Federated Reinforcement Learning: Linear Speedup Under Markovian Sampling

Sajad Khodadadian [1] Pranay Sharma [2] Gauri Joshi [2] Siva Theja Maguluri [1]

Abstract

Since reinforcement learning algorithms are notoriously data-intensive, the task of sampling observations from the environment is usually split across multiple agents. However, transferring these observations from the agents to a central location can be prohibitively expensive in terms of the communication cost, and it can also compromise the privacy of each agent's local behavior policy. In this paper, we consider a federated reinforcement learning framework where multiple agents collaboratively learn a global model, without sharing their individual data and policies. Each agent maintains a local copy of the model and updates it using locally sampled data. Although having N agents enables the sampling of N times more data, it is not clear if it leads to proportional convergence speedup. We propose federated versions of on-policy TD, off-policy TD and Q-learning, and analyze their convergence. For all these algorithms, to the best of our knowledge, we are the first to consider Markovian noise and multiple local updates, and prove a linear convergence speedup with respect to the number of agents. To obtain these results, we show that federated TD and Q-learning are special cases of a general framework for federated stochastic approximation with Markovian noise, and we leverage this framework to provide a unified convergence analysis that applies to all the algorithms.

[1]H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA [2]Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA,. Correspondence to: Sajad Khodadadian <skhodadadian3@gatech.edu>.

Proceedings of the 39 th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1. Introduction

Reinforcement Learning (RL) is an online sequential decision-making paradigm that is typically modeled as a Markov Decision Process (MDP) (Sutton & Barto, 2018). In an RL task, the agent aims to learn the optimal policy of the MDP that maximizes long-term reward, without knowledge of its parameters. The agent performs this task by repeatedly interacting with the environment according to a behavior policy, which in turn provides data samples that can be used to improve the policy. This MDP-based RL framework has a vast array of applications including self-driving cars (Yurtsever et al., 2020), robotic systems (Kober et al., 2013), games (Silver et al., 2016), UAV-based surveillance (Yun et al., 2022), and Internet of Things (IoT) (Lim et al., 2020).

Due to the high-dimensional state and action spaces that are typical in these applications, RL algorithms are extremely data hungry (Duan et al., 2016; Kalashnikov et al., 2018; Akkaya et al., 2019), and training RL models with limited data can result in low accuracy and high output variance (Islam et al., 2017; Xu et al., 2021). However, generating massive amounts of training data sequentially can be extremely time consuming (Nair et al., 2015). Hence, many practical implementations of RL algorithms from Atari domain to Cyber-Physical Systems rely on parallel sampling of the data from the environment using multiple agents (Mnih et al., 2016; Espeholt et al., 2018; Chen et al., 2021a; Xu et al., 2021). It was empirically shown in (Mnih et al., 2016) that the federated version of these algorithms yields faster training time and improved accuracy. A naive approach would be to transfer all the agents' locally collected data to a central server that uses it for training. However, in applications such as IoT (Chen & Giannakis, 2018), autonomous driving (Shalev-Shwartz et al., 2016) and robotics (Kalashnikov et al., 2018), communicating high-dimensional data over low bandwidth network link can be prohibitively slow. Moreover, sharing individual data of the agents with the server might also be undesirable due to privacy concerns (Yang et al., 2019; Mothukuri et al., 2021).

Federated Learning (FL) (Kairouz et al., 2019) is an emerging distributed learning framework, where multiple agents seek to collaboratively train a shared model, while complying with the privacy and data confidentiality requirements

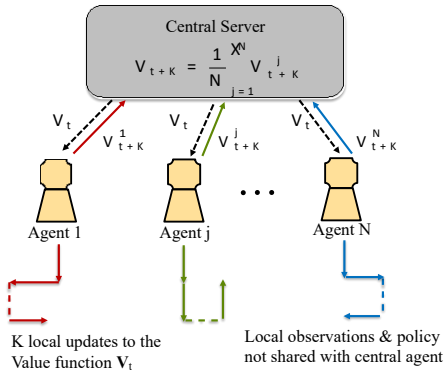

Figure 1. Schematic representation of FRL where N agents follow a Markovian trajectory and synchronize their parameter every K time steps.

(Qi et al., 2021; Yang et al., 2019). The key idea is that the agents collect data, use on-device computation capabilities to locally train the model, and only share the model updates with the central server. Not sharing data reduces communication cost and also alleviates privacy concerns.

Recently, there is a growing interest in employing FL for RL algorithms (also known as FRL) (Nadiger et al., 2019; Liu et al., 2019; Ren et al., 2019; Xu et al., 2021; Zhang et al., 2022). Unlike standard supervised learning where data is collected before training begins, in FRL, each agent collects data by following its own Markovian trajectory, while simultaneously updating the model parameters.

To ensure convergence, after every K time steps, the agents communicate with the central server to synchronize their parameters (see Figure 1). Intuitively, using more agents and a higher synchronization frequency should improve the convergence of training algorithm. However, the following questions remain to be concretely answered:

1. With N agents, do we get an N-fold (linear) speedup in the convergence of FRL algorithms?
2. How does the convergence speed and the final error scale with synchronization frequency K ?

While these questions are well-studied (Wang & Joshi, 2021; Stich, 2018; Qu et al., 2020; Li et al., 2019) in federated supervised learning, only a few works (Wai, 2020; Shen et al., 2020) have attempted to answer them in the context of FRL. However, none of them have established the convergence analysis of FRL algorithms by considering Markovian local trajectories and multiple local updates (see Table 1).

In this paper, we tackle this challenging open problem and answer both the questions listed above. We propose communication-efficient federated versions of on-policy TD, off-policy TD, and Q-learning algorithms. In addition, we are the first to establish the convergence bounds for these algorithms in the realistic Markovian setting, showing a linear speedup in the number of agents. Previous works (Liu & Olshevsky, 2021; Shen et al., 2020) on distributed RL have only shown such a speedup by assuming i.i.d. noise. Moreover, based on experiments, (Shen et al., 2020) conjectures that linear speedup may be possible under the realistic Markov noise setting, which we establish analytically. The main contributions and organization of the paper are summarized below.

- In the on-policy setting, in Section 4 we propose and analyze federated TD-learning with linear function approximation, where the agents' goal is to evaluate a common policy using on-policy samples collected in parallel from their environments. The agents only share the updated value function (not data) with the central server, thus saving communication cost. We prove a linear convergence speedup with the number of agents and also characterize the impact of communication frequency on the convergence.
- In the off-policy setting, in Section 5 we propose and analyze the federated off-policy TD-learning and federated Q-learning algorithms. Again, we establish a linear speedup in their convergence with respect to the number of agents and characterize the impact of synchronization frequency on the convergence. Since every agent samples data using a private policy and only communicates the updated value or Q-function, off-policy FRL helps keep both the data as well as the policy private.
- In Section 6, we propose a general Federated Stochastic Approximation framework with Markovian noise (FedSAM) which subsumes both federated TD-learning and federated Q-learning algorithms proposed above. Considering Markovian sampling noise poses a significant challenge in the analysis of this algorithm. The convergence result for FedSAM serves as a workhorse that enables us to analyze both federated TD-learning and federated Q-learning. We characterize the convergence of FedSAM with a refined analysis of general stochastic approximation algorithms, fundamentally improving upon prior work.

2. Related Work

Single node TD-learning and Q-learning. Most existing RL literature is focused on designing and analyzing algorithms that run at a single computing node. In the on-policy setting, the asymptotic convergence of TD-learning was established in (Tsitsiklis & Van Roy, 1997; Tadić, 2001; Borkar, 2009), and the finite-sample bounds were studied

Table 1. Comparison of sample complexity results for federated supervised learning (local SGD) and reinforcement learning algorithms. The possible distributed architectures are: 1) Worker-server, with a central server that coordinates with N agents; 2) Decentralized, where each agent directly communicates with its neighboring agents, without a central server; and 3) Shared memory, where each agent modifies a subset of the parameters of a global model held in a shared memory, that is accessible to all agents. (Recht et al., 2011).

Algorithm	Architecture	References	Local Updates	Markov Noise	Linear Speedup
Local SGD	Worker-server	(Khaled et al., 2020)	✓	7	✓
Local SGD	Worker-server	(Spiridonoff et al., 2021)	✓	7	✓
TD(0)	Worker-server	(Liu & Olshevsky, 2021)	✓	7	✓
Stochastic Approximation	Decentralized	(Wai, 2020)	7	✓	7
A3C-TD(0)	Shared memory	(Shen et al., 2020)	7	7	✓
A3C-TD(0)	Shared memory	(Shen et al., 2020)	7	✓	7
TD & Q-learning	Worker-server	This paper	✓	✓	✓

in (Dalal et al., 2018; Lakshminarayanan & Szepesvari, 2018; Bhandari et al., 2018; Srikant & Ying, 2019; Hu & Syed, 2019; Chen et al., 2021c). In the off-policy set-ting, (Maei, 2018; Zhang et al., 2020) study the asymptotic and (Chen et al., 2020a; 2021c) characterize the finite time bound of TD-learning. The Q-learning algorithm was first proposed in (Watkins & Dayan, 1992). There has been a long line of work to establish the convergence properties of Q-learning. In particular, (Tsitsiklis, 1994; Jaakkola et al., 1994; Bertsekas & Tsitsiklis, 1996b; Borkar & Meyn, 2000; Borkar, 2009) characterize the asymptotic convergence of Q-learning, (Beck & Srikant, 2012b; 2013; Wainwright, 2019; Chen et al., 2020a; 2021c) study the finite-sample convergence bound in the mean-square sense, and (Even-Dar & Mansour, 2004; Li et al., 2020; Qu & Wierman, 2020) study the high-probability convergence bounds of Q-learning.

Federated Learning with i.i.d. Noise. When multiple agents are used to expedite sample collection, transferring the samples to a central server for the purpose of training can be costly in applications with high-dimensional data (Shao et al., 2019) and it may also compromise the agents' privacy. Federated Learning (FL) is an emerging distributed optimization paradigm (Konečnỳ et al., 2016; Kairouz et al., 2019) that utilizes local computation at the agents to train models, such that only model updates, not data, is shared with the central server. In local Stochastic Gradient Descent (Local SGD or FedAvg) (McMahan et al., 2017; Stich, 2018), the core algorithm in FL, locally trained models are periodically averaged by the central server in order to achieve consensus among the agents at a reduced communication cost. While the convergence of local SGD has been extensively studied in prior work (Khaled et al., 2020; Spiridonoff et al., 2021; Qu et al., 2020; Koloskova et al., 2020), these works assume i.i.d. noise in the gradients, which is acceptable for SGD but too restrictive for RL algorithms.

Distributed and Multi-agent RL. Some recent works have analyzed distributed and multi-agent RL algorithms in the presence of Markovian noise in various settings such as decentralized stochastic approximation (Doan et al., 2019; Sun et al., 2020; Wai, 2020; Zeng et al., 2020), TD learning with linear function approximation (Wang et al., 2020a), and off-policy TD in actor-critic algorithms (Chen et al., 2021e;f). However, all these works consider decentralized settings, where the agents communicate with their neighbors after every local update. On the other hand, we consider a federated setting, with each agent performing multiple local updates between successive communication rounds, thereby resulting in communication savings. In (Shen et al., 2020), a parallel implementation of asynchronous advantage actor-critic (A3C) algorithm (which does not have local updates) has been proposed under both i.i.d. and Markov sampling. However, the authors prove a linear speedup only for the i.i.d. case, and an almost linear speedup is observed experimentally for the Markovian case.

3. Preliminaries: Single Node Setting

We model our RL setting with a Markov Decision Process (MDP) with 5 tuples $(S; A; P; R;)$, where S and A are finite sets of states and actions, P is the set of transition probabilities, R is the reward function, and 2 $(0; 1)$ denotes the discount factor. At each time step t, the system is in some state S_t, and the agent takes some action A_t according to a policy (jS_t) in hand, which results in reward $R(S_t; A_t)$ for the agent. In the next time step, the system transitions to a new state S_{t+1} according to the state transition probability $P(jS_t; A_t)$. This series of states and actions $(S_t; A_t)_{t0}$ constructs a Markov chain, which is the source of the Markovian noise in RL. Throughout this paper we assume that this Markov chain is irreducible and aperiodic (also known as ergodic). It is known that this Markov

chain asymptotically converges to a steady state, and we denote its stationary distribution with .

To measure the long-term reward achieved by following policy , we define the value function

$$V(s) = E\left[\sum_{t=0}^{1} {}^{t} R(S_t; A_t) | S_0 = s; A_t \ (|S_t) \right] : \quad (1)$$

Equation (1) is the tabular representation of the value function. Sometimes, however, the size of the state space $|S|$ is large, and storing $V(s)$ for all $s \ 2 \ S$ is computationally infeasible. Hence, a low dimensional vector $v \ 2 \ R^d$, where $d \ |S|$, can be used to approximate the value function as $V(s) \ (s)^{>} v$ (Tsitsiklis & Van Roy, 1997). Here $(s) \ 2 \ R^d$ is a given feature vector corresponding to the state s. Using a low-dimensional vector v to approximate a high-dimensional vector $(V(s))_{s2S}$ is referred to as the function approximation paradigm in RL. For each $(s; a)$ pair, we also define the Q-function, $Q(s; a) = E[\sum_{t=0}^{1} {}^{t} R(S_t; A_t) | S_0 = s; A_0 = a]$, which will be employed in Q-learning.

3.1. Temporal Difference Learning

An intermediate goal in RL is to estimate the value function (either $(V(s))_{s2S}$ or v) corresponding to a particular pol-icy using data collected from the environment. This task is denoted as policy evaluation and one of the commonly-used approaches to accomplish this is Temporal Difference (TD)-learning (Sutton, 1988). TD-learning is an iterative algorithm where the elements of a d (or $|S|$, in the tabular setting) dimensional vector is updated until it converges to v (or V). This evaluated value function can be employed in different RL algorithms such as actor-critic (Konda & Tsitsiklis, 2000). In the on-policy function approximation setting, the update of the n-step TD-learning is as follows

$$\begin{aligned} &\text{Sample} \quad A_{t+n} \ (|S_{t+n}); S_{t+n+1} \ P(|S_{t+n}; A_{t+n}) \\ &\text{update} \quad v_{t+1} = v_t + (S_t) \sum_{l=t}^{t+n \ 1} {}^{l \ t} (R(S_l; A_l) \\ &\qquad + (S_{l+1})^{>} v_t \quad (S_l)^{>} v_t); \end{aligned} \quad (2)$$

where is the step size. Note that in this setting, the evaluating policy and the sampling policy coincide. In contrast, in the off-policy setting these two policies can in general differ, and we need to account for this difference while running the algorithm. We will further expand on TD-learning and its variants in Sections 4.1 and 5.1.1.

3.2. Control Problem and Q-learning

Assuming some initial distribution on the state space, the average value function corresponding to policy is defined as $V() = E_s[V(s)]$. This scalar quantity is a metric of average long-term rewards achieved by the agent, when it starts from distribution and follows policy . The ultimate goal of the agent is to obtain an optimal policy which results in the maximum long-term rewards, i.e. $2 \arg \max \ V()$. Throughout the paper, we denote the parameters corresponding to the optimal policy with , e.g., $V() \ V()$. The task of obtaining the optimal policy in RL is denoted as the control problem.

Q-learning (Watkins & Dayan, 1992) is one of the most widely used algorithms in RL to solve the control problem. At each time step t, Q-learning preserves a $|S|:|A|$ dimensional table Q_t, and updates it table as $Q_{t+1}(s; a) = Q_t(S_t; A_t) + (R(S_t; A_t) + \max_a Q_t(S_{t+1}; a) \ Q_t(S_t; A_t))$, if $(s; a) = (S_t; A_t)$ and $Q_{t+1}(s; a) = Q_t(s; a)$ otherwise. The $|S||A|$ elements of the vector Q_t are updated iteratively until it converges to Q, corresponding to an optimal policy. Using Q, one can obtain an optimal policy via greedy selection.

3.3. Stochastic Approximation and Finite Sample Bounds

Both TD-learning and Q-learning can be seen as variants of stochastic approximation (Chen et al., 2020b; 2019b;a; 2021d; Tsitsiklis, 1994). While generic stochastic approximation algorithms are studied under i.i.d. noise (Even-Dar & Mansour, 2004; Shah & Xie, 2018; Wainwright, 2019; Liu et al., 2015; Dalal et al., 2018), to apply them for studying RL we need to understand stochastic approximation under Markovian noise (Tsitsiklis, 1994; Qu & Wierman, 2020; Srikant & Ying, 2019; Chen et al., 2021c) which is significantly more challenging.

For a generic stochastic approximation (i.i.d. or Markovian noise) with constant step size , parameter vector x_T, and convergent point x, it can be shown that the algorithm have the following convergence behaviour

$$E[kx_T \quad xk^2] \ C_1(1 \quad C_0)^T + C_2; \quad (3)$$

where $C_0; C_1$; and C_2 are some problem dependent positive constants (Look at Appendix A for a discussion on a lower bound on the convergence of general stochastic approximation). The first term is denoted as the bias and the second term is called the variance. According to this bound, x_T geometrically converges to a ball around x with radius proportional to C_2. Notice that we can always reduce the variance term by reducing the step size , but this will lead to slower convergence in the bias term. In particular, in order to get $E[kx_T \quad xk^2]$, it is easy to see that we need $T \ O(\frac{C_2}{} \log \frac{1}{})$ sample complexity. Now suppose the constant C_2 is large. In this case, the variance term in the bound in (3) is large, and the sample complexity, which is proportional to C_2 will be poor. Notice that by the dis-

cussion in Appendix A, this bound is tight and cannot be improved.

This is where the FL can be employed in order to control the variance term by generating more data. For instance, in federated TD-learning, multiple agents work together to evaluate the value function simultaneously. Due to this collaboration, the agents can estimate the true value function with a lower variance. The same holds for estimating Q in Q-learning.

4. Federated On-policy RL

4.1. On-policy TD-learning with linear function approximation

In this section we describe the TD-learning with linear function approximation and online data samples in the single node setting. In this problem, we consider a full rank feature matrix $\in \mathbb{R}^{|S| \times d}$, and we denote s-th row of this matrix with (s), $s \in S$. The goal is to find $v \in \mathbb{R}^d$ which solves the following fixed point equation:

$$v = \Pi((T^{(n)}v)). \tag{4}$$

In equation (4), $\Pi()$ is the projection with respect to the weighted 2-norm, i.e., $\Pi(V) = \arg\min_{v \in \mathbb{R}^d} \|v - V\|$. Here $\|V\| = \sqrt{V^\top V}$ and $$ is a diagonal matrix with diagonal entries corresponding to $$. In equation (4), $(T^{(n)})$ denotes the n-step Bellman operator (Tsitsiklis & Van Roy, 1997). It is known (Tsitsiklis & Van Roy, 1997) that equation (4) has a unique solution v, and v is "close" to the true value function V. n-step TD-learning algorithm, which was shown in (2), is an iterative algorithm to obtain this unique fixed point using samples from the environment. Note that in this algorithm states and actions are sampled over a single trajectory, and hence the noise in updating v_t is Markovian. Furthermore, since the policy which samples the actions and the the evaluating policy are both , this algorithm is on-policy. As described in (Tsitsiklis & Van Roy, 1997; Bertsekas & Tsitsiklis, 1996a), the TD-learning algorithm can be studied under the umbrella of linear stochastic approximation with Markovian noise. More recently, the authors in (Bhandari et al., 2018; Srikant & Ying, 2019) have shown that the update parameter of TD-learning v_t converges to v in the form $\mathbb{E}[\|v_t - v\|^2] \le O((1 - C_0)^t +)$. In the next section we show how FL can improve this result.

4.2. Federated TD-learning with linear function approximation

The federated version of on-policy n-step TD-learning with linear function approximation is shown in Algorithm 1. In this algorithm we consider N agents which collaboratively work together to evaluate v. For each agent $i; i = 1; 2; \dots; N$, we initialize their corresponding parameters $v_0^i = 0$. Furthermore, each agent i samples its initial state S_0^i from some given distribution . In the next time steps, each agent follows a single Markovian trajectory generated by policy , independently from other agents. At each time t, the parameter of each agent i is updated using this independently generated trajectory as $v_{t+1}^i = v_t^i + (S_t^i) E_{t;n}^i$. Finally, in order to ensure convergence to a global optimum, every K time steps all the agents send their parameters to a central server. The central server evaluates the average of these parameters and returns this average to each of the agents. Each agent then continues their update procedure using this average.

Notice that the averaging step is essential to ensure synchronization among the agents. Smaller K results in more frequent synchronization, and hence better convergence guarantees. However, setting smaller K is equivalent to more number of communications between the single agents and the central server, which incurs higher cost. Hence, an intermediate value for K has to be chosen to strike a balance between the communication cost and the accuracy. At the end, the algorithm samples a time step $\hat{T}$ q_T^c, where

$$q_T^c(t) = \frac{c^t}{\sum_{t'=0}^{T-1} c^{t'}} \quad \text{for} \quad t = 0; 1; \dots; T-1 \tag{5}$$

and $c > 1$ is some constant. Since we have $q_T^c(t) > 0$ and $\sum_{t=0}^{T-1} q_T^c(t) = 1$, it is clear that $q_T^c()$ is a probability distribution over the time interval $[0; T-1]$. In Theorem 4.1 we characterize the convergence of this algorithm as a function of , N, and K. Throughout the paper, $\tilde{O}()$ ignores the logarithmic terms.

Algorithm 1 Federated n-step TD (On-policy, Function Approx.)

Input: Policy ;
Initialization: $v_0^i = 0$ and S_0^i and $\{A^i; S_l^i\}_{l=1}^{n-1}$ for $0 \le l \le n-1$ and all i
for $t = 0$ to $T-1$ do
  for $i = 1; \dots; N$ do
    Sample A_{t+n}^i (S_{t+n}^i); S_{t+n+1}^i
    $e_{t;l}^i = \prod(S_{t+n-l}^i, A_{t+n-l}^i)$ (S_{t+n}^i) $+ (S_{t+1}^i)^\top v_t^i - (S_t^i)^\top v_t^i$ for $l = t; \dots; t+n-1$
    $E_{t;n}^i = \sum_{l=t}^{t+n-1} e_{t;l}^i$
    $v_{t+1}^i = v_t^i + (S_t^i) E_{t;n}^i$
  end for
  if $t+1 \bmod K = 0$ then
    $v_{t+1}^i \leftarrow \frac{1}{N} \sum_{j=1}^{N} v_{t+1}^j; \forall i \in [N]$
  end if
end for
Sample $\hat{T} \sim q_T^{TODO}$
Return: $\frac{1}{N} \sum_{i=1}^{N} v_{\hat{T}}^i$

Theorem 4.1. Let $v_{\hat{T}} = \frac{1}{N}\sum_{i=1}^{N} v_{\hat{T}}^i$ denote the average of the parameters across agents at the random time $\hat{T}$. For small enough step size α, and $T = O(\log \frac{1}{\epsilon})$, there exist constant $c_{TDL} \in (0,1)$ (see Section C.1 for precise statement), such that we have

$$\mathbb{E}[\|v_{\hat{T}} - v^*\|_2^2] \le C_1^{TDL}(1-\alpha C^{TDL})^T \|v_0 - v^*\|_2^2 + C_2^{TDL}\frac{\alpha}{N} + C_3^{TDL}(K-1)^2\alpha^2 + C_4^{TDL}\alpha^3.$$

where C_i^{TDL}, $i = 0,1,2,3,4$ are problem dependent constants, and $\alpha = O(\log(1/\epsilon))$. By choosing $\alpha = O(\frac{\log(NT)}{T})$ and $K = T/N$, we achieve $\mathbb{E}[\|v_{\hat{T}} - v^*\|_2^2] \le \epsilon$ within $T = O(\frac{1}{N\epsilon})$ iterations.

For brevity purposes, here we did not show the exact dependence of the constants C_i^{TDL}, $i = 0,1,2,3$ on the problem dependent constants. For a discussion on the detailed expression look at Section C.1 in the appendix.

Theorem 4.1 shows that federated TD-learning with linear function approximation enjoys a linear speedup with respect to the number of agents. Compared to the convergence bound of general stochastic approximation in (3), the bound in Theorem 4.1 has three differences. Firstly, the variance term which is proportional to the step size α is divided with the number of agents N. This will allow us to control the variance (and hence improve sample complexity) by employing more number of agents. Secondly, we have an extra term which is zero with perfect synchronization $K = 1$. Although this term is not divided with N, but it is proportional to α^2, which is one order higher than the variance term in (3). Finally, the last term is of the order $O(\alpha^3)$, which can be handled by choosing small enough step size.

Furthermore, according to the choice of K in Theorem 4.1, after T iterations, the communication cost of federated TD is $T/K = N$. However, by employing federated TD-learning in the naive setting where all the agents communicate with the central server at every time step, the communication cost will be $O(T)$. Hence, we observe that by carefully tuning the hyper parameters of federated TD, we can significantly reduce the communication cost of the overall algorithm, while not loosing performance in terms of the sample complexity.

Finally, federated TD-learning Algorithm 1 preserves the privacy of the agents. In particular, since the single agents only require to share their parameters v_{t+1}^i, the central server will not be exposed to the state-action-reward trajectory generated by each agent. This can be essential in some applications where privacy is an issue (Mothukuri et al., 2021; Truex et al., 2019). Examples of such applications include autonomous driving (Liang et al., 2019; Zhao et al., 2021), Internet of Things (IoT) (Nguyen et al., 2021; Ren et al., 2019; Wang et al., 2020b), and cloud robotics (Liu et al., 2019; Xu et al., 2021).

Remark. In algorithm 1, the randomness in choosing $\hat{T}$ is independent of all the other randomness in the problem. Hence, in a practical setting, one can sample $\hat{T}$ ahead of time, before running the algorithm, and stop the algorithm at time step $\hat{T}$ and output $v_{\hat{T}}$. By this method, we require only a single data point to be saved, which results in the memory complexity of $O(1)$ for the algorithm.

5. Federated Off-Policy RL

On-policy TD-learning requires online sampling from the environment, which might be costly (e.g. robotics (Gu et al., 2017; Levine et al., 2020)), high risk (e.g. self-driving cars (Yurtsever et al., 2020; Maddern et al., 2017)), or unethical (e.g. in clinical trials (Gottesman et al., 2019; Liu et al., 2018; Gottesman et al., 2020)). Off-policy training in RL refers to the paradigm where we use data collected by a fixed behaviour policy to run the algorithm. When employed in federated setting, off-policy RL has privacy advantages as well (Foerster et al., 2016; Qi et al., 2021; Zhuo et al., 2019). In particular, suppose each single agent attains a unique sampling policy, and they do not wish to reveal these policies to the central server. In off-policy FL, agents only transmit sampled data, and hence the sampling policies remain private to each agent.

In Section 5.1 we will discuss off-policy TD-learning and in Section 5.2 we will discuss Q-learning, which is an off-policy control algorithm. For the off-policy algorithms, we only study the tabular setting. Notice that it has been observed that the combination of off-policy sampling and function approximation in RL (also known as deadly triad (Sutton & Barto, 2018)) can result in instability or even divergence (Baird, 1995). Recently there has been some work to overcome deadly triad (Chen et al., 2021b). Extension of our work to function approximation in the off-policy setting is a future research direction.

5.1. Federated Off-Policy TD-learning

In the following, we first discuss single-node off-policy TD-learning, and then we generalize it to the federated setting.

5.1.1. Off-policy TD-learning

In off-policy TD-learning the goal is to evaluate the value function $V = (V(s))_{s\in S}$ corresponding to the policy π using data sampled from some fixed behaviour policy π_b. In this setting, the evaluating policy π and the sampling policy π_b can be arbitrarily different, and we need to account for this difference while performing the evaluation. Although π and π_b can be different, notice that the value function V^π does not depend on π_b. In order to account for this

difference, we introduce the notion of importance sampling as $l^b(s,a) = \frac{(a|s)}{b(a|s)}$ which is employed in the off-policy TD-learning.

Recently, several works studied the finite-time convergence of off-policy TD-learning. In particular, the authors in (Khodadadian et al., 2021; Chen et al., 2021c; 2020b; 2021d) show that, similar to on-policy TD, off-policy TD-learning can be studied under the umbrella of stochastic approximation. Hence, this algorithm enjoys similar convergence behaviour as (3).

5.1.2. Federated off-policy TD-learning

The federated version of n-step off-policy TD-learning is shown in Algorithm 2. In this algorithm, each agent i attains a unique (and private) sampling policy β^i and follows an independent trajectory generated by this policy. Furthermore, at each time step t, each agent i attains a $|\mathcal{S}|$-dimensional vector V_t^i and updates this vector using the samples generated by β^i. In order to account for the off-policy sampling, each agent utilizes $l^{(i)}(S^i, A^i) = \frac{\pi(A^i|S^i)}{\beta^i(A^i|S^i)}$ in the update of their algorithm. We further define $l_{max} = \max_{s,a,i} l^{(i)}(s,a)$, which is a measure of discrepancy between the evaluating policy π and sampling policy β^i of all the agents.

In order to ensure synchronization, all the agents transmit their parameter vectors to the central server every K time steps. The central server returns the average of these vectors to each agent and each agent follows this averaged vector afterwards. Notice that in federated off-policy TD-learning Algorithm 2, each agent share neither their sampled trajectory of state-action-rewards, nor their sampling policy with the central server. This provides two levels of privacy for the single agents. At the end, the algorithm samples a time step $\hat{T} \sim q_T^{c_{TD}}$, where the distribution q_T^c is defined in (5) and $c_{TD} = 1 - \frac{\omega_{TD}}{2}$, where $\omega_{TD} = 1 - \frac{0.5e^{1-4}(2-\min(1-n+1))}{\sqrt{e\left(1+\frac{2-\min(1-n+1)}{2-2\min(1-n+1)}\right)^2}}$. Here, we denote $\min = \min_{s,i} \beta^i(s)$. The constant c_{TD} is carefully chosen to ensure the convergence of Algorithm 2. Furthermore, for small enough step size α, it can be shown that $0 < c_{TD} < 1$.
Theorem 5.1 states the convergence of this Algorithm.

Theorem 5.1. Consider the federated n-step off-policy TD-learning Algorithm 2. Denote $V_{\hat{T}} = \frac{1}{N}\sum_{i=1}^N V_{\hat{T}}^i$. For small enough step size α and large enough T, we have

$$E[\|V_{\hat{T}} - V^*\|_2^2] \leq C_1^{TD_T}\frac{1}{c_{TD}T} + C_2^{TD_T}\alpha^2 + \frac{C_3^{TD_T}}{N(K-1)^2},$$

where $C_1^{TD_T} = C_1^{TD_T}:\frac{l_{max}^{4n-2}}{\min(1-\omega_{TD})}$, $C_2^{TD_T} = C_2^{TD_T}:\frac{l_{max}^{3n-1}|\mathcal{S}|\log^2(|\mathcal{S}|)}{\min^2(1-\omega_{TD})^4}$, $C_3^{TD_T} = C_3^{TD_T}:\frac{l_{max}^{7n-3}|\mathcal{S}|^2\log^2(|\mathcal{S}|)}{\min^4(1-\omega_{TD})^8}$, and $C_i^{TD_T}$, $i = 1,2,3$ are universal problem independent constants. In addition, choosing $\alpha = \frac{8\log(NT)}{T\omega_{TD}}$ and $K = T/N$, we have $E[\|V_{\hat{T}} - V^*\|_2^2]$ after $T = O\left(\frac{1}{N}:\frac{l_{max}^{7n-3}|\mathcal{S}|^2\log^2(|\mathcal{S}|)}{\min^5(1-\omega_{TD})^9}\right)$ iterations.

Algorithm 2 Federated n-step TD (Off-policy Tabular Setting)

Input: Policy π;
Initialization: $V_0^i = 0$ and S_0^i and $\{A^i, S_{1-n}^i\}_{i=1}^N$ for $0 \leq t \leq n-1$ and all i
for $t = 0$ to $T-1$ do
  for $i = 1,\dots,N$ do
    Sample $A_{t+n}^i \sim \beta^i(\cdot|S_{t+n}^i)$; $S_{t+n}^i \sim P(\cdot|S_{t+n}^i, A_{t+n}^i)$
    $\delta_t^i(l) = R(S_l^i, A_l^i) + V_t^i(S_{l+1}^i) - V_t^i(S_l^i)$
    Update $V_{t+1}^i(s) = V_t^i(s) + \sum_{l=t}^{t+n-1}\alpha\gamma^{l-t}l^{(i)}(S_j^i, A_j^i)\delta_t^i(l)$ if $s = S_t^i$ and $V_{t+1}^i(s) = V_t^i(s)$ otherwise.
  end for
  if $t+1 \bmod K = 0$ then
    $V_{t+1}^i \leftarrow \frac{1}{N}\sum_{j=1}^N V_{t+1}^j, \forall i \in [N]$
  end if
end for
Sample $\hat{T} \sim q_T^{c_{TD}}$
Return: $\frac{1}{N}\sum_{i=1}^N V_{\hat{T}}^i$

The proof is given in Section C.2 in the appendix.

Note that similar to on-policy TD-learning Algorithm 1, off-policy TD-learning also enjoys a linear speedup while maintaining a low communication cost. In addition, this algorithm preserves the privacy of the agents by holding both the data and the sampling policy private.

5.2. Federated Q-learning

So far we have discussed policy evaluation problem with on and off-policy samples. Next we aim at solving the control problem by employing the celebrated Q-learning algorithm (Watkins & Dayan, 1992; Tsitsiklis, 1994). In the next section we will explain the Q-learning algorithm. Further, in Section 5.2.2 we will provide a federated version of Q-learning along with its convergence result.

5.2.1. Q-learning

The goal of Q-learning is to evaluate Q, which is the unique Q-function corresponding to the optimal policy. Knowing Q, one can obtain an optimal policy through a greedy selection (Puterman, 2014), and hence resolve the control problem.

Suppose $\{S_t, A_t\}_{t\geq 0}$ is generated by a fixed behaviour policy π_b. At each time step t, Q-learning preserves a $|S|\times|A|$ table Q_t and updates the elements of this table as shown in Section 3.2. By assuming π_b to be an ergodic policy, the asymptotic convergence of Q_t to Q^* has been established in (Bertsekas & Tsitsiklis, 1996b). Furthermore, it can be shown that Q-learning is a special case of stochastic approximation and enjoys a convergence bound similar to (3) (Beck & Srikant, 2012a; Li et al., 2020; Qu & Wierman, 2020; Chen et al., 2021c).

Two points worth mentioning about the Q-learning algorithm. Firstly, Q-learning is an off-policy algorithm in the sense that only samples from a fixed ergodic policy is needed to perform the algorithm. Secondly, as opposed to the TD-learning, the update of the Q-learning is non-linear. This imposes a sharp contrast between the analysis of Q-learning and TD-learning (Chen et al., 2019a).

5.2.2. Federated Q-learning

Algorithm 3 provides the federated version of Q-learning. We characterize its convergence in the following theorem.

Algorithm 3 Federated Q-learning

1: Input: Sampling policy π_b^i for $i = 1, 2, \dots, N$, initial distribution
2: Initialization: $Q_0^i = 0$ and S_0^i for all i
3: **for** $t = 0$ to $T-1$ **do**
4: **for** $i = 1, \dots, N$ **do**
5: Sample $A_t^i \sim \pi_b^i(\cdot|S_t^i)$; $S_{t+1}^i \sim P(\cdot|S_t^i, A_t^i)$
6: Update $Q_{t+1}^i(s,a) = Q_t^i(S_t^i, A_t^i) + \alpha_t\big(R(S_t^i, A_t^i) + \gamma \max_{a'} Q_t(S_{t+1}^i, a') - Q_t(S_t^i, A_t^i)\big)$, if $(s,a) = (S_t^i, A_t^i)$ and $Q_{t+1}^i(s,a) = Q_t(s,a)$ otherwise.
7: **end for**
8: **if** $t+1 \bmod K = 0$ **then**
9: $Q_{t+1}^i = \frac{1}{N}\sum_{j=1}^N Q_{t+1}^j, \forall i \in [N]$
10: **end if**
11: **end for**
12: Sample: $\hat{T} \sim q_T^{QC}$
13: Return: $\frac{1}{N}\sum_{i=1}^N Q_{\hat{T}}^i$

Theorem 5.2. Consider the federated Q-learning Algorithm 3 with $c_Q > \frac{1}{(1-\gamma)^2}$ $(0,1)$, where $\gamma' = 1 - \frac{0.5e^{1-4(2-\min(1-\gamma))}}{\sqrt{e\left(1+\frac{2}{2-\min(1-\gamma)}\right)^{2^2}}}$ and we denote $\mu_{\min} = \min_{s,a,i} \pi_b^i(s)\pi_b^i(a|s)$. Denote $\bar{Q}_{\hat{T}} = \frac{1}{N}\sum_{i=1}^N Q_{\hat{T}}^i$. For small enough step size α and large enough T, we have

$$\mathbb{E}[\|\bar{Q}_{\hat{T}} - Q^*\|_2^2] \leq C_1^Q \frac{1}{c_Q T}\alpha + C_2^Q \alpha^2 + C_3^Q \frac{1}{N}(K-1)\alpha^2,$$

where $C_1^Q = {}_1C^Q \cdot \frac{1}{\mu_{\min}(1-\gamma)^3}$, $C_2^Q = C_2^Q \cdot \frac{|S||A|\log^2(|S||A|)}{\mu_{\min}^2(1-\gamma)^4}$, $C_3^Q = C_3^Q \cdot \frac{|S||A|^2\log^2(|S||A|)}{\mu_{\min}^4(1-\gamma)^8}$, and C_i^Q, $i = 1, 2, 3$ are universal problem independent constants. In addition, choosing $\alpha = \frac{8\log(NT)}{T(1-\gamma)}$ and $K = T/N$, we have $\mathbb{E}[\|\bar{Q}_{\hat{T}} - Q^*\|_2^2]$ within $T = \tilde{\mathcal{O}}\left(\frac{1}{N}\cdot\frac{|S||A|^2\log^2(|S||A|)}{\mu_{\min}^5(1-\gamma)^9}\right)$ iterations.

According to Theorem 5.2, federated Q-learning Algorithm 3, similar to federated off-policy TD-learning, enjoys linear speedup, communication efficiency as well as privacy guarantees. We would like to emphasize that the update of Q-learning is non-linear. Hence the result of Theorem 5.2 cannot be derived from Theorems 4.1 and 5.1.

6. Generalized Federated Stochastic Approximation

In this section we study the convergence of a general federated stochastic approximation for contractive operators, FedSAM, which is presented in Algorithm 4. In this algorithm there are N agents $i = 1, 2, \dots, N$. At each time step $t \geq 0$, each agent i maintains the parameter $\theta_t^i \in \mathbb{R}^d$. At time $t = 0$, all agents initialize their parameters with $\theta_0^i = \theta_0$. Next, at time $t \geq 0$, each agent i updates its param-eter as $\theta_{t+1}^i = \theta_t^i + \alpha_{t+1}\big(G^i(\theta_t^i, y_t^i)\theta_t^i + b^i(y_t^i)\big)$. Here α_t denotes the step size, and y_t^i is a noise which is Marko-vian along the time t, but is independent across the agents i. This notion is defined more concretely in Assumption 6.4. We note that functions $G^i(\cdot,\cdot)$ and $b^i(\cdot)$ are allowed to be dependent on the agent i. This allows us to employ the convergence bound of FedSAM in order to derive the convergence bound of off-policy TD-learning with different behaviour policies across agents. In order to avoid divergence, every K time steps we synchronize the parameters of all the agents as $\theta_t^i = \theta_t = \frac{1}{N}\sum_{j=1}^N \theta_t^j, \forall i \in [N]$. Note that although smaller K corresponds to more frequent synchronization and hence more "accurate" updates, at the same time it results in a higher communication cost, which is not desirable. Hence, in order to determine the optimal choice of synchronization period, it is essential to characterize the dependence of the convergence on K. This is one of the results which we will derive in Theorem B.1. Finally, the algorithm samples $\hat{T} \sim q_T^c$, where $q_T^c(t) = \frac{\alpha_t \sum_{t'=t}^{T-1}\alpha_{t'}}{\sum_{t'=0}^{T-1}\alpha_{t'}}$ and outputs $\theta_{\hat{T}}$. This sampling scheme is essential for the convergence of overall algorithm. We further make some assumptions regarding the underlying process.

First, we assume that the expectation of $G^i(\cdot, y^i)$ geometrically converges to some function $G^i(\cdot)$ and the expectation of $b^i(y^i)$ geometrically converges to 0. In particular, we have the following assumption.

Assumption 6.1. For every agent i, there exist a function $G^i(\cdot)$ such that we have

$$\lim_{t\to\infty} \mathbb{E}[G^i(\cdot, y_t^i)] = G^i(\cdot)$$
$$\lim_{t\to\infty} \mathbb{E}[b^i(y_t^i)] = 0. \tag{6}$$

Algorithm 4 Federated Stochastic Approximation with Markovian Noise (FedSAM)

Input: c_{FSAM}; T; $_0$; K; $^i_0 = _0$ for all $i = 1; : : : ; N$.
for $t = 0$ to $T - 1$ do
  $^i_{t+1} = ^i_t + G^i(^i_t; y^i_t) + b^i_t(y^i_t)$; $8i \in [N]$
  if $t + 1 \bmod K = 0$ then
    $^i_{t+1} = _{t+1}$, $\frac{1}{N}\sum_{j=1}^{N} {}^j_{t+1}$; $8 i \in [N]$
  end if
end for
Sample $\hat{T} \sim q_{c_{FSAM}}(T)$.
Return: $\frac{1}{N}\sum_{i=1}^{N} {}^i_{\hat{T}}$

Furthermore, there exists $m_1; m_2 \geq 0$ and $\in [0; 1)$, such that for every $i = 1; 2; : : : ; N$,

$$\begin{aligned} \|G^i(\,) - \mathbb{E}[G^i(\,; y^i)]\|_c &\leq m_1\|\,\|_c\, ^t \\ \|\mathbb{E}[b^i(y^i_t)]\|_c &\leq m_2\, ^t; \end{aligned} \tag{7}$$

where $\|\cdot\|_c$ is a given norm.

Next, we assume a contraction property on the expected operator $G^i(\,)$.

Assumption 6.2. We assume all expected operators $G^i(\,)$ are contraction mappings with respect to $\|\cdot\|_c$ with contraction factor $_c \in (0; 1)$. That is, for all $i = 1; 2; : : : ; N$,

$$\|G^i(_1) - G^i(_2)\|_c \leq _c\|_1 - _2\|_c; \quad 8\ _1; _2 \in \mathbb{R}^d.$$

Next, we consider some Lipschitz and boundedness properties on $G^i(\,;\,)$ and $b^i(\,)$.

Assumption 6.3. For all $i = 1; : : : ; N$, there exist constants A_1, A_2 and B such that

1. $\|G^i(_1; y^i) - G^i(_2; y^i)\|_c \leq A_1\|_1 - _2\|_c$, for all $_1; _2; y^i$.
2. $\|G^i(\,; y^i)\|_c \leq A_2\|\,\|_c$ for all $\,; y^i$.
3. $\|b^i(y^i)\|_c \leq B$ for all y^i.

Remark. By Assumption 6.2 and due to the Banach fixed point theorem, $G^i(\,)$ has a unique fixed point for all $i = 1; 2; : : : ; N$. Furthermore, by Assumption 6.3, we have $G^i(0; y) = 0$. Hence the point 0 is the unique fixed point of $G^i(\,)$.

Finally, we impose an assumption on the random data y^i_t.

Assumption 6.4. We assume that the Markovian noise y^i_t (Markovian with respect to time t) is independent across agents i. In other words, for all measurable functions $f(\,)$ and $g(\,)$, we assume the following

$$\mathbb{E}_{t\ r}[f(y^i_t)\ g(y_t)^j] = \mathbb{E}_{t\ r}[f(y_t)]\ \mathbb{E}_{t\ r}[g(y_t)]^j$$

for all $r \leq t; i = j$.

Theorem 6.1 states the convergence of Algorithm 4.

Theorem 6.1. Consider the federated stochastic approximation Algorithm 4 with $c_{FSAM} = 1$ $'_2 \in (0; 1)$ ($'_2$ is defined in Equation (14) in the appendix), and synchronization frequency K. Denote $_t = \frac{1}{N}\sum_{i=1}^{N} {}^i_t$, and consider $\hat{}$ as the output of this algorithm after T iterations. Assume $= d2 \log \frac{1}{}e$. For $T \geq \max\{K + ; 2\}$ and small enough step size , we have

$$\mathbb{E}[\|_T - \hat{}\|_c^2] \leq C_1 c^T \frac{^2}{^{2+1}} + C_2 \frac{^2}{^{2+1}_{FSAM}} + C_3(K\ ^2 \overline{1})^2 + C_4 {}^3 {}^2; \tag{8}$$

where C_i, $i = 1; 2; 3; 4$ are some constants which are specified precisely in Appendix B, and are independent of $K; ; N$. Choosing $= \frac{8\log(NT)}{T}$ and $K = T/N$, we get $T = \tilde{O}\left(\frac{1}{N}\right)$ sample complexity for achieving $\mathbb{E}[\|_T - \hat{}\|_c^2]$.

Theorem 6.1 establishes the convergence of $\hat{}$ to zero in the expected mean-squared sense. The first term in (8) converges geometrically to zero as T grows. The second term is proportional to similar as (3). However, the number of agents N in the denominator ensures linear speedup, meaning that for small enough (such that $/N$ is the dominant term), the sample complexity of each individual agent, relative to a centralized system, is reduced by a factor of N. The third term has quadratic dependence on , and is zero when we have perfect synchronization, i.e. $K = 1$. The last term is proportional to ^{3}, and has the weakest dependency on the step size . For $K > 1$ we can merge the last two terms by upper bounding ^{3} ^{2}. The current upper bound, however, is tighter since with $K = 1$ (i.e. perfect synchronization) we have no term in the order ^{2}. Note that similar bounds (sans the last ^{3} term) have been established for the simpler i.i.d. noise case in the federated setting (Khaled et al., 2020; Koloskova et al., 2020). Consequently, we achieve the same sample complexity results for the more general federated setting with Markov noise.

Remark. The bound in Theorem 6.1 holds only after $T > \max\{K + ; 2\}$ and for all synchronization periods $K \geq 1$. At $K = 1$ the third term in the bound goes away, and we will be left only with the first order term, which is linearly decreasing with respect to the number of agents N, and the third order term $\tilde{O}(^3)$. The last term, however, is not tight and can be further improved to be of the order $O(^j); j > 3$. However, for that we need to assume larger , which means the bound only hold after a longer waiting time. In particular, by choosing $= dr \log e$, we can get $O(^{2r\ 1})$ for the last term (see the proof of Lemma B.2).

Acknowledgment

This work was partially supported by NSF awards CCF-1944993, CCF-2045694, CNS-2112471, CMMI-2112533, EPCN-2144316, and an award from Raytheon technologies.

References

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., Schneider, J., Tezak, N., Tworek, J., Welinder, P., Weng, L., Yuan, Q., Zaremba, W., and Zhang, L. Solving Rubik's cube with a robot hand. Preprint arXiv:1910.07113, 2019.

Baird, L. Residual algorithms: Reinforcement learning with function approximation. In Machine Learning Proceedings 1995, pp. 30–37. Elsevier, 1995.

Banach, S. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. Fund. math, 3(1):133–181, 1922.

Beck, A. First-order methods in optimization, volume 25. SIAM, 2017.

Beck, C. L. and Srikant, R. Error bounds for constant step-size Q-learning. Syst. Control. Lett., 61:1203–1208, 2012a.

Beck, C. L. and Srikant, R. Error bounds for constant step-size Q-learning. Systems & control letters, 61(12): 1203–1208, 2012b.

Beck, C. L. and Srikant, R. Improved upper bounds on the expected error in constant step-size Q-learning. In 2013 American Control Conference, pp. 1926–1931. IEEE, 2013.

Bertsekas, D. P. and Tsitsiklis, J. N. Neuro-dynamic programming. Athena Scientific, 1996a.

Bertsekas, D. P. and Tsitsiklis, J. N. Neuro-dynamic programming. Athena Scientific, 1996b.

Bertsekas, D. P., Bertsekas, D. P., Bertsekas, D. P., and Bertsekas, D. P. Dynamic programming and optimal control, volume 2. Athena scientific Belmont, MA, 1995.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In Conference on learning theory, pp. 1691–1692. PMLR, 2018.

Borkar, V. S. Stochastic approximation: a dynamical systems viewpoint, volume 48. Springer, 2009.

Borkar, V. S. and Meyn, S. P. The ODE method for convergence of stochastic approximation and reinforcement learning. SIAM Journal on Control and Optimization, 38 (2):447–469, 2000.

Chen, T. and Giannakis, G. B. Bandit convex optimization for scalable and dynamic iot management. IEEE Internet of Things Journal, 6(1):1276–1286, 2018.

Chen, T., Zhang, K., Giannakis, G. B., and Basar, T. Communication-efficient policy gradient methods for distributed reinforcement learning. IEEE Transactions on Control of Network Systems, 2021a.

Chen, Z., Zhang, S., Doan, T. T., Maguluri, S. T., and Clarke, J.-P. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. Under review by Automatica, Preprint arXiv:1905.11425, 2019a.

Chen, Z., Zhang, S., Doan, T. T., Maguluri, S. T., and Clarke, J.-P. Performance of Q-learning with linear function approximation: Stability and finite-time analysis. In OptRL Workshop at NeuRIPS 2019, 2019b.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. Finite-sample analysis of stochastic approximation using smooth convex envelopes. Under Review at Mathematics of Operations Research, Preprint arXiv:2002.00874, 2020a.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. Finite-sample analysis of stochastic approximation using smooth convex envelopes. In Advances in Neural Information Processing Systems, 2020b. URL https://proceedings.neurips.cc/paper/2020/file/5d44ee6f2c3f71b73125876103c8f6c4-Paper.pdf.

Chen, Z., Khodadadian, S., and Maguluri, S. T. Finite-Sample Analysis of Off-Policy Natural Actor-Critic with Linear Function Approximation. Preprint arXiv:2105.12540, 2021b. Submitted to NeurIPS 2021.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. A Lyapunov Theory for Finite-Sample Guarantees of Asynchronous Q-Learning and TD-Learning Variants. Under review by JMLR, Preprint arXiv:2102.01567, 2021c.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. Finite-Sample Analysis of Off-Policy TD-Learning via Generalized Bellman Operators. Preprint arXiv:2106.12729, 2021d.

Chen, Z., Zhou, Y., and Chen, R. Multi-agent off-policy td learning: Finite-time analysis with near-optimal sam-ple complexity and communication complexity. arXiv preprint arXiv:2103.13147, 2021e.

Chen, Z., Zhou, Y., Chen, R., and Zou, S. Sample and communication-efficient decentralized actor-critic algorithms with finite-time analysis. arXiv preprint arXiv:2109.03699, 2021f.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. Finite sample analysis for TD(0) with function approximation. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

Doan, T., Maguluri, S., and Romberg, J. Finite-Time Analysis of Distributed TD(0) with Linear Function Approximation on Multi-Agent Reinforcement Learning. In International Conference on Machine Learning, pp. 1626–1635, 2019.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., and Abbeel, P. Rl ^{2}: Fast reinforcement learning via slow reinforcement learning. arXiv preprint arXiv:1611.02779, 2016.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., Legg, S., and Kavukcuoglu, K. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. In International Conference on Machine Learning, pp. 1406–1415, 2018.

Even-Dar, E. and Mansour, Y. Learning Rates for Q-Learning. J. Mach. Learn. Res., 5:1–25, 2004. ISSN 1532-4435.

Foerster, J. N., Assael, Y. M., De Freitas, N., and Whiteson, S. Learning to communicate with deep multi-agent reinforcement learning. arXiv preprint arXiv:1605.06676, 2016.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. Guidelines for reinforcement learning in healthcare. Nature medicine, 25(1):16–18, 2019.

Gottesman, O., Futoma, J., Liu, Y., Parbhoo, S., Celi, L., Brunskill, E., and Doshi-Velez, F. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In International Conference on Machine Learning, pp. 3658–3667. PMLR, 2020.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In 2017 IEEE international conference on robotics and automation (ICRA), pp. 3389–3396. IEEE, 2017.

Hu, B. and Syed, U. A. Characterizing the exact behaviors of temporal difference learning algorithms using markov jump linear system theory. arXiv preprint arXiv:1906.06781, 2019.

Islam, R., Henderson, P., Gomrokchi, M., and Precup, D. Reproducibility of benchmarked deep reinforcement learning tasks for continuous control. arXiv preprint arXiv:1708.04133, 2017.

Jaakkola, T., Jordan, M. I., and Singh, S. P. Convergence of stochastic iterative dynamic programming algorithms. In Advances in neural information processing systems, pp. 703–710, 1994.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. Preprint arXiv:1912.04977, 2019.

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., and Levine, S. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. Preprint arXiv:1806.10293, 2018.

Khaled, A., Mishchenko, K., and Richtárik, P. Tighter theory for local sgd on identical and heterogeneous data. In International Conference on Artificial Intelligence and Statistics, pp. 4519–4529. PMLR, 2020.

Khodadadian, S., Chen, Z., and Maguluri, S. T. Finite-Sample Analysis of Off-Policy Natural Actor-Critic Algorithm. In International Conference on Machine Learning, 2021.

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

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. A unified theory of decentralized sgd with changing topology and local updates. In International Conference on Machine Learning, pp. 5381–5393. PMLR, 2020.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. In Advances in neural information processing systems, pp. 1008–1014, 2000.

Konečnỳ, J., McMahan, H. B., Ramage, D., and Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610.02527, 2016.

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

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. Preprint arXiv:2005.01643, 2020.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Sample Complexity of Asynchronous Q-Learning: Sharper Analysis and Variance Reduction. Advances in neural information processing systems, 2020.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. arXiv preprint arXiv:1907.02189, 2019.

Liang, X., Liu, Y., Chen, T., Liu, M., and Yang, Q. Federated transfer reinforcement learning for autonomous driving. arXiv preprint arXiv:1910.06001, 2019.

Lim, H.-K., Kim, J.-B., Heo, J.-S., and Han, Y.-H. Federated reinforcement learning for training control policies on multiple iot devices. Sensors, 20(5):1359, 2020.

Liu, B., Liu, J., Ghavamzadeh, M., Mahadevan, S., and Petrik, M. Finite-sample analysis of proximal gradient TD algorithms. In Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, pp. 504–513, 2015.

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

Liu, R. and Olshevsky, A. Distributed td (0) with almost no communication. arXiv preprint arXiv:2104.07855, 2021.

Liu, Y., Gottesman, O., Raghu, A., Komorowski, M., Faisal, A. A., Doshi-Velez, F., and Brunskill, E. Representation Balancing MDPs for Off-policy Policy Evaluation. Advances in Neural Information Processing Systems, 31: 2644–2653, 2018.

Maddern, W., Pascoe, G., Linegar, C., and Newman, P. 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research, 36(1):3–15, 2017.

Maei, H. R. Convergent actor-critic algorithms under off-policy training and function approximation. Preprint arXiv:1802.07842, 2018.

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

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In International conference on machine learning, pp. 1928–1937, 2016.

Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., and Srivastava, G. A survey on security and privacy of federated learning. Future Generation Computer Systems, 115:619–640, 2021.

Nadiger, C., Kumar, A., and Abdelhak, S. Federated reinforcement learning for fast personalization. In 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), pp. 123–127. IEEE, 2019.

Nair, A., Srinivasan, P., Blackwell, S., Alcicek, C., Fearon, R., De Maria, A., Panneershelvam, V., Suleyman, M., Beattie, C., Petersen, S., et al. Massively parallel methods for deep reinforcement learning. arXiv preprint arXiv:1507.04296, 2015.

Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., and Poor, H. V. Federated learning for internet of things: A comprehensive survey. arXiv preprint arXiv:2104.07914, 2021.

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

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

Qu, G. and Wierman, A. Finite-Time Analysis of Asynchronous Stochastic Approximation and Q-Learning. In Conference on Learning Theory, pp. 3185–3205. PMLR, 2020.

Qu, Z., Lin, K., Kalagnanam, J., Li, Z., Zhou, J., and Zhou, Z. Federated learning's blessing: Fedavg has linear speedup. arXiv preprint arXiv:2007.05690, 2020.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In Proceedings of the 29th International Coference on International Conference on Machine Learning, pp. 1571–1578, 2012.

Recht, B., Re, C., Wright, S., and Niu, F. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q. (eds.), Advances in Neural Information Processing Systems, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/218a0aefd1d1a4be65601cc6ddc1520e-Paper.pdf.

Ren, J., Wang, H., Hou, T., Zheng, S., and Tang, C. Federated learning-based computation offloading optimization in edge computing-supported internet of things. IEEE Access, 7:69194–69201, 2019.

Shah, D. and Xie, Q. Q-learning with nearest neighbors. In Advances in Neural Information Processing Systems, pp. 3111–3121, 2018.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. Safe, multi-agent, reinforcement learning for autonomous driving. Preprint arXiv:1610.03295, 2016.

Shalev-Shwartz, S. et al. Online learning and online convex optimization. Foundations and Trends® in Machine Learning, 4(2):107–194, 2012.

Shao, K., Tang, Z., Zhu, Y., Li, N., and Zhao, D. A survey of deep reinforcement learning in video games. arXiv preprint arXiv:1912.10944, 2019.

Shen, H., Zhang, K., Hong, M., and Chen, T. Asynchronous advantage actor critic: Non-asymptotic analysis and linear speedup. arXiv preprint arXiv:2012.15511, 2020.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587):484, 2016.

Spiridonoff, A., Olshevsky, A., and Paschalidis, I. C. Communication-efficient sgd: From local sgd to one-shot averaging. In Advances in Neural Information Processing Systems, volume 34, 2021.

Srikant, R. and Ying, L. Finite-time error bounds for linear stochastic approximation and TD learning. In Conference on Learning Theory, pp. 2803–2830. PMLR, 2019.

Stich, S. U. Local sgd converges fast and communicates little. In International Conference on Learning Representations, 2018.

Sun, J., Wang, G., Giannakis, G. B., Yang, Q., and Yang, Z. Finite-time analysis of decentralized temporal-difference learning with linear function approximation. In International Conference on Artificial Intelligence and Statistics, pp. 4485–4495. PMLR, 2020.

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

Sutton, R. S. and Barto, A. G. Reinforcement learning: An introduction. MIT press, 2018.

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

Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., and Zhou, Y. A hybrid approach to privacy-preserving federated learning. In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, pp. 1–11, 2019.

Tsitsiklis, J. N. Asynchronous stochastic approximation and Q-learning. Machine learning, 16(3):185–202, 1994.

Tsitsiklis, J. N. and Van Roy, B. Analysis of temporal-difference learning with function approximation. In Advances in neural information processing systems, pp. 1075–1081, 1997.

Wai, H.-T. On the convergence of consensus algorithms with markovian noise and gradient bias. In 2020 59th IEEE Conference on Decision and Control (CDC), pp. 4897–4902. IEEE, 2020.

Wainwright, M. J. Stochastic approximation with cone-contractive operators: Sharp ℓ_1-bounds for Q-learning. Preprint arXiv:1905.06265, 2019.

Wang, G., Lu, S., Giannakis, G., Tesauro, G., and Sun, J. Decentralized TD Tracking with Linear Function Approximation and its Finite-Time Analysis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), Advances in Neural Information Processing Systems, volume 33, pp. 13762–13772. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/file/9ec51f6eb240fb631a35864e13737bca-Paper.pdf.

Wang, J. and Joshi, G. Cooperative sgd: A unified framework for the design and analysis of local-update sgd algorithms. Journal of Machine Learning Research, 22(213): 1–50, 2021.

Wang, X., Wang, C., Li, X., Leung, V. C., and Taleb, T. Federated deep reinforcement learning for internet of things with decentralized cooperative edge caching. IEEE Internet of Things Journal, 7(10):9441–9455, 2020b.

Watkins, C. J. and Dayan, P. Q-learning. Machine learning, 8(3-4):279–292, 1992.

Xu, M., Peng, J., Gupta, B., Kang, J., Xiong, Z., Li, Z., and Abd El-Latif, A. A. Multi-agent federated reinforcement learning for secure incentive mechanism in intelligent cyber-physical systems. IEEE Internet of Things Journal, 2021.

Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T., and Yu, H. Federated learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 13(3):1–207, 2019.

Yun, W. J., Park, S., Kim, J., Shin, M., Jung, S., Mohaisen, A., and Kim, J.-H. Cooperative multi-agent deep reinforcement learning for reliable surveillance via autonomous multi-uav control. IEEE Transactions on Industrial Informatics, 2022.

Yurtsever, E., Lambert, J., Carballo, A., and Takeda, K. A survey of autonomous driving: Common practices and emerging technologies. IEEE Access, 8:58443–58469, 2020.

Zeng, S., Doan, T. T., and Romberg, J. Finite-time analysis of decentralized stochastic approximation with applications in multi-agent and multi-task learning. arXiv preprint arXiv:2010.15088, 2020.

Zhang, S., Liu, B., Yao, H., and Whiteson, S. Provably convergent two-timescale off-policy actor-critic with function approximation. In International Conference on Machine Learning, pp. 11204–11213. PMLR, 2020.

Zhang, S. Q., Lin, J., and Zhang, Q. A multi-agent reinforcement learning approach for efficient client selection in federated learning. arXiv preprint arXiv:2201.02932, 2022.

Zhao, L., Ran, Y., Wang, H., Wang, J., and Luo, J. Towards cooperative caching for vehicular networks with multi-level federated reinforcement learning. In ICC 2021-IEEE International Conference on Communications, pp. 1–6. IEEE, 2021.

Zhuo, H. H., Feng, W., Lin, Y., Xu, Q., and Yang, Q. Federated deep reinforcement learning. arXiv preprint arXiv:1901.08277, 2019.

Appendices

The appendices are organized as follows. In section A we discuss the lower bound on the convergence of general stochastic approximation. In Section B we derive the convergence bound of FedSAM algorithm. Next, we employ the results in Section B to derive the convergence bounds of federated TD-learning in Section C and federated Q-learning in Section D.

A. Lower Bound on the Convergence of General Stochastic Approximation

In this section we discuss the convergence of general stochastic approximation. In this discussion we provide a simple stochastic approximation with iid noise which can give insight into the general convergence bound in (3). In particular, we show that the convergence bound in (3) is tight and cannot be improved.

Consider a one dimensional random variable X with zero mean $E[X] = 0$ and bounded variance $E[X^2] = \sigma^2$. Consider the following update

$$x_{t+1} = x_t + \alpha(X_t - x_t); \quad t \geq 0; \tag{9}$$

where we start with some fixed deterministic x_0 and X_t is a an iid sample of the random variable X. It is easy to see that the update (9) is a special case of the update of the general stochastic approximation with the fixed point $x = 0$.

By expanding the update (9), we have

$$x_t = (1-\alpha)^t x_0 + \sum_{k=0}^{t-1} \alpha(1-\alpha)^{t-k-1} X_k.$$

Hence, we have

$$x_t^2 = (1-\alpha)^{2t} x_0^2 + \left(\sum_{k=0}^{t-1} \alpha(1-\alpha)^{t-k-1} X_k\right)^2 + 2(1-\alpha)^t x_0 \sum_{k=0}^{t-1} \alpha(1-\alpha)^{t-k-1} X_k.$$

Taking expectation on both sides, and using the zero mean property of X_k, we have

$$\begin{aligned}
E[x_t^2] &= (1-\alpha)^{2t} x_0^2 + \alpha^2 E\left[\left(\sum_{k=0}^{t-1} (1-\alpha)^{t-k-1} X_k\right)^2\right] \\
&= (1-\alpha)^{2t} x_0^2 + \alpha^2 E\left[\sum_{k=0}^{t-1} (1-\alpha)^{2(t-k-1)} X_k^2 + \sum_{k,k'=0; k=k'}^{t-1} (1-\alpha)^{2t-k-k'-2} X_k X_{k'}\right] \\
&= (1-\alpha)^{2t} x_0^2 + \alpha^2 \left[\sum_{k=0}^{t-1} (1-\alpha)^{2(t-k-1)} + \sum_{k,k'=0; k=k'}^{t-1} (1-\alpha)^{2t-k-k'-2} E[X_k]E[X_{k'}]\right] && \text{(iid property)} \\
&= (1-\alpha)^{2t} x_0^2 + \alpha^2 \sum_{k=0}^{t-1} (1-\alpha)^{2(t-k-1)} && \text{(zero mean)} \\
&= (1-\alpha)^{2t} x_0^2 + \alpha^2 \sigma^2 \frac{1-(1-\alpha)^{2t}}{1-(1-\alpha)^2} \\
&= x_0^2 (1-\alpha)^{2t} + \alpha\sigma^2 \frac{1-(1-\alpha)^{2t}}{2-\alpha} \\
&= \underbrace{\left(x_0^2 - \frac{\alpha\sigma^2}{2-\alpha}\right)(1-\alpha)^{2t}}_{T_1:\text{bias}} + \underbrace{\frac{\alpha\sigma^2}{2-\alpha}}_{T_2:\text{variance}}
\end{aligned} \tag{10}$$

It is clear that (10) has the same form as the bound in (3) with T_1 as the geometric term which converges to zero as $t \to \infty$, and T_2 term proportional to the step size α. In addition, note that in the above derivation, we did not use any inequality, and hence the bounds in (10) as well as (3) are tight.

B. Analysis of Federated Stochastic Approximation

First, we restate Theorem 6.1 with explicit expressions of the different constants.

Theorem B.1. Consider the FedSAM Algorithm 4 with $c_{FSAM} = 1 - \frac{\alpha_2}{2}$ (α_2 is defined in (14)), and suppose Assumptions 6.1, 6.2, 6.3, 6.4 are satisfied. Consider small enough step size α which satisfies the assumptions in (21), (23), (32), (35). Furthermore, denote $\tau = \lceil 2 \log e \rceil$, and take large enough T such that $T > \max\{K + \tau, 2\tau\}$. Then, the output of the FedSAM Algorithm 4, $\hat{\theta}_T$, $\frac{1}{N}\sum_{i=1}^{N} \theta_T^i$, satisfies

$$\mathbb{E}[\|\hat{\theta}_T - \theta_c\|^2] \le C_1 \alpha^{\tau+1} \left(1 - \frac{\alpha_2}{2}\right)^{T-\tau} + C_2 \frac{\alpha}{N} + C_3 (K-\tau-1)^2 \alpha^4 + C_4 \alpha^3, \tag{11}$$

where $C_1 = 16 u_{cm}^2 M_0 (\log \frac{1}{\alpha} + \frac{1}{\alpha_2})$, $M_0 = \frac{1}{l_{cm}^2}\frac{1}{C_1^2} B + (A_2 + 1)\left(\|\theta_0 - \theta_c\| + \frac{B}{2C_1}\right)^2 + \|\theta_0 - \theta_c\|^2$, $C_2 = 8 u_{cm} C_8 \tau^2 + C_{12}/\alpha_2$, and $C_3 = 80 B^2 C_{17} u_{cm} \left(1 + \frac{2B(1-\alpha)}{\alpha_2 \alpha \tau}\right)$ and $C_4 = 8 u_{cm} C_7 \tau^2 + C_{11} + 0.5 C_3 C_9 \tau^2 + C_3 C_{10} + 2 C_1 C_3 C_{10} + 3 A_1 C_3 + C_{13}/\alpha_2$. Here the constants $u_{cm}, l_{cm}, \alpha_2, C_1, C_3, C_7, C_8, C_9, C_{10}, C_{11}, C_{12}, C_{13}, C_{17}, B$ are problem dependent constants which are defined in the following proposition and lemmas.

Next, we characterize the sample complexity of the FedSAM Algorithm 4, where we establish a linear speedup in the convergence of the algorithm.

Corollary B.1.1. Consider FedSAM Algorithm 4 with fixed number of iterations T and step size $\alpha = \frac{8 \log(NT)}{\alpha_2 T}$. Suppose T is large enough, such that α satisfies the requirements of the step size in Theorem B.1 and $T > 4$. Also take $K = T/N$. Then we have $\mathbb{E}[\|\theta_T - \theta_c\|^2]$ after $T = O\left(\frac{1}{N\epsilon}\right)$ iterations.

Corollary B.1.1 establishes the sample and communication complexity of FedSAM Algorithm 4. The $O(1/(N\epsilon))$ sample complexity shows the linear speedup with increasing N. Another aspect of the cost is the number of communications required between the agents and the central server. According to Corollary B.1.1, we need $T/N = \tilde{O}(N)$ rounds of communications in order to reach an ϵ-optimal solution. Hence, even in the presence of Markov noise, the required number of communications is independent of the desired final accuracy ϵ, and grows linearly with the number of agents. Our result generalizes the existing result achieved for the simpler i.i.d. noise case in (Khaled et al., 2020; Spiridonoff et al., 2021).

In the following sections, we discuss the proof of Theorem B.1. In Section B.1, we introduce some notations and preliminary results to facilitate our Lyapunov-function based analysis. Next, in Section B.2, we state some primary propositions, which are then used to prove Theorem B.1. In Sections B.3,B.4,B.5, we prove the aforementioned propositions. Along the way, we state several intermediate results, which are stated and proved in Sections B.6 and B.7, respectively.

Throughout the appendix we have several sets of constants. The constants $C_i, i = 1, \ldots, 17$ are problem dependent constants which we define recursively. The final constants which appears in the resulting bound in Theorem B.1 are shown as $C_i, i = 1, \ldots, 4$. Finally, the constant c is used in the sampling of the time step T.

B.1. Preliminaries

We define the following notations:

- θ_t, $\frac{1}{N}\sum_{i=1}^{N} \theta_t^i$: virtual sequence of average (across agents) parameter.
- $\boldsymbol{\theta}_t = [\theta_t^1, \ldots, \theta_t^N]$: set of local parameters at individual nodes.
- $Y_t = [y_t^1, \ldots, y_t^N]$: Markov chains at individual nodes.
- π^i: the stationary distribution of y_t^i as $t \to \infty$.
- $G(\boldsymbol{\theta}_t, Y_t)$, $\frac{1}{N}\sum_{i=1}^{N} G^i(\theta_t^i, y_t^i)$: average of the noisy local operators at the individual local parameters.
- $\bar{G}(\theta_t, Y_t)$, $\frac{1}{N}\sum_{i=1}^{N} G^i(\theta_t, y_t^i)$: average of the noisy local operators at the average parameter.
- $b(Y_t)$, $\frac{1}{N}\sum_{i=1}^{N} b^i(y_t^i)$: average of Markovian noise.

- $G(\theta_t) \triangleq \frac{1}{N}\sum_{i=1}^{N} G^i(\theta_t)$: average of expected operators evaluated at the local parameter.
- $\bar{G}(\theta_t) \triangleq \frac{1}{N}\sum_{i=1}^{N} G^i(\theta_t)$: average of expected operator evaluated at the average parameter.
- $\theta^* = 0$ is the unique fixed point satisfying $\theta^* = G^i(\theta^*)$ for all $i = 1, \dots, N$. Note that this follows directly from the assumptions. In particular, by Assumption 6.2, $G^i(\cdot)$ is a contraction, and hence by Banach fixed point theorem (Banach, 1922), there exist a unique fixed point of this operator. Furthermore, by Assumption 6.3, we have

$$\|G^i(0)\|_c = \|\mathbb{E}_{y^i}[G^i(0; y)]\|_c \le \mathbb{E}_{y^i}\|G^i(0; y)\|_c \le A_2 \mathbb{E}_{y^i}[\|0\|_c] = 0.$$

  Hence $G^i(0) = 0$ and 0 is the unique fixed point of the operator $G^i(\cdot)$.
- $\theta_t^i \triangleq \|\theta_t^i - \theta^*\|_c$; $\theta_t \triangleq \frac{1}{N}\sum_{i=1}^{N} \theta_t^i$; $\bar{\theta}_t \triangleq \frac{1}{N}\sum_{i=1}^{N} \theta_t^i$; $\Theta_t \triangleq \sum_{i=1}^{N} (\theta_t^i)^2$: measures of synchronization error.

Throughout this proof we assume $\|\cdot\| \triangleq \|\cdot\|_c$ as some given norm. $\mathbb{E}_t[\cdot] \triangleq \mathbb{E}[\cdot|\mathcal{F}_t]$, where $\mathcal{F}_t$ is the sigma-algebra generated by $\{r_r^i, g_r^i\}_{i=1,\dots,N;\, r=1,\dots,t}$. Unless specified otherwise, $\|\cdot\|$ denotes the Euclidean norm.

Generalized Moreau Envelope: Consider the norm $\|\cdot\|_c$ which appears in Assumptions 6.1-6.3. Square of this norm need not be smooth. Inspired by (Chen et al., 2020a), we use the Generalized Moreau Envelope as a Lyapunov function for the analysis of the convergence of Algorithm 4. The Generalized Moreau Envelope of $f(\cdot)$ with respect to $g(\cdot)$, for $\lambda > 0$, is defined as

$$M_{f,\lambda}^{g}(x) = \min_{u \in \mathbb{R}^d} \; f(u) + \frac{1}{\lambda} g(x - u) \; : \tag{12}$$

Let $f(x) = \frac{1}{2}\|x\|_c^2$ and $g(x) = \frac{1}{2}\|x\|_s^2$, which is L-smooth with respect to $\|\cdot\|_s$ norm. For this choice of f, g, $M_{f,\lambda}^{g}(\cdot)$ is essentially a smooth approximation to f, which is henceforth denoted with the simpler notation $M_\lambda(\cdot)$. Also, due to the equivalence of norms, there exist $l_{cs}, u_{cs} > 0$ such that

$$l_{cs}\|\cdot\|_s \le \|\cdot\|_c \le u_{cs}\|\cdot\|_s. \tag{13}$$

We next summarize the properties of $M_\lambda(\cdot)$ in the following proposition, which were established in (Chen et al., 2020a).

Proposition B.1 ((Chen et al., 2020b)). The function $M_\lambda(\cdot)$ satisfies the following properties.

(1) $M_\lambda(\cdot)$ is convex, and $\frac{L}{\lambda}$-smooth with respect to $\|\cdot\|_s$. That is, $M_\lambda(y) \le M_\lambda(x) + \langle \nabla M_\lambda(x), y - x\rangle + \frac{L}{2\lambda}\|x - y\|_s^2$ for all $x, y \in \mathbb{R}^d$.

(2) There exists a norm, denoted by $\|\cdot\|_m$, such that $M_\lambda(x) = \frac{1}{2}\|x\|_m^2$.

(3) Let $\gamma_{cm} = (1 + \gamma_{cs}^2)^{1/2}$ and $u_{cm} = (1 + u_{cs}^2)^{1/2}$. Then it holds that $\gamma_{cm}\|\cdot\|_m \le \|\cdot\|_c \le u_{cm}\|\cdot\|_m$.

By Proposition B.1, we can use $M_\lambda(\cdot)$ as a smooth surrogate for $\frac{1}{2}\|\cdot\|_c^2$. Furthermore, we denote

$$\rho_1 = \frac{1 + \frac{u_{cs}^2}{\gamma_{cs}^2}}{1 + \frac{\lambda L}{\gamma_{cs}^2}}; \quad \rho_2 = 1 - \frac{\lambda}{u_{cs}^2}; \quad \text{and} \quad \rho_3 = \frac{114\lambda^2 L(1 + u_{cs}^2)}{\gamma_{cs}^2}. \tag{14}$$

Note that by choosing $\lambda > 0$ small enough, we can ensure $\rho_2 \in (0, 1)$.

B.2. Proof of Theorem B.1

In this section, first we state three key results (Propositions B.2, B.3, B.4). These are then used to prove Theorem B.1. The first step of the proof is to characterize the one-step drift of the Lyapunov function $M_\lambda(\cdot)$, with the parameters generated by the FedSAM Algorithm 4, which is formally stated in the following proposition.

Proposition B.2 (One-step drift - I). Consider the update of the FedSAM Algorithm 4. Suppose the assumptions 6.1, 6.2, 6.3, and 6.4 are satisfied. Consider $\lambda = \alpha^2 \log \alpha$ and $t \ge 2$, we have

$$\mathbb{E}_{t-2}[M_\lambda(\theta_{t+1})]$$

$$1 \quad 2'_2 \quad {}_1 \quad {}_2 \quad \frac{2LA_2}{2} \quad {}_2 \frac{L(A_1+1)}{I^2_{cs}} \quad {}_3 \frac{{}_2 + {}_2 6Lu^2_m}{I^2_{cs}} \quad {}_4 m_1 + \frac{(A_2+1)^2}{I^2_{cs}} E_{t\ 2}[M(t)]_{cs} \quad {}_{cs} \quad {}_{cs}$$

$$+ {}^2 \underbrace{\frac{m_2 L}{I^2_{cs}}}_{C_3} E_{t\ 2}[k_{t\ t}\ k_c] \tag{15}$$

$$+ \underbrace{\frac{L}{I^2_{cs}} \quad \frac{L}{2\ I^2_{cs}} + A_2 \quad 1 + \frac{u^2_{cm}}{2^2} + (A_1+1) \quad \frac{3u^2_{cm}}{2\ 3^2} + 2 + 3m_1{}^2}_{C_4} \quad E_{t\ 2}[k_{t\ t}\ k^2]_c \tag{16}$$

$$+ \underbrace{\frac{L - \frac{m}{2} + 3u^2_{cm}(A_1+31) + A_2 + \frac{m}{2I} + \frac{3}{2}\frac{2}{2}}{t]} \quad \frac{Lu^2_{cs}}{2^2_{cs}} E_{t\ 2}[A^2_1}_{C_5} \tag{17}$$

$$+ \underbrace{\frac{L}{2I^2_{cs}} \quad \frac{3u^{pm}}{(18)} + 3 \quad (A_1+1) + m_1 \quad E_{t\ 2}[}_{C_6} \tag{18}$$

$$+ \underbrace{\frac{m^2_2 u^2_{cm} L^2}{2_2 I^4_{cs}}}_{C_7} \quad {}^4 \tag{19}$$

$$+ \underbrace{\frac{1}{2} \quad 1 + \quad {}^2 3L^2}_{C_8} E_{t\ 2}\, kb(Y_t)k^2; \tag{20}$$

where $_1$, $_2$, $_3$, and $_4$ are arbitrary positive constants.

Proof. The proof of Proposition B.2 is presented in full detail in Section B.3. □

Before discussing the bound in its full generality, we discuss a few special cases.

- Perfect synchronization ($K = 1$) with i.i.d. noise: since $_t = 0$ for all t, $= 0$ (independence across time), and $C_7 = 0$ (see Lemmas B.2 and B.5 in Section B.6), the terms (15), (16) (17),(18), (19) will not appear in the bound, which is the form we get for centralized systems with i.i.d. noise (Rakhlin et al., 2012).
- Infrequent synchronization ($K > 1$) with i.i.d. noise: $= 0$, and $C_7 = 0$, the terms (15), (16),(18), (19) will not appear in the bound, which is the form we get for federated stochastic optimization with i.i.d. noise (Khaled et al., 2020).
- Perfect synchronization ($K = 1$) with Markov noise: since $_t = 0$ for all t, the bound in Proposition B.2 generalizes the results in (Srikant & Ying, 2019; Chen et al., 2021c).

Next, we substitute the bound on $k_{t\ t}\ k_c$ (Lemma B.6) and $k_{t\ t}\ k^2$ (Lemma B.7) to further bound the one-step drift. Establishing a tight bound for these two quantities are essential to ensure linear speedup.

Proposition B.3 (One-step drift - II). Consider the update of the FedSAM Algorithm 4. Suppose Assumptions 6.1, 6.2, 6.3, and 6.4 are satisfied. Define $C_{15}() = \frac{6m_1 Lu^2_{cm}}{I^2_{cs}} + \frac{6L(A_2+1)^2 u^2_m}{2\ I^2_{cs}} + 144(A^2_2 + 1)C\ u^2_{cm}{}^2$. For

$$\min \quad \frac{1}{360(A_2 + 1)^2 C_{15}()}; \tag{21}$$

$= d2 \log e$, and $t\ 2$, we have

$$E_{t\ 2}[M(_{t+1})] \quad (1 \quad '_2)\, E_{t\ 2}[M(_t)] + C_{14}()^4 + C_{16}()^2 + C_{17} \quad \frac{X^t}{N} E_{t\ 2}[\quad {}_k]; \ k = t$$

where $C_{14}() = C_7 + C_{11} + 0{:}5C_3^2C_9^2 + C_3C_{10} + 2C_1C_3C_{10} + 3A_1C_3 + C_{13} + C_8\frac{u_{cD}^2}{l_{cD}^2}2m_2^{22}{}^2$, $C_{16}() = C_8\frac{u_{cD}^2B^2}{l_{cD}^2} + \frac{1}{2} + C_{12}{}^2$ and $C_{17} = (3A_1C_3 + 8A^2C_4 + C_5 + C_6)$. Here we define $C_9 = \frac{8}{l_{cD}}\frac{u_{cB}}{}$, $C_{10} = \frac{8m_2u_{cD}}{l_{cD}(1 \quad)}$, $C_{11} = \frac{8C_1^2C_3^2u_{cm}^2}{\frac{2}{6}}$, $C_{12} = \frac{8C_4u_{cD}^2B^2}{l_{cD}^2}$, $C_{13} = \frac{14C_4u_{cD}^2m_2^2}{l_{cD}^2(1 \quad)}$.

Proof. The proof of Proposition B.3 is presented in full detail in Section B.4. □

Conditional expectation $E_{t\ 2}[]$. The conditional expectation $E_{t\ 2}[]$ used in Proposition B.3 is essential when dealing with Markovian noise. The idea of using conditional expectation to deal with Markovian noise is not novel per se. In the previous work (Bhandari et al., 2018; Srikant & Ying, 2019; Chen et al., 2021c), conditioning on t is sufficient to establish the convergence results. Due to the mixing property (Assumption 6.1), the Markov chain geometrically converges to its stationary distribution. Therefore, choosing "large enough" , and conditioning on t , one can ensure that the Markov chain at time t is "almost in steady state." However, in federated setting, conditioning on t results in bounds that are too loose. In particular, consider the differences $k_t \quad _t k_c$ and $k_t \quad _t k^2$ in (15) and (16) respectively. In the centralized setting, as in (Bhandari et al., 2018; Srikant & Ying, 2019; Chen et al., 2021c), these terms can be bounded deterministically to yield ^{2} bound. However, in the federated setting, this crude bound does not result in linear speedup in N. In this work, to achieve a finer bound on $k_t \quad _t k_c$, we go steps further back in time. This ensures that the difference behaves almost like the difference of average of i.i.d. random variables, resulting in a tighter bound (see Lemma B.6). By exploiting the conditional expectation $E_{t\ 2}[]$, we derive a refined analysis to bound this term as $O(^2=N + {}^4)$, which guarantees a linear speedup (see Lemmas B.6 and B.7).

Taking total expectation in Proposition B.3 (using tower property), we get

$$E[M(_{t+1})] \quad (1 \quad '_2)\,E[M(_t)] + C_{14}()^4 + C_{16}()^{\ 2} + C_{17}\sum_{k=t}^{t}\frac{}{N}\,E[\quad _k] \tag{22}$$

To understand the bound in Proposition B.3, consider the case of $K = 1$ (i.e. full synchronization). In this case we have $_i = 0$ for all i, and the bound in Proposition B.3 simplifies to $E[M(_{t+1})] \quad (1 \quad '_2)\,E[M(_t)] + C_{14}()^3 + C_{16}()^{\ 2}\frac{}{N}$. This recursion is sufficient to achieve linear speedup. However, the bound in Proposition B.3 also include terms which are proportional to the error due to synchronization. In order to ensure convergence along with linear speedup, we need to further upper bound this term with terms which are of the order $O(^3(K \quad 1))$ and $M(_i)$. The following lemma is the next important contribution of the paper where we establish such a bound for weighted sum of the synchronization error. Notice that the weights fw_tg are carefully chosen to ensure the best rate of convergence for the overall algorithm.

Proposition B.4 (Synchronization Error). Suppose $T > K + .$ For such that

$$\begin{aligned} {}^2 \quad & \min\ \frac{1}{\quad _2};\ \frac{\ln(5=4)}{2(1 + \tilde{A}_1)(K \quad 1)^2};\\ & {}^2(\log() + 1)\ _{2K^2};\ \frac{v}{\quad}\\ & {}^2(\log() + 1)^3\exp\ '_2\ \ 2\log + 1\ _{4K}\exp\ \ '_{\frac{v}{2}}\ \sqrt{\frac{\ln(5=4)}{2\quad)\ (1 + \tilde{A}_1}}\ ! \end{aligned} \tag{23}$$

where $v = \frac{'_2}{80\tilde{A}_2C_{17}u_{cm}^2}$, the weighted consensus error satisfies

$$\begin{aligned} \frac{2C_{17}}{2W_T}\sum_{t=2}^{T}w_t\ \sum_{'=t}^{t}E\ \ \Big\| & \sim \frac{10C\quad}{\quad ^2B^2\ '_2}\ \frac{4m_2}{1 + B(1\quad)}\ \frac{1}{\quad ^{17}}\ (\ + 1)(K \quad 1) + \frac{1}{2W_T}\sum_{t=0}^{T}w_tEM(_t): \end{aligned} \tag{24}$$

Proof. The proof is presented in Section B.5. □

Finally, by incorporating the results in Propositions B.2, B.3, and B.4, we can establish the convergence of FedSAM in Theorem B.1.

Proof of Theorem B.1. Assume $w_0 = 1$, and consider the weights w_t generated by the recursion $w_t = w_{t-1}(1-\frac{\alpha^2}{2})^{-1}$. Multiplying both sides of (22) by $\frac{2w_t}{\alpha}$ and rearranging the terms, we get

$$\begin{aligned} w_t\mathbb{E}\|M(t)\| \le & \frac{2w_t}{\alpha}\Big(\frac{1}{\alpha}-\frac{\alpha}{2}\Big)\mathbb{E}\|M(t)\| - \frac{2w_t}{\alpha}\mathbb{E}\|M(t+1)\| + \frac{2w_t}{\alpha}C_{14}(\alpha)^4 + C_{16}(\alpha)\Big)N\alpha^2 + \frac{2w_tC_{17}}{\alpha^2}\sum_{i=t}^{t+\tau}\mathbb{E}[\\ = & \frac{2}{\alpha}[w_{t-1}\mathbb{E}\|M(t)\| - w_t\mathbb{E}\|M(t+1)\|] + \alpha^2\big(C_{14}(\alpha)^3 + C_{16}(\alpha)\big)N + \alpha^2\frac{2C_{17}}{\alpha^2}\sum_{i=t}^{t+\tau}w_t\|\\ & \|]; \end{aligned} \tag{25}$$

where we use $w_{t-1} = w_t(1-\frac{\alpha^2}{2})$. Summing (25) over $t = 2$ to T (define $W_T = \sum_{t=2}^{T} w_t$), we get

$$\begin{aligned} & \frac{1}{W_T}\sum_{t=2}^{T} w_t\mathbb{E}\|M(t)\| \\ & \le \frac{2}{\alpha W_T}[w_1\mathbb{E}\|M(2)\| - w_T\mathbb{E}\|M(T+1)\|] + \alpha^2\big(C_{14}(\alpha)^3 + C_{16}(\alpha)\big)N - \frac{1}{W_T}\sum_{t=2}^{T} w_t \\ & + \frac{1}{W_T}\sum_{t=2}^{T}\frac{2w_tC_{17}}{\alpha^2}\sum_{i=t}^{t+\tau}\\ & \frac{2w_1}{\alpha}\mathbb{E}\|M(2)\| + \frac{2}{\alpha}\sum_{t=2}^{T}\frac{2C_{17}}{\alpha^2}\sum_{i=t}^{t+\tau}w_t + \alpha^2\big(C_{14}(\alpha)^3 + C_{16}(\alpha)\big)N + \alpha^2\frac{1}{W_T}\sum_{t=2}^{T} w_t\sum_{i=t}^{t+\tau}\\ & \frac{2}{\alpha}\sum_{t=2}^{T}(\|M(t)\| - \|M(t+1)\|)w_t = \frac{2}{\alpha}\Big(1-\frac{\alpha^2}{2}\Big)^{T-1}\mathbb{E}\|M(2)\| + \alpha^2\big(C_{14}(\alpha)^3 + C_{16}(\alpha)\big)N + \alpha^2\frac{2C_{17}}{\alpha^2}\sum_{t=2}^{T}w_t\sum_{i=t}^{t+\tau}\\ & . \end{aligned} \tag{26}$$

Substituting the bound on $\frac{1}{W_T}\sum_{t=2}^{T} w_t\sum_{i=t}^{t+\tau}$ from Proposition B.4 into (26), we get

$$\begin{aligned} \frac{1}{W_T}\sum_{t=2}^{T} w_t\mathbb{E}\|M(t)\| \le & \frac{1}{\alpha}\Big(1-\frac{\alpha^2}{2}\Big)^{T-1}\mathbb{E}\|M(2)\| + \alpha^2\big(C_{14}(\alpha)^3 + C_{16}(\alpha)\big)N \\ & + \alpha^2B^2\Big(1+\frac{4m_2}{B(1-\rho)}\Big)(\tau+1)(K-1) + \frac{1}{2W_T}\sum_{t=0}^{T} w_t\mathbb{E}\|M(t)\|\Big) \\ \frac{1}{W_T}\sum_{t=2}^{T} w_t\mathbb{E}[\|M(t)\|] \le & \Big(1-\frac{\alpha^2}{4}\Big)^{T-1}\|M_0\| + \alpha^2\big(C_{14}(\alpha)^3 + C_{16}(\alpha)\big)N \\ & + \alpha^2B^2\Big(1+\frac{4m_2}{B(1-\rho)}\Big)(\tau+1)(K-1) + 2\Big(1-\frac{\alpha^2}{2}\Big)^{T-1}\|M_0\|; \end{aligned} \tag{27}$$

where M_0 is a problem dependent constant and is defined in Lemma B.6. To simplify (27), we define $C_{18}(\alpha) = B^2\big(1+\frac{4m_2}{B(1-\rho)}\big)$, $C_{19}(\alpha) = \alpha^2 C_{16}(\alpha)$. We have

$$\frac{1}{W_T}\sum_{t=2}^{T} w_t\mathbb{E}\|M(t)\| \le 4\Big(1-\frac{\alpha^2}{2}\Big)^{T-1}\|M_0\| + \frac{4C_{14}(\alpha)^3}{\alpha} + 2\alpha C_{18}(\tau+1)(K-1) + C_{19}(\alpha). \tag{28}$$

Furthermore, define $\tilde{W}_T = \sum_{t=0}^{T} w_t - W_T$. By definition of $\hat{T}$ we have $\mathbb{E}[\|M(\hat{T})\|] = \frac{1}{\tilde{W}_T}\sum_{t=0}^{T} w_t\mathbb{E}\|M(t)\|$, and hence

$$\mathbb{E}[\|M(\hat{T})\|] = \frac{1}{\tilde{W}_T}\sum_{t=0}^{1} w_t\mathbb{E}\|M(t)\| + \frac{1}{\tilde{W}_T}\sum_{t=2}^{T} w_t\mathbb{E}\|M(t)\|$$

$$\frac{M_0}{\tilde{W}_{t=0}}\sum_{t=0}^{T} w_t + \frac{1}{T\tilde{W}_{t=2}}\sum_{t=2}^{T} w_t E\|M(t)\|^T \qquad \text{(Lemma B.6)}$$

$$\frac{2M_0 w_{2\ 1}}{\tilde{W}_{t=2}} + \frac{1}{\tilde{W}}\sum_{t=2}^{T} w_t E\|M(t)\|^T \qquad (w_t \ w_{t+1} \text{ for all } t \ 0)$$

$$\frac{2M_0 w_{2\ 1}}{w_{t=2}} + \frac{1}{\tilde{W}}\sum_{t=2}^{T} w_t E\|M(t)\|^T$$

$$= 2M_0\ 1\ \frac{'_2 T^{\ 2+1}}{2} + \frac{1}{\tilde{W}_{t=2}}\sum_{t=2}^{T} w_t E\|M(t)\|^T$$

$$2M_0\ 1\ \frac{'_2 T^{\ 2+1}}{2} + \frac{1}{W}\sum_{t=2}^{T} w_t E\|M(t)\|^T$$

$$4 + {}_{'2}\frac{4}{} M_0\ 1\ \frac{2^{2}}{2'} + {}^{T^{2+1}} + 4\ \frac{C_{14}(\ )_3}{'_2} + C_{18}(\ )(K\ \ 1)^2 + C_{19}(\ )\ \ \frac{}{N} \qquad \text{(by (28))}$$

$$= C_{20}(\ )\ \ 1\ \ \frac{'_2 T^{\ 2+1}}{2} + \frac{4C_{14}(\ )_3}{'_2} + C_{18}(\ )(K\ \ 1)^2 + C_{19}(\ )\ ;_2\ \frac{}{N}$$

where $C_{20}(;) = 4M_0 + {}_{'2}\ ^{4M_0}$. Furthermore, by Proposition B.1, we have $M(\hat{}) = {}_2 k\hat{} k_m \ \frac{1}{2u}\ \frac{}{T}\ k^2_T k^2$, and hence

$$E[k_T k_{\hat{c}}]_2 C_1(;)\ 1\ \frac{'_2 T^{\ 2+1}}{2} + C_2(\ ) + C_3(\ )(K\ \ \frac{1}{N})^2 + C_4(\ )^3 \qquad (29)$$

where $C_1(;) = 2u_{cm}{}^2 C_{20}(;)$, $C_2(\ ) = 2u_{cm}C_{19}(\ )$, $C_3(\ ) = 2u_{cm}C_{18}(\ )$ and $C_4(\ ) = 2u_{cm}:\ ^{4C_{14}(\ )}.\ ^{2}\ \frac{}{'_2}$

Finally, note that by definition of , we have $= d2\log e\ 1 + 2\log\ = 1 + 2(\ln\ )(\log e) = 1 + (2\log\ \frac{1}{e})\ln\ \frac{1}{1} + (2\log\ _e)$. Hence, we have $C_1(;)\ 2u_{cm}M_0(4 + \log\ _e + \ ^2) \ 16u_{cm}M_0(\log\ \frac{1}{e} + \frac{4}{2}) = C_1 :_2$ where $C_1 = 16u_{cm}^2 M_0(\log\ \frac{1}{e} + {}_{'2})$.

Furthermore, we have $C_2(\ ) = 2u_{cm}C_{19}(\ ) = 2u_{cm}{}_{'2}\ C_{16}(\ ) = \frac{8u_{cm}(C_8 + \frac{1}{2} + C_{12})}{2} \ \frac{8u_{cm}\ C_8 + \frac{1}{2} + C_{12}}{2} = C_2$, where we $\frac{^2(\ _2)}{'_2}$ $_2$ denote to emphasize the dependence of on , and $C_2 = \frac{8u_{cm}\ C_8 + \frac{1}{2} + C_{12}}{2}.\ \frac{^2(\ _2)}{'_2}$ Note that we have $= O(\log(1=))$.

In addition, $C_3(\ ) = 2u_{cm}C_{18}(\ ) = \frac{80B^2 C_{17}u_{cm}(1+{}_{B(1\ )})}{'_2}$: C_3, where C_3 And lastly, $= \frac{80B^2 C_{17}u^2_{cm}(1+{}_{B(1\ )})}{'_2}$.

$C_4(\ ) = \frac{8u_{cm}(C_7 + C_{11} + 0{:}5C_3C_9 + C_3C_{10} + 2C_1C_3C_{10} + 3A_1C_3 + C_{13})_2}{}$ $C_4{}^2$, where $C_4 = 8u_{cm}\ C_7 + C_{11} + 0{:}5C^2C^2 + C_3C_{10} + 2C_1C_3C_{10} + 3A_1C_3 + C_{13}\ =' _2$. □

Next we will state the proof of Corollary B.1.1. $^2{}_2\frac{\tilde{}\ {}^2\ \frac{4m_2}{}}{{}_2\ {}_2\ '_2}$ $\frac{\tilde{}\ \frac{4m_2}{}}{'_2}$

Proof of Corollary B.1.1. By this choice of step size, for large enough T, will be small enough and can satisfy the requirements of the step size in (21), (23), (32), (35). Furthermore, the first term in (29) will be

$$C_1\frac{1}{2})\ 1\ \frac{'_2 T^{\ 2+1}}{2} \quad C\ \ {}_{1-}e^{\ \frac{4\log(NT)}{T}(T\ \ {}_2+1)} = C_1\frac{'_2 T}{8\log(NT)}e^{\ \log((NT)^4)(1+\frac{1}{T}}$$

$$= \frac{C_1{}'_2 T}{8\log(NT)}\ \frac{1}{N^4T^4}^{(1+\frac{1\ \ 2}{T})}$$

$$\frac{C_1{}'_2 T}{8\log(NT)}\ \frac{1}{N^4T^4}^{0:5} \qquad \text{(Assumption on } T\text{)}$$

$$= \frac{C_1{}'_2 T}{8\log(NT)}\frac{1}{N^2T^2}$$

$$= \tilde{O}\left(\frac{C_1'^2}{NT}\right) :$$

Furthermore, for the second term we have

$$C^{N\text{—}} = \tilde{O}\left(\frac{C_2='^2_2}{NT}\right) :$$

Finally, for the third and the fourth terms we have

$$C_3(K \quad 1)^2 + C_4^{32} = O(C_3='^2_2 + C_4='^3_2)\,\frac{\log^2(NT)}{T^2} = O\left(\frac{(C_3='^2_2 + C_4='^2_2)}{NT}\right) :$$

Upper bounding (29) with , we get $\tilde{O}\left(\frac{C_1'^2_2+C_2='^2_2+C_3='^2_2+C_4='^2_2}{NT}\right)$. Hence, we need to have $T = \tilde{O}\left(\frac{C_1'^2_2+C_2='^2_2+C_3='^2_2+C_4='^3}{N^2}\right)$ number of iterations to get to a ball around the optimum with radius . □

B.3. Proof of Proposition B.2

The update of the virtual parameter sequence $f_t g$ can be written as follows

$$_{t+1} = _t + (G(_t; Y_t) \quad _t + b(Y_t)) : \tag{30}$$

Using $\frac{p\ 1}{}$-smoothness of $M()$ (Proposition B.1), we get

$$\begin{aligned}
M(_{t+1}) \;&\le\; M(_t) + hrM(_t); _{t+1} \quad _t i + \frac{L}{2} k_{t+1} \text{—} _t k_s^2 \qquad \text{(Smoothness of } M()\text{)} \\
&= M(_t) + hrM(_t); G(_t; Y_t) \quad _t + b(Y_t)i + \frac{L^2}{2} kG(_t; Y_t) \quad _t + b(Y_t)k_s^2 \\
&= M(_t) + \underbrace{rM(_t); G(_t) \quad _t}_{T_1:\ \text{Expected update}} + \underbrace{hrM(_t); b(Y_t)i}_{T_2:\ \text{Error due to Markovian noise } b(Y_t)} \\
&\quad + \underbrace{hrM(_t); G(_t; Y_t) \quad G(_t)i}_{T_3:\ \text{Error due to Markovian noise } Y_k} + \underbrace{hrM(_t); G(_t) \quad G(_t)i}_{T_4:\ \text{Error due local updates}} : \\
&\quad + \underbrace{\frac{L^2}{2} kG(_t; Y_t) \quad _t + b(Y_t)k_s^2}_{T_5:\ \text{Error due to noise and discretization}}
\end{aligned} \tag{31}$$

The inequality in (31) characterizes one step drift of the Lyapunov function $M(_t)$. The term T_1 is responsible for negative drift of the overall recursion. T_2 and T_3 appear due to the presence of the Markovian noises $b^i(y^i)$ and $G^i(^i; y^i)$ in the update of Algorithm 4. T_4 appears due to the mismatch between the parameters of the agents $^i; i = 1; : : : ; N$. Finally, T_5 appears due to discretization error in the smoothness upper bound. Next, we state bounds on $T_1; T_2; T_3; T_4; T_5$ in the following intermediate lemmas.

Lemma B.1. For all $2 R^d$, the operator $G()$ satisfies the following

$$T_1 = rM(); G() \qquad 2'_2 M() :$$

Lemma B.1 guarantees the negative drift in the one step recursion analysis of Proposition B.2. It follows from the Moreau envelope construction (Chen et al., 2021c) and the contraction property of the operators $G^i(); i = 1; : : : ; N$ (Assumption 6.2).

Lemma B.2. Consider the iteration t of the Algorithm 4, and consider $= d2 \log e$. We have

$$E_t \quad [T_2] = E_t \quad hrM(_t); b(Y_t)i$$

$$\frac{L}{2}\frac{2}{l_{cs}}E_t\left[\left\|k_t - {}_t k_c\right\| + \frac{2}{2}E_t\left[\left\|kb(Y_t)k_c\right\|^2\right]\right.$$
$$\left. + \frac{m_2 L}{l_{cs}^2}E_t\left[\left\|k_t - {}_t k_c\right\| + \frac{1}{2}\right]\right. \frac{Lm_2{}^2u_{cm}{}^2}{l^2} + {}^2E_t\left[M(t)\right];$$

where $_1$ is an arbitrary positive constant.

In the i.i.d. noise setting, $E[T_2] = 0$. In Markov noise setting, going back steps (which introduces $_t$) enables us to use Markov chain mixing property (Assumption 6.1).

Lemma B.3. For any t 0, denote $T_3 = hrM(t); G(t; Y_t) \quad G(t)i$. For any $<$ t, we have $E_t[T_3]$

$$\left.{}^2\frac{2LA_2}{l_{cs}^2} + {}^2\frac{L(A_1+1)}{l_{cs}^2} + \frac{6m_1Lu_{cm}}{3}\frac{E_t\left[M(\ )\right]}{l_{cs}^2}\right.^2$$
$$+ \frac{2LA_2}{l_{cs}^2}\left(\frac{1}{2} + \frac{u_{cm}^2}{2^2_2} + \frac{L(A_1+1)}{l_{cs}^2}\left(\frac{3u_{cm}^2}{2^2_3} + 2\right) + \frac{3m_1L}{l_{cs}^2}E_t\left[\left\|k_t - {}_t k\right\|^2\right]\right.$$
$$+ \frac{3u_{cm}^2}{2\ {}^2_3} + \frac{3}{2}\frac{L(A_1+1)}{l_{cs}^2} + \frac{LA_2}{l_c^2}E_t\left[\right.$$
$$\left.{}_t\right] + \frac{{}_{cm}}{3u_3^2} + \frac{}{3}\frac{}{2^2}\frac{L(A_1+1)}{l_{cs}} + \frac{}{2}\frac{m_1^2L}{l_{cs}}E_t\left[\right.\left.\right];{}_t$$

where $_2$ and $_3$ are arbitrary positive constants.

Lemma B.4. For any t 0, we have

$$T_4 = hrM(t); G(t) \quad G(t)i \ {}_4M(t) + {}^2\frac{L^2u_{cm}^2}{2l_{cs}^2\ {}^2\ {}^4_4}$$
$$_t;$$

where $_4$ is an arbitrary positive constant.

Lemma B.5. For any 0 $<$ t, we have

$$T_5 = \left\|G(t; Y_t) \quad {}_t + b(Y_t)k^2\right. {}_s\frac{6(A_2+1)^2u_{cm}^2}{l_t^2}M(t) + \frac{3A_1}{3}\left(\frac{}{l^2}\right)^2\left(\frac{}{l_{cs}^2}\right)^2$$
$$+ kb(Y_t)k_c : {}_{cs}{}_{cs}$$

Substituting the bounds in Lemmas B.1, B.2, B.3, B.4, B.5, and taking expectation, we get the final bound in Proposition B.2.

B.4. Proof of Proposition B.3

First we state the following two intermediate lemmas, which are proved in Section B.7.

Lemma B.6. Suppose $=$ $d2\log e$ and

$$\min\left(\frac{1}{12\sqrt{A_2^2+1}}; \frac{1}{8(A_2+1)}\right) : \tag{32}$$

For any 0 t 2 we have the following

$$M(t)\left(1 - {}_{cm}\frac{1}{2}\right)\left(\frac{1}{C_1^2}B + (A_2+1)\left(k_0k_c + \frac{B}{2C_1}\right)^2 + k_0k^2{}_c\right)^{!}M_0 \tag{33}$$

Furthermore, for any t 2, we have the following

$$E_t{}_2[k_t - {}_t k_c] \ 4C_1E^{t}{}_2[k_tk^{c}] + 8\left(\frac{B}{l_{cD}} + \frac{u_{cD}}{l_{cD}}\right)\left(1 + \frac{u_{cD}}{2C_1}\right)\frac{8m_2}{N}$$
$$+ 6A_1 \quad \sum_{i=t}^{t} E_t{}_2[{}_i]: \tag{34}$$

Lemma B.7. Suppose = d2 log e and

$$\min \quad \frac{1}{C_1}; \frac{1}{4C_2}, \frac{1}{40C_1^{\ 2}} \quad ; \tag{35}$$

where $C_1 = 3\sqrt{A_2^2 + 1}$ and $C_2 = 3C_1 + 8$. We have the following

$$E_{t\ 2}[k_t \quad _t \quad k^2] \quad 8^2_c C^2 E_{t\ 2} k_t k_1^2 + 8\frac{_{cD}}{} \frac{B^2}{c} {}^{22}{}_{cD} \frac{u^2}{l^2} \frac{}{N}$$
$$+ \ 14\frac{u^2_{cD}}{l^2} \frac{m_2^{24}}{1} + \frac{}{2} 8^2 A^2 \quad \overset{X}{\underset{i=0}{}} E_{t\ 2}{}^{2}{}_{i:cD\ i=0}$$

In Lemma B.6, we define $C_9 \ , \ \frac{8u_{cD}B}{l_{cD}}; C_{10} \ , \ \frac{8m_2 u_{cD}}{l_{cD}(1 \quad)}$. Hence, we can bound the term in (15) as

$$^2C_3 E_{t\ 2}[k_t \quad _t \quad k_c]$$
$$^2C_3 \quad 4C_1 E_{t\ 2}[k_t k_c] + C_9 \frac{}{} + C_{10}{}^2(1 + 2C_1) + 6A_1 \quad \overset{X}{\underset{k=t}{^t}} E_{t\ 2}[k] \quad !$$
$$^2C_3 \quad 4C_1 u_{cm} E_{t\ 2} \quad \overset{h}{} \frac{}{2M(t)} + C_9 \frac{}{N} + C_{10}{}^2(1 + 2C_1) + 6A_1 \quad \overset{X^t}{\underset{k=t}{}} E_{t\ 2}[k] \quad !$$

(Proposition B.1)

$$= \frac{1}{6} 4C_1 C_3 u_{cm}{}^{5=2} \quad \overset{p}{} \overset{p}{} 2E_{t\ 2}[M(t)]$$
$$+ \ ^2C_3 \quad C_9 \frac{}{N} + C_{10}{}^2(1 + 2C_1) + 6A_1 \quad \overset{X}{\underset{k=t}{}} E_{t\ 2}[k] \quad !$$

($\overset{p}{}$: is concave, $_6 > 0$)

$$8\frac{^2C_1^2 u_{cm}^2}{C} \frac{3}{2}{}^{52} + {}^2E_{t\ 2}[M(t)] + C_3 C_9 \frac{}{} + C_3 C_{10} \frac{}{N}{}^3 (1 + 2C_1)^{\ 6} \quad _t$$
$$+ \ 6A_1 C_3 \quad \overset{X}{\underset{k=t}{}} E_{t\ 2} \quad \frac{1}{2}{}^5 + \frac{1}{2} \quad _k \qquad \text{(Young's inequality)}$$

$$8C^2_{\frac{1}{2}} C_1^2 u_{cm}^2 {}^{52} + {}^2E_{t\ 2}[M(t)] + C_3 C_9 \frac{}{} + {}^4[C_3 C_{10}(1 + 2C_1) + 3A_1 C_3(+ 1)] \underbrace{\quad}_{C_6}$$
$$+ \ 3A_1 C_3 \quad \overset{X}{\underset{k=t}{}} E_{t\ 2}[\quad _k] \qquad \text{(By (58))}$$

$$C_{11}{}^{52} + {}^2E_{t\ 2}[M(t)] + \frac{}{2} C^2 C^{24} + \frac{1}{2}{}^3 \frac{}{N} + {}^4[C_3 C_{10}(1 + 2C_1) + 3A_1 C_3(+ 1)] + 3A_1 C_3 \quad \overset{X}{\underset{k=t}{}} E_t$$
$$_{2}[\quad _k]: \tag{36}$$

Furthermore, using Lemma B.7, (58) and Proposition B.1, the term in (16) can be bounded as follows:

$$C_4 E_{t\ 2} k_t \quad _t \quad k^2 \ 16^2_c C^2 C_4 u^2_m E_{t\ 2}[M(t)] + 8^2 A^2 C_4 \quad \overset{X}{\underset{t\ k=0}{}} E_{t\ 2}[\quad _1$$
$$_t \quad _k]$$
$$+ \ \underbrace{\frac{8C_4 u_{cD}^2 B^2}{l^2_{cD}} \frac{}{N}{}^2}_{C_{12}} + \underbrace{\frac{14C_4 u_{cD}^2 m_2^2}{l^2_{cD}(1 \quad)^2}}_{C_{13}} \quad :^4 \tag{37}$$

Inserting the upper bounds in (36) and (37) in the upper bound in Proposition B.2, we have

$$
\begin{aligned}
&E_{t-2}[M(t+1)] \\
&\quad \le \Big(1 - \tfrac{3}{2}\alpha\beta + \alpha^2 \tfrac{2LA_2}{l_{cs}^2} + \alpha^2 \tfrac{L(A_1+1)}{l_{cs}^2} + \alpha^2 c_{s4} c_{s6}\Big) \\
&\qquad + \alpha^2 \tfrac{6Lu_{cm}^2}{l_{cs}^2} m_1 + \alpha^2 \tfrac{6L(A_2+1)^2 u_{cm}^2}{l_{cs}^2} + \alpha^2 16^2 C^2 C_4 u_{cm}^2 m_1 \Big) E_{t-2}[M(t)] \\
&\qquad + C_5 E_{t-2}[\quad_t] + C_6 E_{t-2}[\quad_t] \\
&\qquad + \Big(C_7 + C_{11}\alpha^2 + 0.5C^2C^2\alpha^3 + C_3C_{10} + 2C_1C_3C_{10}\alpha + 3A_1C_3(\alpha+1) + C_{13} + C_8 \tfrac{u_{cD}^2}{l_{cs}^2} 2m_2^2 \alpha^4\Big)}_{C_{14}(\alpha)} \\
&\qquad + C_8 \tfrac{u_{cD}^2 B^2}{l_{cs}^2} \alpha^4 + \alpha^2 C_{12}^2 \tfrac{\alpha}{N} + 3A_1C_3 + 8\alpha^2A^2C_4 \sum_{k=t-1}^{t} E_{t-2}[\quad_k]. \quad (38)
\end{aligned}
$$

We define $C_{14}^0(\alpha) \equiv C_{14}(\alpha)$, $\alpha_0 = \min\Big(\sqrt{\tfrac{\beta^2}{10}}, \tfrac{l_{cs}^2}{2LA^2}, \tfrac{\beta^2}{10}, \tfrac{l_{cs}^2}{L(A^2+1)}\Big)$, and denote

$$
C_{15}(\alpha) = \frac{6m_1Lu_{cm}\alpha}{l_{cs}^2} + \frac{6L(A_2+1)^2u_{cm}^2\alpha}{2\,l_{cs}^2} + 144(A_2+1)C_4u_{cm}^2\alpha^2. \quad (39)
$$

This yields

$$
\begin{aligned}
&E_{t-2}[M(t+1)] \le \Big(1 - \tfrac{3}{2}\alpha\beta + 2\alpha^2 C_{15}(\alpha)\Big) E_{t-2}[M(t)] + C_8 \tfrac{u_{cD}^2 B^2}{l_{cs}^2}\alpha^4 + \alpha^2 C_{12}^2 \tfrac{\alpha}{N} \\
&\qquad + C_{14}(\alpha)^4 + (3A_1C_3 + 8\alpha^2A^2C_4 + \alpha C_5 + C_6) \sum_{k=t-1}^{t} E_{t-2}[\quad_k] \\
&\quad \le \Big(1 - \tfrac{3}{2}\alpha\beta + 2\alpha^2 C_{15}(\alpha)\Big) E_{t-2}[M(t)] + C_{14}(\alpha)^4 + C_{16}(\alpha)\tfrac{\alpha}{N} + C_{17}\sum_{k=t-1}^{t} E_{t-2}[\quad_k], \quad (40)
\end{aligned}
$$

where $C_{16}(\alpha) = C_8 \tfrac{u_{cD}^2 B^2}{l_{cs}^2}\alpha^2 + \alpha C_{12}^2$ and $C_{17} = (3A_1C_3 + 8A_1C_4\alpha + \alpha C_5 + C_6)$. Due to $\alpha \le \frac{1}{2C_{15}(\alpha)}$, we have $1 - \tfrac{3}{2}\alpha\beta + 2\alpha^2 C_{15}(\alpha) \le 1 - \tfrac{3}{4}\alpha\beta$. This completes the proof.

B.5. Proof of Proposition B.4

First we state the following lemma, which characterizes a bound on the expectation of the synchronization error $\quad_t$.

Lemma B.8. Suppose Assumptions 6.1, 6.3 holds and the step size α satisfies $\alpha \le \frac{\sqrt{\ln(5/4)}}{2(1+A^2)(K-1)^2}$. Then, for $sK \le t \le (s+1)K - 1$, where $s = \lfloor t/K \rfloor$, the network consensus error $\quad_t \triangleq \frac{1}{N}\sum_{i=1}^{N} \|x_t^i - x_t\|_c^2$ satisfies

$$
E\quad_t \le 5^2(t - sK)B^2\alpha^2\Big(1 + \frac{4m_2}{1-\sqrt{\alpha\beta}}\Big) + 5^2(t - sK)A_2\alpha^2 \sum_{t'=sK}^{t-1} E\|x_{t'} - x_{t'}\|_c^2, \quad (41)
$$

where $\tilde{A}_1 = \frac{2A_1^2u_{c2}^2}{l_{c2}^2}$; $\tilde{A}_2 = \frac{2A_2^2u_{c2}^2}{l_{c2}^2}$; $\tilde{B} = \frac{u_{c2}}{l_{c2}}B$. Here, A_1, A_2, B are the constants defined in Assumption 6.3, and l_{c2}, u_{c2} are constants involved in the equivalence of the norms: $l_{c2}\|x\|_2 \le \|x\|_c \le u_{c2}\|x\|_2$.

Due to the assumption on the step size, the bound in Lemma B.8 hold. Substituting the bound on $\quad$ from Lemma B.8, we get

$$
\frac{2C_{17}}{\beta}\sum_{t=2}^{T}\Big(w_t \sum_{\tau=t}^{t} E\Big[\quad_\tau\Big]\Big)
$$

$$\frac{2C_{17}}{\alpha_2 W_T}\sum_{t=2}^{T} w_t \sum_{\tau=t-\tau_{t-1}}^{t-1} 5^2(\tau - s_{\tau,K})B^2\alpha_{t-1}^2 + B(1-\frac{4m_2}{4})\alpha_{t-1}^2 + 5^2(\tau - s_{\tau,K})A_2\alpha_{s_{\tau,K}}^2 \sum_{t'=s_{\tau,K}}^{\tau-1} \mathbb{E}\|x_{t'} - x_{s_{\tau,K}}\|_c^2. \tag{42}$$

where $s_\tau = b\tau/Kc$. Hence, $s_{\tau,K}$ denotes the last time instant before τ when synchronization happened. The first term in (42) can be upper bounded as follows.

$$\begin{aligned}
&\frac{2C_{17}}{\alpha_2 W_T}\sum_{t=2}^{T} w_t \sum_{\tau=t-\tau_{t-1}}^{t-1} 5^2(\tau - s_{\tau,K})B^2\alpha_{t-1}^2 + B(1-\frac{4m_2}{4})\alpha_{t-1}^2 \\
&= 5^2B^2\alpha_1^2 + B(1-\frac{4m_2}{4})\alpha_1^2 \frac{2C_{17}}{\alpha_2 W_T}\sum_{t=2}^{T} w_t \sum_{\tau=t-\tau_{t-1}}^{t-1} (\tau - s_{\tau,K})^2 \\
&\quad + B(1-\frac{4m_2}{4})\frac{2C_{17}}{\alpha_2 4m_2 W_T}\sum_{t=2}^{T} w_t \sum_{\tau=t-\tau_{t-1}}^{t-1} (K-1) \qquad (\tau - s_{\tau,K} \le K-1) \\
&= 5^2B^2\alpha_1^2 + B(1-\frac{4m_2}{4})\frac{1}{\alpha_2}\Big[\frac{2C_{17}}{W_T}\sum_{t=2}^{T} w_t (K-1)\Big] \\
&\le 5^2B^2\alpha_1^2 + B(1-\frac{4m_2}{4})\frac{10C_{17}}{\alpha_2}(K-1).
\end{aligned} \tag{43}$$

Next, we compute the second term in (42).

$$\begin{aligned}
&\frac{2C_{17}}{\alpha_T W_T}\sum_{t=2}^{T} w_t \sum_{\tau=t-\tau_{t-1}}^{t-1} 5^2(\tau - s_{\tau,K})A_2\alpha_{s_{\tau,K}}^2 \sum_{t'=s_{\tau,K}}^{\tau-1} \|x_{t'} - x_{s_{\tau,K}}\|_c^2 \\
&\le 2A_2\alpha_1^2\frac{20C_{17}u_{cm}^2}{\alpha_T W_T}\sum_{t=2}^{T} w_t \sum_{\tau=t-\tau_{t-1}}^{t-1} (\tau - s_{\tau,K})\sum_{t'=s_{\tau,K}}^{\tau-1} M(t')^2
\end{aligned} \tag{Proposition B.1}$$

$$\alpha_2 \le \frac{20C_{17}u_{cm}^2}{\alpha_2 W_T}\Bigg[\underbrace{\sum_{t=2}^{K} w_t \sum_{\tau=t-\tau_{t-1}}^{t-1} (\tau - s_{\tau,K})\sum_{t'=s_{\tau,K}}^{\tau-1} M(t')}_{I_1} + \underbrace{\sum_{t=K+1}^{T} w_t \sum_{\tau=t-\tau_{t-1}}^{t-1} (\tau - s_{\tau,K})\sum_{t'=s_{\tau,K}}^{\tau-1} M(t')}_{I_2}\Bigg] \tag{44}$$

where if $K < 2$, $I_1 = 0$. Next, we bound $I_1; I_2$ separately.

$$\begin{aligned}
I_1 &= \sum_{t=2}^{K} w_t \sum_{\tau=t-\tau_{t-1}}^{t-1} (\tau - s_{\tau,K})\sum_{t'=s_{\tau,K}}^{\tau-1} M(t') \\
&\le (K-1)\sum_{t=2}^{K} w_t \sum_{\tau=t-\tau_{t-1}}^{t-1}\sum_{t'=0}^{\tau-1} M(t') \qquad (\text{for } \tau < K,\ s_\tau = 0;\ \text{for } \tau = K,\ \tau - s_{\tau,K} = 0) \\
&\le (K-1)(\tau+1)\sum_{t=2}^{K} w_t \sum_{t'=0}^{\tau-1} M(t') \\
&\le (K-1)(K-2+1)(\tau+1)w_K \sum_{t=0}^{K-1} M(t);
\end{aligned} \tag{45}$$

where, (45) follows since $w_{t-1} \ge w_t, \forall\, t$. Next, to bound I_2 in (44), we again split it into two terms.

$$I_2 = \underbrace{\sum_{t=K+1}^{K+\tau} w_t \sum_{\tau=t-\tau_{t-1}}^{t-1} (\tau - s_{\tau,K})\sum_{t'=s_{\tau,K}}^{\tau-1} M(t')}_{I_3} + \underbrace{\sum_{t=K+\tau+1}^{T} w_t \sum_{\tau=t-\tau_{t-1}}^{t-1} (\tau - s_{\tau,K})\sum_{t'=s_{\tau,K}}^{\tau-1} M(t')}_{I_4}.$$

First, we bound I_3.

$$I_3 \le \sum_{t=K+1}^{K+} w_t \left[\sum_{\tau=t-K}^{t-1} \left(\sum_{\tau'=s_\tau K}^{\tau-1} M(\tau') + \sum_{\tau=K+1}^{t-1} \sum_{\tau'=s_\tau K}^{\tau-1} M(\tau') \right) \right] \quad (s_\tau = 0 \text{ for } \tau < K,\ s_\tau \ge 1 \text{ for } K \le \tau \le K+)$$

$$\le \sum_{t=K+1}^{K+} w_t \left[K \sum_{\tau'=0}^{t-K-1} M(\tau') + (t-K) \sum_{\tau'=K}^{t-1} M(\tau') \right]$$

$$\le w_{K+} K^2 \sum_{t=0}^{K-1} M(t) + {}^3 w_{K+} \sum_{t=K}^{K+-1} M(t). \quad (46)$$

Next, we bound I_4, assuming $t_0 K + \le T < (t_0+1)K + $, where t_0 is a non-negative integer.

$$I_4 = \sum_{t=K++1}^{T} w_t \underbrace{\left(\sum_{\tau=t-K}^{t-1} \sum_{\tau'=s_\tau K}^{\tau-1} M(\tau') \right)}_{K-1}$$

$$(K-1) \sum_{t=K++1}^{T} w_t \sum_{\tau'=s_\tau K}^{t-1} M(\tau')$$

$$= (K-1) w_{K++1} \sum_{\tau'=s_\tau K}^{K++1-1} M(\tau') + \cdots + w_{K++K-1} \sum_{\tau'=s_\tau K}^{2K+-2} M(\tau')$$

$$+ (K-1) w_{2K+} \sum_{\tau'=s_\tau K}^{2K+-1} M(\tau') + \cdots + w_{3K+-1} \sum_{\tau'=s_\tau K}^{3K+-2} M(\tau')$$

$$+ \cdots + (K-1) w_{t_0 K+} \sum_{\tau'=s_\tau K}^{t_0 K+-1} M(\tau') + \cdots + w_T \sum_{\tau'=s_\tau K}^{T-1} M(\tau')$$

$$(K-1)(+1) w_{K++1} \sum_{\tau'=K}^{K+} M(\tau') + \cdots + w_{2K+-1} \sum_{\tau'=K}^{2K+-2} M(\tau')$$

$$+ (K-1)(+1) w_{2K+} \sum_{\tau'=2K}^{2K+-1} M(\tau') + \cdots + w_{3K+-1} \sum_{\tau'=2K}^{3K+-2} M(\tau')$$

$$+ \cdots + (K-1)(+1) w_{t_0 K+} \sum_{\tau'=t_0 K}^{t_0 K+-1} M(\tau') + \cdots + w_T \sum_{\tau'=t_0 K}^{T-1} M(\tau')$$

$$(K-1)K(+1) w_{2K+-1} \sum_{\tau'=K}^{2K+-2} M(\tau') + w_{3K+-1} \sum_{\tau'=2K}^{3K+-2} M(\tau') + \cdots + w_T \sum_{\tau'=t_0 K}^{T-1} M(\tau')$$

$$(K-1)K(+1) w_{2K+-1} \left(\sum_{\tau'=K}^{2K+-2} M(\tau') + \sum_{\tau'=2K}^{3K+-2} M(\tau') \right)$$

$$+ w_{3K+-1} \left(\sum_{\tau'=2K}^{3K+-2} M(\tau') + \sum_{\tau'=3K}^{4K+-2} M(\tau') \right)$$

$$+ \cdots + w_{t_0 K+-1} \left(\sum_{\tau'=(t_0-1)K}^{t_0 K+-2} M(\tau') + \sum_{\tau'=t_0 K}^{(t_0+1)K+-2} M(\tau') \right) + w_T \sum_{\tau'=t_0 K}^{T-1} M(\tau')$$

$$(K-1)K(+1) \frac{(-1)}{K} \left[\sum_{\tau'=K}^{2K+-2} w_{\tau'+K+-1} M(\tau') + \sum_{\tau'=2K}^{3K+-2} w_{\tau'+K+-1} M(\tau') + \cdots + w_T \sum_{\tau'=t_0 K}^{T-1} M(\tau') \right]$$

$$(K-1)(\tau+1)\left[\sum_{\tau=K}^{T-1} w_{\tau+K+1}M(\tau)+w_T\sum_{\tau=T-K+1}^{T-1}M(\tau)\right]$$
$$=(K-1)(\tau+1)\left[\sum_{\tau=K}^{T-1}\frac{w_\tau}{1-\frac{\alpha_{\tau-K+1}^2}{2}}M(\tau)+\sum_{\tau=T-K+1}^{T-1}\frac{w_\tau}{1-\frac{\alpha_T^2}{2}}M(\tau)\right] \quad (w_t = w_{t+1}(1-\tfrac{\alpha_{t+1}^2}{2}))$$
$$(K-1)(\tau+1)\sum_{t=K}^{T-1}\frac{w_t}{1-\frac{\alpha_{t-K+1}^2}{2}}M(t): \tag{47}$$

Using the bounds on I_3, I_4 from (46) and (47) respectively, we can bound I_2.

$$I_2 \le w_{K+1}K^2\sum_{t=0}^{K-1}M(t)+3w_{K+1}\sum_{t=K}^{T-1}M(t)+(K-1)(\tau+1)\sum_{t=K}^{T-1}\frac{w_t}{1-\frac{\alpha_{t-K+1}^2}{2}}M(t). \tag{48}$$

Substituting the bounds on I_1, I_2 from (45), (48) respectively, into (44), we get

$$\frac{2C_{17}}{\alpha_T W_{T-2}}\sum_{t=2}^{T}w_t\sum_{\tau=t}^{T}\alpha_{\tau-K}^2\sum_{s=0}^{\tau-K-1}\|\theta_s-\theta_{s+K}\|_2^2$$
$$+\alpha^2A_2^2\frac{20C_{17}u_{\max}^2}{\alpha_T W_T}\left[(K-1)(K+2)(\tau+1)w_K\sum_{t=0}^{K-1}M(t)+K^2w_{K+1}\sum_{t=0}^{T-K-1}M(t)\right]$$
$$+\alpha^2A_2^2\frac{20C_{17}u_{\max}^2}{\alpha_T^2W_T}\left[3w_{K+1}\sum_{t=K}^{T-1}M(t)+(K-1)(\tau+1)\sum_{t=K}^{T-1}\frac{w_t}{1-\frac{\alpha_{t-K+1}^2}{2}}M(t)\right] \tag{49}$$

We analyze the terms in (49) separately. First, for the terms with $\sum_{t=0}^{K-1}M(t)$,

$$\frac{1}{W_T}\alpha^2A_2^2\frac{20C_{17}u_{\max}^2}{\alpha_T^2}K(\tau+1)[(K+2)w_K+w_{K+1}]\sum_{t=0}^{K-1}M(t)$$
$$=\frac{1}{W_T}\alpha^2A_2^2\frac{20C_1\alpha_T u_{\max}^2}{\alpha_T^2}K(\tau+1)\sum_{t=0}^{K-1}\left[(K+2)\frac{w_t}{1-\frac{\alpha_{t+K}^2}{2}}+\frac{w_t}{1-\frac{\alpha_{t+2K}^2}{2}}\right]M(t)$$
$$=\frac{1}{W_T}\alpha^2A_2^2\frac{20C_{17}u_{\max}^2}{\alpha_T^2}K(\tau+1)\sum_{t=0}^{K-1}\left[(K+2)+\frac{1}{1-\frac{\alpha_{t+K}^2}{2}}\frac{w_t}{1-\frac{\alpha_{t+K}^2}{2}}\right]M(t)$$
$$\alpha^2A_2^2\frac{20C_{17}u_{\max}^2}{\alpha_T^2}K(\tau+1)\left[(K+2)+\frac{1}{1-\frac{\alpha_0^2}{2}}\right]\frac{1}{W_T}\sum_{t=0}^{K-1}w_tM(t)^2$$
$$\frac{1}{2W_T}\sum_{t=0}^{K-1}w_tM(t); \tag{50}$$

where (50) holds since we choose α small enough such that

$$\alpha^2A_2^2\frac{20C_{17}u_{\max}^2}{\alpha_T^2}(K+2)+\frac{1}{1-\frac{\alpha_0^2}{2}}K(\tau+1)\alpha^2 \le \frac{1}{2}.$$

To get this, we use the inequality $1-x \ge \exp\left(-\frac{x}{1-x}\right)$ for $x<1$, $\frac{\alpha_0^2}{2}\le\frac{1}{2}$ and $\tau+1<2\log\tau+1$, we get $\frac{1}{(1-\frac{\alpha_0^2}{2})}\exp\left(\alpha_0^2\log\tau+1\right)$. For (50) to hold, it is sufficient that

$$\alpha^2 \le \frac{\alpha_0^2}{80C_{17}u_{\max}A_2^2K^2(\log(\tau)+1)}\min\left(\frac{1}{K},\frac{1}{2(\log(\tau)+1)\exp\left(\alpha_0^2\log\tau+1\right)}\right) \tag{51}$$

Next, for the remaining terms in the third line of (49), by the assumption on the step size, we have

$$\begin{aligned} {}^2A_2 &\le \frac{20C_{17}u_{cm}^2}{\varphi_2}\left(1-\frac{\varphi_2}{2}\right)^{\tau-1} \le \frac{1}{4}, \\ {}^2A_2 &\le \frac{20C_{17}^2u_{cm}^2(\tau+1)(K-1)}{\varphi_2}\left(1-\frac{\varphi_2}{2}\right)^{K+1} \le \frac{1}{4}; \end{aligned} \tag{52}$$

and hence we get

$$\frac{1}{W_T}{}^2A_2 \le \frac{20C_{17}u_{cm}^2}{\varphi_2}\left[\frac{3w_{K+1}}{1-\frac{\varphi_2}{2}}M(\tau)+\frac{(\tau+1)(K-1)^T}{1-\frac{\varphi_2}{2}}\sum_{t=K+1}^{\tau}w_tM(t)\right] \le \frac{1}{2W_T}\sum_{t=K}^{T}w_tM(t). \tag{53}$$

Substituting (50), (53) in (49), we get

$$\frac{2C_{17}}{\varphi_T W_T}\sum_{t=2}^{T}w_t\sum_{\tau=t}^{t}5^2(\tau-s,K)A_2 \le \sum_{t=s,K}^{\tau-1}k_tk_c^2 - \frac{1}{2W_T}\sum_{t=0}^{T}w_tM(t). \tag{54}$$

Finally, substituting the bounds in (43), (54) into (42), we get

$$\begin{aligned} &\frac{2C_{17}}{\varphi_2 W_T}\sum_{t=2}^{T}w_t\sum_{\tau=t}^{t}E \\ &{}^2B^2\le 17C_1+B(1-\frac{4(m_2+1)(K-1)}{2}+\frac{w_1EM(t)}{2W_T}. \end{aligned} \tag{55}$$

B.6. Auxiliary Lemmas

The following lemma is of central importance in proving linear speedup of FedSAM.

Lemma B.9. Let l_{cD} and u_{cD} be constants that satisfy $l_{cD}k\ k_D \le k\ k_c \le u_{cD}k\ k_D$, where $k\ k_D = \sqrt{x^\top Dx}$ for some positive definite matrix $D \succ 0$. Note that for any $D \succ 0$, these constants always exist due to norm equivalence. Furthermore, in case the norm kxk_c is defined in the form $\sqrt{x^\top Dx}$ for some $D \succ 0$, we take $l_{cD} = u_{cD} = 1$. We have

$$E_{t-r}[kb(Y_t)k_c] \le \frac{u_{cD}}{l_{cD}}\left(\rho^r+\frac{B}{N}\right)+2m_2\rho^r \tag{56}$$

$$E_{t-r}[kb(Y_t)k_c^2] \le \frac{u_{cD}^2}{l_{cD}^2}\left(\rho^r+\frac{B^2}{N}\right)+2m_2^2\rho^{2r}. \tag{57}$$

Lemma B.9 is essential in characterizing the linear speedup in Theorem B.1. This lemma characterizes the bound on the conditional expectation of $kb(Y_t)k_c$ and $kb(Y_t)k_c^2$, conditioned on r time steps before. In order to understand this lemma, consider the bound in (57). For the sake of understanding, suppose the noise Y_t is i.i.d. In this case we will end up with the first term which is proportional to $1=N$. This is precisely the linear reduction of the variance of sum of N i.i.d. random variables. Furthermore, in order to extend the i.i.d. noise setting to the more general Markovian noise, we need to pay an extra price by adding the exponentially decreasing term to the first variance term.

Lemma B.10. The following hold

$$\begin{aligned} &kG(t;Y_t) \quad G(t;Y_t)k_c \quad A_{1t}^2 \quad A_1{}^2 \quad {}^2 \quad {}^2 \\ &{}_t \quad {}^2 \\ &{}_t: \end{aligned} \tag{58}$$

Lemma B.11. For the generalized Moreau Envelope defined in (12), it holds that

$$\begin{aligned} &r\,M_f^{\ ;g}(x) \quad \overset{?}{=} kxk_m\ ; \\ &D \qquad\qquad\quad E \\ &\quad r\,M_f{}^{;g}(x);x \quad 2M_f{}^{;g}(x): \end{aligned}$$

B.7. Proof of Lemmas

Proof of Lemma B.1. Using Cauchy-Schwarz inequality, we have

$$\langle \nabla M(\theta), G(\theta) \rangle \le \|\nabla M(\theta)\|_m^* \|G(\theta)\|_m \tag{59}$$

$$\underbrace{\langle \nabla M(\theta), G(\theta) \rangle}_{T_{11}} \le \|\nabla M(\theta)\|_m^* \underbrace{\|G(\theta)\|_m}_{T_{12}}$$

where $\|\cdot\|_m^*$ denotes the dual of the norm $\|\cdot\|_m$. Furthermore, by Proposition B.1 we have (Lemma B.11)

$$T_{11} \le \|\nabla M(\theta)\|_m^* \|G(\theta)\|_m \tag{a}$$

$$\le \|\nabla M(\theta)\|_m^* \left\| \frac{1}{N} \sum_{i=1}^N G^i(\theta) \right\|_c \qquad \text{(By (13))}$$

$$\le \|\nabla M(\theta)\|_m^* \frac{1}{N} \sum_{i=1}^N \|G^i(\theta)\|_c \qquad \text{(triangle inequality)}$$

$$\le \|\nabla M(\theta)\|_m^* \frac{c_m u_c}{l_c} \qquad \text{(Assumption 6.2, Proposition B.1)}$$

$$= \frac{2 u_c c_m}{l_c} \|\nabla M(\theta)\|_m^* \tag{60}$$

Furthermore, by the convexity of the $\|\cdot\|_m$ norm (Lemma B.11), we have

$$T_{12} = \|\nabla M(\theta)\|_m \le 2 \|\nabla M(\theta)\|_m^* \tag{61}$$

Hence, using (60), (61) in (59) we get

$$\|\nabla M(\theta)\|_m^* \ge \frac{l_c}{2 u_c c_m} \langle \nabla M(\theta), G(\theta) \rangle \tag{62}$$

where φ_2 is defined in (14). □

Proof of Lemma B.2. Given some $t < \tau$, $T_2 = \langle \nabla M(\theta_t), b(Y_t) \rangle$ can be written as follows:

$$T_2 = \langle \nabla M(\theta_t) - \nabla M(\theta_{t-\tau}), b(Y_t) \rangle + \langle \nabla M(\theta_{t-\tau}), b(Y_t) \rangle$$

$$\le \|\nabla M(\theta_t) - \nabla M(\theta_{t-\tau})\|_s^* \|b(Y_t)\|_s + \langle \nabla M(\theta_{t-\tau}), b(Y_t) \rangle \qquad \text{(Cauchy–Schwarz)}$$

$$\le \frac{L}{l_{cs}} \|\theta_t - \theta_{t-\tau}\|_c \|b(Y_t)\|_{cs} + \langle \nabla M(\theta_{t-\tau}), b(Y_t) \rangle \qquad \text{(Proposition B.1)}$$

$$\le \frac{1}{2} \left(\frac{L}{l_{cs}} \right)^2 \|\theta_t - \theta_{t-\tau}\|_c^2 + \frac{1}{2} \|b(Y_t)\|_{cs}^2 + \langle \nabla M(\theta_{t-\tau}), b(Y_t) \rangle : \tag{63}$$

Taking expectation on both sides, we have

$$\mathbb{E}_t[T_2] \le \frac{L^2 l_{cs}^2 \alpha_t^2}{2} \mathbb{E}_t \|\theta_t - \theta_{t-\tau}\|_c^2 + \frac{1}{2} \mathbb{E}_t \|b(Y_t)\|_{cs}^2 + \mathbb{E}_t \langle \nabla M(\theta_{t-\tau}), b(Y_t) \rangle$$

$$\le \frac{L^2 l_{cs}^2 \alpha_t^2}{2} \mathbb{E}_t \|\theta_t - \theta_{t-\tau}\|_c^2 + \frac{1}{2} \mathbb{E}_t \|b(Y_t)\|_{cs}^2 + \underbrace{\|\nabla M(\theta_{t-\tau})\|_s^* \|\mathbb{E}_t[b(Y_t)]\|_s}_{T_{21}} : \tag{64}$$

For T_{21}, we have

$$T_{21} \le l_{cs} \|\nabla M(\theta_{t-\tau})\|_s^* \|\mathbb{E}_t[b(Y_t)]\|_c$$

$$\le [\|\nabla M(\theta_t)\|_s^* + \|\nabla M(\theta_t) - \nabla M(\theta_{t-\tau})\|_s^*] m_2 \alpha_t \tau \qquad \text{(Assumption 6.1)}$$

$$\le \frac{m_2 \alpha_t \tau}{l_{cs}} [\|\nabla M(\theta_t)\|_s^* + L \|\theta_t - \theta_{t-\tau}\|_c] \qquad \text{(assumption on } \alpha_t\text{)}$$

$$\frac{m_2^2 L}{l^2}\|k_t - {}_t k_s + k\|rM(t) - rM(0)k^2 \quad \text{(Proposition B.1)}$$

$$\frac{m_2^2 L}{l^2}[\|k_t - {}_t k_c\|^2 + \|k_t k_c\|^2] \quad \text{(By (13))}$$

$$\frac{m_2^2 L}{l^2}\|k_t - {}_t k_c\|^2 + \frac{m_2^2 u_{cm} L}{l^2}\|k_t k_m\|^2 \quad \text{(Proposition B.1; } {}_1 > 0\text{)}$$

$$\frac{m_2^2 L}{l^2}\|k_t - {}_t k\|^2 + \frac{1}{2}\|k_t k_c\|^2 + \frac{m_2^2 u_{cm} L}{l^2}\|^2 + {}^2 M_1(t). \quad (65)$$

Substituting (65) in (64), we get

$$E_t[T_2] \le \frac{L}{2}\|^2 E_t\|k_t - {}_t k\|^2 + \frac{L}{2}\|^2 E_t\|k_t - {}_t k_c\|^2 + \frac{L}{l_c}E_t\|k_t - {}_t k_c\|^2 + \frac{m_2^2 u_{cm} L}{l^2}\|^2 + {}^2 E_t[M(t)]. \quad (66)$$

Proof of Lemma B.3. □

$$\begin{aligned} T_3 &= \langle rM(t); G(t; Y_t) - G(t)\rangle \\ &= \underbrace{\langle rM(t) - rM(t-); G(t; Y_t) - G(t)\rangle}_{T_{31}} \\ &\quad + \underbrace{\langle rM(t-); G(t; Y_t) - G(t-; Y_t) + G(t-) - G(t)\rangle}_{T_{32}} \\ &\quad + \underbrace{\langle rM(t-); G(t-; Y_t) - G(t-)\rangle}_{T_{33}}. \end{aligned} \quad (66)$$

Next, we bound all three terms individually.

I. Bound on T_{31}:

$$T_{31} = \frac{1}{N}\sum_{i=1}^{N} \underbrace{\|rM(t) - rM(t-)\|}_{T_{311}} \underbrace{\|G^i(t; {}^i_t; y^i_t) - G^i(t; {}^i_t)\|}_{T_{312}}. \quad (67)$$

For T_{311}, we have

$$T_{311} = \|rM(t) - rM(t-)\| \le \frac{L}{cs}\|k_t - {}_t k_c\|; \quad (68)$$

where the inequality follows from Proposition B.1 and (13). For T_{312}, we have,

$$\begin{aligned} T_{312} &= \|G^i(t; {}^i_t; y^i_t) - E_{y^i} G^i(t; {}^i_t; y)\| \\ &\le \frac{1}{cs}\|G^i(t; {}^i_t; y^i_t) + E_{y^i} G^i(t; {}^i_t; y)\| \quad \text{(Triangle and Jensen's inequality)} \\ &\le \frac{1}{cs}[A_2\|k_t k_c\| + A_2\|k_t k_c\|] \quad \text{(Assumption 6.3)} \\ &\le \frac{2A_2}{cs}\|k_t - {}_t k_c\| + \frac{2A_2}{cs}\|k_t k_c\| \quad \text{(triangle inequality)} \\ &= \frac{2A_2}{cs}\|k_t - {}_t k_c\| + \frac{2A_2}{cs}\|k_t k_c\|. \end{aligned} \quad (69)$$

Substituting (68) and (69) in (67), we get

$$
\begin{aligned}
T_{31} &\le \frac{2LA_2}{l^2}\left[\frac{1}{N}\sum_{i=1}^{N} k_t\,\|\theta_t-\theta_c^{i}\|+k_t\|\theta_c-\theta_t^{cs}\|\right] && \\
&\le \frac{2LA_2}{l_{cs}^2}\left[\frac{1}{N}\sum_{i=1}^{N} k_t\,\|\theta_t-\theta_c\|+\frac{u_{cm}}{2}k_t\|\theta_t-\theta_c\|^2+\frac{k_t}{2u_{cm}}k_m^2\right] && \text{(Proposition B.1)}\\
&\le \frac{2LA_2}{l_{cs}^2}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}k_t\|\theta_t-\theta_c\|^2+\frac{1}{2}\left(\frac{u_{cm}}{2}\right)^2k_t\|\theta_t-\theta_c\|^2+\frac{k_t}{2u_{cm}}k_m^2\right]=\frac{2LA_2}{l_{cs}^2}\left[\frac{1}{2}+\frac{u_{cm}^2}{8}\right]k_t\|\theta_t-\theta_c\|^2+\frac{LA_2}{u_{cm}l_{cs}^2}k_t^2k_m^2 && \text{(Young's inequality)}\\
&\quad+\frac{u_{cm}}{l_{cs}^2}k_t\|\theta_t-\theta_c\|\frac{u_{cm}}{2}\|\theta_t-\theta_c\|^2+\frac{k_t}{2u_{cm}}k_m^2+2M(t)+\frac{1}{2}\|\theta_t-\theta_c\|^2+\frac{1}{2}\|\theta_t-\theta_c\|^2 && (70)
\end{aligned}
$$

II. Bound on T_{32}:

$$
\begin{aligned}
T_{32} &= \frac{1}{N}\sum_{i=1}^{N}\Big\langle rM(\theta_t,\omega_t);G^i(\theta_t,\omega_t)\Big\rangle_{\mu}-\Big\langle G^i(\theta_t,\omega_t),G^i(\theta_t,\omega_t)\Big\rangle_{\mu}\\
&\le \underbrace{\frac{1}{N}\sum_{i=1}^{N}\|rM(\theta_t,\omega_t)\|_{\mu}\,\|G^i(\theta_t,\omega_t)-G^i(\theta_t,\omega_t)\|_{\mu}}_{T_{321}}+\underbrace{\frac{1}{N}\sum_{i=1}^{N}\|rM(\theta_t,\omega_t)\|_{\mu}\,\|G^i(\theta_t,\omega_t)-G^i(\theta_t,\omega_t)\|_{\mu}}_{T_{322}} \qquad (71)
\end{aligned}
$$

where the last inequality is by Hölder's inequlaity. For T_{321}, we have

$$
\begin{aligned}
T_{321} &= \|rM(\theta_t,\omega_t)-rM(0,0)\|_{\mu}\,\|\theta_t-0\|_{\mu} && \text{(Proposition B.1)}\\
&\le \frac{L}{l}\left[\|\theta_t\|\,\|\omega_t-\omega_c\|+\|\theta_t-\theta_c\|\,\|\omega_c\|\right] && (72)
\end{aligned}
$$

For T_{322}, we have

$$
\begin{aligned}
T_{322} &= \|G^i(\theta_t,\omega_t)-G^i(\theta_t,\omega_t)\|_{\mu}+\|G^i(\theta_t,\omega_t)-G^i(\theta_t,\omega_t)\|_{\mu}\\
&\le \frac{1}{l_{cs}}\|G^i(\theta_t,\omega_t)-G^i(\theta_t,\omega_t)\|_{\mu}+\|G^i(\theta_t,\omega_t)-G^i(\theta_t,\omega_t)\|_{\mu} && \text{(triangle inequality)}\\
&\le \frac{1}{l_{cs}}A_1\|\theta_t-\theta_c^{i}\|+c_k^{i}\|\theta_t-\theta_c\| && \text{(Assumptions 6.2 and 6.3)}\\
&\le \frac{A_1+1}{l_{cs}}\|\theta_t-\theta_c\|+\|\theta_t-\theta_c\|+\|\theta_c-\theta_t\| && (c_k^{i}\le 1;\ \text{triangle inequality})\\
&= \frac{A_1+1}{l_{cs}}\|\theta_t-\theta_c\|+\|\theta_t-\theta_c\|+\|\theta_t-\theta_t^{i}\| && (73)
\end{aligned}
$$

Substituting (72) and (73) in (71), we get

$$
\begin{aligned}
T_{32} &\le \frac{L(A_1+1)}{l^2}\left[\frac{1}{N}\sum_{i=1}^{N}3k_t\|\theta_c-\theta_t\|+k_t\|\theta_t-\theta_t^{i}\|+\|\theta_c-\theta_t\|\right]\\
&\quad+\frac{1}{N}\sum_{i=1}^{3N}\|\theta_t-\theta_c\|\,\|\theta_t-\theta_t^{i}\|+k_t\|\theta_c-\theta_t\|+\|\theta_t-\theta_t^{i}\|\Big]\\
&\le \frac{L(A_1+1)}{l_{cs}^2}\left[3k_tk_m+\frac{3}{2N}\sum_{i=1}^{N}\|\theta_t-\theta_c\|+k_t\|\theta_c-\theta_t\|\right]^2\\
&\quad+\frac{1}{2}k_t\|\theta_t-\theta_c\|+2\left[\frac{1}{2N}\sum_{i=1}^{3N}\|\theta_t-\theta_t^{i}\|+k_t\|\theta_c-\theta_t\|\right]^2 && \text{(Young's inequality)}
\end{aligned}
$$

$$\frac{L(A\,l+1)^2}{2}M(t)+\frac{3u_{cm}}{3}+\frac{2k_t^2}{2^{\frac{2}{3}}}{}_t\,k^2+\frac{3u_{cm}}{c}+\frac{3}{N}\frac{(i)^{22}}{2^{\frac{2}{3}}}+(\frac{i}{2})^2\frac{1}{i=1}\frac{L(A\,l+1)^2}{t}M(t)+{}_t{}^{-}{}_{2_3}+2k_t$$
$$t\;k_c+\frac{3u}{\frac{1}{2}^{\frac{3}{3}}_{cs}}+\frac{3}{3}\Big[\frac{3u_{cm}^2}{2}\qquad 2\qquad\frac{2}{2^{\frac{2}{2}}_{cm}}\qquad\frac{}{2}$$
$$_t+{}_t\;]: \tag{74}$$

III. Bound on T_{33}: Taking expectation on both sides of T_{33}, we have

$$E_t\;[T_{33}] =$$
$$rM(t\;);E_t\;[G(t\;;Y_t)]\quad G(t\;)$$
$$=\frac{1}{N}\sum_{i=1}^{N} rM(t\;);E_t\;[G^i(t\;;y_t)]\quad G^i(t\;) \qquad \text{(By (13))}$$
$$\frac{1}{l_{cs}}\frac{1}{N}\sum_{i=1}^{N}k\,rM(t\;)k^?\;E_{t_s}\;[G^i(^i\;;y^i)]\;{}_t\;G^i(^i_t\;)^c \qquad \text{(Assumption 6.1)}$$
$$\frac{1}{l}\frac{1}{N}\sum_{i=1}^{N}k\,rM(t\;)k^?\;m_1k^i\;{}_t\;k_c{}^{cs} \qquad \text{(assumption on)}$$
$$\frac{m_1}{l_{cs}}\frac{1}{N}\sum_{i=1}^{N}k\,rM(t\;)k^?\;k^i\;{}_s\;{}_t\;k_c \qquad \text{(By Proposition B.1 and (13))}$$
$$\frac{m_1L}{l^2}\frac{1}{N}\sum_{i=1}^{N}k\,{}_t\;k\;k_t\;k_c{}^{cs\;i} \qquad \text{(tringle inequality)}$$
$$\frac{m_1L}{l_{cs}^2}\frac{1}{N}\sum_{i=1}^{N}k\,{}_t\;k\;k_t\;k_c+k_t\;\;{}_t\;k_c\;\;{}^i \qquad \text{(Young's inequality)}$$
$$\#\quad\frac{m_1}{l_c{}^2_s}k_t\;k^2+\frac{1}{N^c}\sum_{i=1}^{N}\frac{1}{2}k_t\;k^2+\frac{1}{2}{}^t{}_c\;{}_t{}^{-}{}_c\;m_1L\,E_t\;{}^i\;3^{\;2}$$
$$k_tk^2+k_t\;{}_t\;k^2+\frac{1}{l_{cs}^2}{}^t{}_c\;{}_c\;\frac{}{2} \qquad ((a+b)^2\;2a^2+2b^2)$$
$$=\frac{m_1L}{l^2:_{cs}}E_t\;\;6u_{cm}M_2(t)+3\,k_t\;{}_t\;k^2+\frac{1}{c}\;{}_c\;\frac{}{2} \tag{75}$$

Substituting the bounds on $T_{31};T_{32};T_{33}$ from (70), (74), (75), in (66), we get the result. □

Proof of Lemma B.4. Denote $T_4 = \langle rM(t);G(t)\quad G(t)\rangle$. By the Cauchy–Schwarz inequality, we have

$$T_4\;\underbrace{krM(t)k}_{\{z\}}\;\underbrace{k\hat{G}_s(t)\quad G(t)k}_{T_{42}:\{z\}}{}_{T_{41}}{}_s \tag{76}$$

For T_{41} we have

$$krM(t)k_s\;{}^?{=}krM(t)\quad rM(0)k_s\;{}^?$$
$$L\,k_t{-}k_s \qquad \text{(Proposition B.1)}$$
$$L\frac{}{k_t}k_{c\,cs} \qquad \text{(By (13))}$$
$$Lu_{cm}\frac{}{k_t}k^{m\;cs} \qquad \text{(Proposition B.1)}$$
$$=\frac{Lu_{cm}}{l_{cs}}p\;\underline{2M(t):} \tag{77}$$

For T_{42} we have

$$\|G(\theta_t) - G(\theta_t)\|_s = \Big\| \frac{1}{N}\sum_{i=1}^N G^i(\theta^i)_t - G^i(\theta_t) - \frac{1}{N}\sum_{i=1}^N G^i(\theta^i)_t - G^i(\theta_t)\Big\|_{cs} \quad \text{(Jensen's inequality)}$$

$$\frac{1}{N}\sum_{i=1}^N \big\| G^i(\theta^i)_t - G^i(\theta_t)\big\|_{cs} \le \sum_{i=1}^N \frac{\|\theta^i_t - \theta_t\|_{cs}}{N} \quad (78)$$

where (78) follows from Assumption 6.2. Combining (77) and (78), for an arbitrary $\epsilon_4 > 0$, we get

$$T_4 \le \frac{L u_{cm}\sqrt{p}}{\sqrt{2}M(t)_c}\sum_{i=1}^N \frac{\|\theta^i_t - \theta_t\|_{cs}}{N} \le \frac{1}{N}\sum_{i=1}^N \epsilon_4 2M(t) + \frac{1}{\epsilon_4}\frac{L u_{cm}}{\sqrt{2}}\|\theta^i_t - \theta_t\|^2_{cs} \quad (c_1)$$

$$\le \epsilon_4 2M(t) + \frac{1}{\epsilon_4}\frac{L^2 u^2_{cm}}{2}\|\theta^i_t - \theta_t\|^2_{cs} \quad \text{(Young's inequality)}$$

$$= \epsilon_4 2M(t) + \frac{L^2 u^2_{cm}}{2\epsilon_4}\alpha^2_t. \quad (79)$$

□

Proof of Lemma B.5. Denote $T_5 = \|G(t;Y_t)\theta_t + b(Y_t)\|^2_s$. We have

$$T_5 = \|G(t;Y_t) + G(t;Y_t) - G(t;Y_t)\theta_t + b(Y_t)\|^2_s \le 3\|G(t;Y_t)\theta_t\|^2_s + 3\|G(t;Y_t) - G(t;Y_t)\|^2_s\|\theta_t\|^2 + 3\|b(Y_t)\|^2_s : T_{51} + T_{52} \quad (80)$$

For T_{51} we have

$$T_{51} \le \|G(t;Y_t)\|^2_{cs}(\|\theta_t\|_c + \|\theta_t\|_c)^2 \le \|G(t;Y_t)\|^2_{cs}(\|\theta_t\|_c + \|\theta_t\|_c)^2 \quad \text{(By (13))}$$

$$\le (A_2\|\theta_t\|_c + \|\theta_t\|_c)^2 = 2(A_2 + 1)^2 u^2_m M(t)_c : \quad \text{(Assumption 6.3)} \quad (81)$$

For T_{52} we have

$$T_{52} = \Big\| \frac{1}{N}\sum_{i=1}^N (G^i(\theta^i;y^i) - G^i(\theta_t;y^i))_t - \frac{1}{N}\sum_{i=1}^N G^i(\theta^i;y^i)_t - G^i(\theta_t;y^i)\Big\|^2_s \quad (\|\cdot\|_s^2 \text{ is convex})$$

$$\frac{1}{N}\sum_{i=1}^N \|G^i(\theta^i;y^i)_t - G^i(\theta_t;y^i)\|^2_{cs} \le \frac{1}{N}\sum_{i=1}^N \|\theta^i_t\|^2_c \quad \text{(By (13))}$$

$$\frac{A_1^2}{2}\frac{1}{N}\sum_{i=1}^N \|\theta^i_t\|^2_c \le \frac{A_1^2}{2}\|\theta_t\|^2_{cs} \quad \text{(Assumption 6.3)}$$

$$= \frac{A_1^2}{2}\alpha^2_t : \quad (82)$$

Using the bounds in (81), (82), we get

$$T_5 \le \frac{6(A_2+1)^2 u^2_{cm}}{\sqrt{2}}M(t) + \frac{3A_1^2}{2}\alpha^2_t + 3\|b(Y_t)\|^2_s \quad (83)$$

where the last inequality is by the assumption on .

□

Proof of Lemma B.6. Define $\bar{\theta}_t = \frac{1}{N}\sum_{i=1}^N \theta_t^i$. By the update rule of Algorithm 4, if $t+1 \bmod K = 0$, we have

$$\begin{aligned}
\bar{\theta}_{t+1} &= \frac{1}{N}\sum_{i=1}^N \theta_{t+1}^i = \frac{1}{N}\sum_{i=1}^N \Big(\theta_t^i + \alpha_t\big(G^i(\theta_t^i, y_t^i)\theta_t^i + b^i(y_t^i)\big)\Big) \\
&= \bar{\theta}_t + \frac{\alpha_t}{N}\sum_{i=1}^N G^i(\theta_t^i, y_t^i)\theta_t^i + \alpha_t \bar{b}_t \\
&\le \bar{\theta}_t + \frac{\alpha_t}{N}\sum_{i=1}^N \|G^i(\theta_t^i, y_t^i)\|\,\|\theta_t^i\| + \alpha_t\|\bar{b}_t\| && \text{(triangle inequality)} \\
&\le \bar{\theta}_t + \frac{\alpha_t}{N}\sum_{i=1}^N (A_2+1)\|\theta_t^i\| + B\alpha_t && \text{(Assumption 6.3)} \\
&= (1+(A_2+1)\alpha_t)\bar{\theta}_t + B\alpha_t.
\end{aligned} \tag{84}$$

Furthermore, if $t+1 \bmod K = 0$, we have $\theta_{t+1}^j = \frac{1}{N}\sum_{i=1}^N (\theta_t^i + \alpha_t(G^j(\theta_t^i, y_t^j)\theta_t^i + b^j(y_t^j)))$, and hence

$$\begin{aligned}
\bar{\theta}_{t+1} &= \frac{1}{N}\sum_{i=1}^N \theta_{t+1}^i = \frac{1}{N}\sum_{i=1}^N \frac{1}{N}\sum_{j=1}^N \Big(\theta_t^i + \alpha_t\big(G^j(\theta_t^i, y_t^j)\theta_t^i + b^j(y_t^j)\big)\Big) \\
&\le \bar{\theta}_t + \frac{\alpha_t}{N}\sum_{i=1}^N \|G^i(\theta_t^i, y_t^i)\|\,\|\theta_t^i\| + \alpha_t\|\bar{b}_t\| && \text{(triangle inequality)}
\end{aligned}$$

and the same bound as in (84) holds. By recursive application of (84), we get

$$\begin{aligned}
\|\bar{\theta}_{t+1}\| &\le (1+(A_2+1)\alpha)^{t+1}\|\theta_0\| + B\sum_{\tau=0}^{t}(1+(A_2+1)\alpha)^{\tau} \\
&= (1+(A_2+1)\alpha)^{t+1}\|\theta_0\| + B\,\frac{(1+(A_2+1)\alpha)^{t+1}-1}{A_2+1}.
\end{aligned} \tag{85}$$

Notice that for $x \le \log 2$, we have $(1+x)^{t+1} \le 1+2x(t+1)$. If $0 \le t \le 2\tau-1$, by the assumption on α, we have $(1+(A_2+1)\alpha)^{t+1} \le (1+(A_2+1)\alpha)^{2\tau} \le 1+4(A_2+1)\tau\alpha \le 2$. Hence, we have

$$\|\bar{\theta}_t\| \le 2\|\theta_0\| + B\,\frac{4(A_2+1)}{A_2+1} = 2\|\theta_0\| + 4B. \tag{86}$$

Furthermore, we have

$$\begin{aligned}
\|\theta_{t+1}^i - \bar{\theta}_t\| &= \|G(\theta_t^i, y_t^i)\theta_t^i + b(y_t^i)\| \le \|G(\theta_t^i, y_t^i)\|\,\|\theta_t^i\| + \|b(y_t^i)\| \\
&\le \|\theta_t^i - \bar{\theta}_t\| + \|\bar{\theta}_t\| + \|b(y_t^i)\| && \text{(triangle inequality)} \\
&= \frac{1}{N}\sum_{j=1}^N \big(\|G^j(\theta_t^i, y_t^j)\theta_t^i\| - \|G^j(\theta_t^i, y_t^j)\bar{\theta}_t\|\big) + \|\bar{\theta}_t\| + \|b(y_t^i)\| \\
&\le \frac{1}{N}\sum_{j=1}^N \|G^j(\theta_t^i, y_t^j)\|\,\|\theta_t^i - \bar{\theta}_t\| + \|\bar{\theta}_t\| + B \\
&\le \frac{1}{N}\sum_{j=1}^N (A_2+1)\|\theta_t^i - \bar{\theta}_t\| + \|\bar{\theta}_t\| + B && \text{(convexity of norm)} \\
&= (A_2+1)\|\theta_t^i - \bar{\theta}_t\| + \|\bar{\theta}_t\| + B.
\end{aligned}$$

(Assumption 6.3)

(87)

Suppose $0 \le t \le 2\tau$. We have

$$\|\theta_t^i - \bar{\theta}_t\| \le \Big\|\sum_{k=0}^{t-1}(\theta_{k+1}^i - \theta_k^i)\Big\| \le \sum_{k=0}^{t-1}\|\theta_{k+1}^i - \theta_k^i\| \quad \text{(triangle inequality)}$$

$$\left[\sum_{k=0}^{t-1}(A_2+1)^k\right]+B \qquad \text{(By (87))}$$

$$Bt+(A_2+1)\sum_{k=0}^{t-1}(2_0+4B) \qquad \text{(By (86))}$$

$$2(B+(A_2+1)(2_0+4B))$$

$$C_1\frac{1}{}\quad B+(A_2+1)\ _0+\frac{B}{2C_1}\quad : \qquad (t\ 2) \tag{88}$$

Furthermore, by Proposition B.1, we have $M(_t)=\ _2{}^1k_tk_m$, and hence for any $0\ t\ 2$ we have

$$\begin{aligned}
&M(_t)\quad \frac{1}{2l^2}k_tk_c{}^{cm}{}^2\\
&=\frac{1}{2l^2}(k_t\quad _0+\ _0k_c)^2{}^{cm}\\
&2l_{cm}\frac{1}{2}(k_t\quad _0k_c+k_0k_c)^2 \qquad \text{(triangle inequality)}\\
&l_{cm}^{\frac{1}{2}}(k_t\quad _0k_c+k_0{}^1k^2)\\
&l_{cm^2}\frac{1}{}\quad \frac{1}{C_1^2}\quad B+(A_2+1)\quad k_0k_c+\frac{B}{2C_1}\quad ^2\quad +k_0k_c{}^2\quad ;
\end{aligned}$$

which proves the first claim.

Next we prove the second claim. By the update rule in 30, we have

$$\begin{aligned}
k_{l+1}\quad {}_lk_c=&{}^{22}kG(_l;Y_l)\quad _l+b(Y_l)k^2\\
&3^2kG(_l;Y_l)\quad {}_lk^2+3^2kb(Y_l)k^2+3^2k_cG(_l;Y_l)\quad G(_l;Y_l)k^2\ 6^2(kG(_l;Y_l)k^2+\\
&k_lk_c)+3^2kb(Y_l)k^2+3^2k_c{}^2G(_l;Y_l)\quad G(_l;Y_l)k^2\\
=&6^2\ @\ \frac{1}{N}\sum_{i=1}^{N}G^i(_l;y^i)_l\ +k_lk^2\quad _cA+3^2kb(Y_l)k^2+3^2kG(_l;Y_l)\quad G(_l;Y_l)k^2\ 0\\
&6^2\ @\ \frac{1}{N}\sum_{i=1}^{N}G^i(_l;y^i)_{c\,l}\quad \Big)^2+k_lk^2{}_cA+3^2kb(Y_l)k^2+{}_c3^2kG(_l;Y_l)\quad G(_l;Y_l)k^2 \qquad \text{(convexity of norm)}\\
&6^2\ @\ \sum_{i=1}^{N}A_2k_lk_c\quad \Big)^2+k_lk^2{}_cA+3^2kb(Y_l)k^2+{}_c3^2kG(_l;Y_l)\quad G(_l;Y_l)k^2 \qquad \text{(Assumption 6.3)}\\
=&6^2(A^2+\ _2{}^1)k_lk^2+{}_c3^2kb(Y_l)k^2+3^2{}_c^1kG(_l;Y_l)\quad G(_l;Y_l)k^2\\
&6^2(A^2+\ _2{}^1\overline{N}k_lk^2+3{}_c^2kb(Y_l)k^2+3^2{}_cA^{22};\qquad {}_1\ _l
\end{aligned} \tag{89}$$

where (89) follows from Lemma B.10. Taking square root on both sides, we get

$$k_{l+1}\quad {}_lk_c\ 3\quad \sqrt{A_2+1k_lk_c+2kb(Y_l)k_c+2A_1}_l: \tag{90}$$

Combining the above inequality with the fact that $k_{l+1}k_c\quad k_lk_c\ \ k_{l+1}\quad {}_lk_c$, we get

$$k_{l+1}k_c\ (1+3\quad \sqrt{A_2+1)k_lk_c+2kb(Y_l)k_c+2A_1}_l$$

$$= (1 + C_1)k_l k_c + 2kb(Y_l)k_c + 2A_1 l; \tag{91}$$

where we denote $C_1 = 3\sqrt{A_2^2 + 1}$. Assuming $t \quad l \quad t$, and taking expectation on both sides, we have

$$\begin{aligned}
&E_{t\ 2}[k_{l+1}k_c] \quad (1 + C_1)E_{t\ 2}[k_l k_c] + 2E_{t\ 2}[kb(Y_l)k_c] + 2A_1 E_{t\ 2}[l] \\
&\quad (1 + C_1)E_{t\ 2}[k_l k_c] + 2\frac{u_{cD}}{l_{cD}} + \frac{2m_2}{B}\Big)^{l\ t+2} + 2A_1 E_{t\ 2}[l]_{cD} && \text{(Lemma B.9)} \\
&\quad (1 + C_1)E_{t\ 2}[k_l k_c] + 2\frac{u_{cD}}{l_{cD}} + \frac{2m_2}{B}\Big)^{l\ t+} + 2A_1 E_{t\ 2}[l]_{cD} && \text{(Assumption on)} \\
&= (1 + C_1)E_{t\ 2}[k_l k_c] + c_t(l) + 2A_1 E_{t\ 2}[l]; && (92)
\end{aligned}$$

where $c_t(l) = 2\frac{u_{cD}}{l_{cD}} \frac{B}{N} + 2m_2)^{l\ t+}$. By applying this inequality recursively, we have

$$\begin{aligned}
&E_{t\ 2}[k_{l+1}k_c] \quad (1 + C_1)E_{t\ 2}[k_l k_c] + c_t(l) + 2A_1 E_{t\ 2}[l] \\
&\quad (1 + C_1)\left[(1 + C_1)E_{t\ 2}[k_{l\ 1}k_c] + c_t(l \quad 1) + 2A_1 E_{t\ 2}[l\ 1]\right] + c_t(l) + 2A_1 E_{t\ 2}[l] \quad (1 + C_1)^{l+1\ t+} \\
&E_{t\ 2}[k_t \quad k_c] + \sum_{k=t}^{l} (1 + C_1)^{l\ k} c_t(k) + 2A_1 E_{t\ 2}\Bigg[\sum_{=t}^{l} (1 + C_1)^{l\ k}{}_k\Bigg] \\
&(1 + C_1)^{+1} E_{t\ 2}k_t \quad k_c + \underbrace{\sum_{=t}^{k} (1 + C_1)^{t\ k} c_t(k)}_{T_1} + 2A_1 E_{t\ 2}\underbrace{\sum_{k=t}^{k} (1 + C_1)^{t\ k}{}_k}_{T_2} :
\end{aligned} \tag{93}$$

We study T_1 and T_2 in (93) separately. For T_1 we have

$$\begin{aligned}
T_1 &= 2\frac{u_{cD}}{l_{cD}} \sum_{k=0}^{} (1 + C_1)^{k} \frac{B}{N} + 2m_2{}^{k} \\
&= 2\frac{u_{cD}}{l_{cD}} \frac{}{N}\frac{(1 + C_1)^{+1} \quad 1}{C_1} + 2m_2(1 + C_1)^{k} \\
&\quad 2\frac{u_{cD}}{l_{cD}} \frac{}{N}\frac{(1 + C_1)^{+1} \quad 1}{C_1} + 2m_2(1 + C_1)^{k} && (> 0)\ 2\frac{u_{cD}}{l_{cD}} \frac{B}{N}\frac{(1 + C_1)^{+1} \quad 1}{C_1} + 2m_2(1 + C_1)\frac{1}{1} : && (94)
\end{aligned}$$

Notice that for $x \quad \log 2$, we have $(1 + x)^{+1} \quad 1 + 2x(+ 1)$. By the assumption on , we have $(1 + C_1)^{+1} \quad 1 + 2C_1(+ 1) \quad 1 + 4C_1 \quad 2$ and $(1 + C_1) \quad 1 + 2C_1 \quad 1 + 1=2 \quad 2$. Hence, we have

$$T_1 \quad 2\frac{u_{cD}}{l_{cD}} \frac{B}{N} 2(+ 1) + \frac{4m_2}{1} : \tag{95}$$

Furthermore, for the term T_2 we have

$$\begin{aligned}
T_2 &= \sum_{k=0}^{} (1 + C_1)^{k}{}_{t+k} \sum_{k=0}^{} (1 + C_1)_{t+k} && \text{(due to } > 0) \\
&\sum_{k=0}^{} (1 + 2C_1)_{t+k} \quad 2 \sum_{k=0}^{} {}_{t\ k} : && (96)
\end{aligned}$$

Subtituting (95), (96) in (93), for every $t \quad l \quad t$, we get

$$E_{t\ 2}[k_l k_c] \quad 2E_{t\ 2}[k_t \quad k_c] + \frac{2u_{cD}}{l} \frac{4B}{} + \frac{4m_2}{1} + 4A_1 \frac{}{1} {}_{t\ k} : \tag{97}$$

But we have

$$\mathbb{E}_{t-2}\|k_t - k_c\| \le \mathbb{E}_{t-2}\Big[\sum_{i=t}^{t-1}\|k_{i+1} - k_c\|\Big] \quad \text{(triangle inequality)}$$

$$\le \sum_{i=t}^{t-1}\mathbb{E}_{t-2}[C_1\|k_i - k_c\| + 2kb(Y_i)\|k_c\| + 2A_{1i}] \quad \text{(by (90))}$$

$$\le C_1\Big(2\mathbb{E}_{t-2}[\|k_t - k_c\|] + \frac{2u_{cD}}{l}\Big(\frac{4B}{p} + \frac{4m_2}{N}\Big) + 4A_1\frac{1}{1}\Big)\sum_{j=0}^{t-1}\mathbb{E}_{t-2}[\|t-j\|]\Big) \quad \text{(by (97))}$$

$$+ 2\sum_{i=t}^{t-1}\frac{u_{cD}}{l}\Big(\frac{pB}{N} + 2m_2\Big)^{i-t+2} \quad \text{(Lemma B.9)}$$

$$+ 2A_1\sum_{i=t}^{t-1}\mathbb{E}_{t-2}[\|i\|]$$

$$\le C_1\Big(2\mathbb{E}_{t-2}[\|k_t - k_c\|] + \frac{2u_{cD}}{l}\Big(\frac{4B}{p} + \frac{4m_2}{N}\Big) + 4A_1\frac{1}{1}\Big)\sum_{j=0}^{t-1}$$

$$+ 2\frac{l_{cD}u}{cD}\Big(\frac{B}{N} + 4^2\frac{l_{cD}u}{cD}m_2\Big)\frac{1}{1} + 2A_1\sum_{i=t}^{t-1}\mathbb{E}_{t-2}[\|i\|] \quad \text{(assumption on)}$$

$$2C_1\mathbb{E}_{t-2}[\|k_t - k_c\|] + \frac{u_{cD}}{l_{cD}}\Big(\frac{B}{N} + 8C_1p^2 + 2 + \frac{8m_2}{C}\Big)\frac{1^2}{1} + \frac{u_{cD}4m_2^2}{1}$$

$$+ (4A_1^2C_1 + 2A_1)\sum_{i=t}^{t-1}\mathbb{E}_{t-2}[\|i\|] \quad (\|i\| \le 0)$$

$$2C_1\mathbb{E}_{t-2}[\|k_t - k_c\|] + 4\frac{u_{cD}}{l}\Big(\frac{B}{N} + \frac{4m_2}{p}\Big)^2(1 + 2C_1) + 3A_1\sum_{i=t}^{t-1}\mathbb{E}_{t-2}[\|i\|]: \quad (98)$$

Furthermore, by triangle inequality, we have $\|k_t - k_c\| \le \|k_t - k_{t-1}\| + \|k_{t-1} - k_c\| + \|k_t - k_c\|$. By assumption on , we have $2C_1\|k_t - k_c\| \le 2C_1\|k_t - k_{t-1}\| + 2C_1\|k_{t-1} - k_c\| + 2C_1\|k_t - k_c\| \le 0.5\|k_t - k_{t-1}\| + 2C_1\|k_{t-1} - k_c\| + 2C_1\|k_t - k_c\|$. By taking expectation on both sides, and substituting it in (98), we get (34).

Proof of Lemma B.7. From (91) we have

$$\|k_{l+1} - k_c\|^2 \le (1 + C_1)^2\|k_l - k_c\|^2 + 4^2kb(Y_l)\|k_c\|^2 + 4^2A_{1l}^2$$

$$+ \underbrace{4(1 + C_1)\|k_l - k_c\|kb(Y_l)\|k_c\|}_{T_1} + \underbrace{4A_1(1 + C_1)\|k_l - k_c\| + 8^2A_{1l}kb(Y_l)\|k_c\|}_{T_2} : \quad (99)$$

For T_1 we have

$$T_1 = 2\sqrt{(1 + C_1)}\|k_l - k_c\| \; 2\sqrt{(1 + C_1)}kb(Y_l)\|k_c\|$$

$$\le 2(1 + C_1)\|k_l - k_c\| + 2(1 + C_1)kb(Y_l)\|k_c\|^2 \quad (ab \le \tfrac{1}{2}a^2 + \tfrac{1}{2}b^2)$$

$$\le 4\|k_l - k_c\| + 4kb(Y_l)\|k_c\|; \quad (100)$$

where the last inequality is by the assumption on . Analogously for T_2 we have

$$T_2 = 2\sqrt{(1 + C_1)}\|k_l - k_c\| \; 2A_1\sqrt{(1 + C_1)}$$

$$\le 4\|k_l - k_c\|^2 + 4A_{1l}^2: \quad (101)$$

For T_3 we have

$$T_3 = 2\alpha_t b(Y_t)\alpha_c \, 4A_1 t \, 2^2 \alpha_t b(Y_t)\alpha^2 + \alpha_c 8^2 A^2 2 : \quad (102)$$

Combining the bounds in (100), (101), and (102), and noting that $(1 + C_1)^2 \le 1 + 3C_1$, from (99) we have

$$\begin{aligned}
\alpha_{t+1}\alpha_c^2 &\le (1 + 3C_1)\alpha_t\alpha_c + 4\alpha^2 2\alpha_t b(Y_t)\alpha_c + 4\alpha^2 A^2 1 t \\
&\quad + 4\alpha_t\alpha^2 + \alpha_c 4\alpha_t b(Y_t)\alpha^2 + 4\alpha_t\alpha^2 + 4A^2 \alpha_c 2 + 2^2 \alpha_t b(Y_t)\alpha^2 + 8^2 A^2 2 \\
&\le (1 + (3C_1 + 8))\alpha_t\alpha^2 + (6^2 + 4)\alpha_t b(Y_t)\alpha^2 + A^2(12^2 + 4)^2 \\
&\le (1 + C_2)\alpha_t\alpha_c + 10\alpha_t b(Y_t)\alpha_c + 16\alpha^2 A^2 1 t : \quad (103)
\end{aligned}$$

where $C_2 = 3C_1 + 8$. Taking expectation on both sides, we have

$$\begin{aligned}
\mathbb{E}\alpha_{t+1}\alpha^2 &\le (1 + C_2)\mathbb{E}\alpha_t\alpha^2 + 10\mathbb{E}\alpha_t b(Y_t)\alpha^2 + 16A^2\mathbb{E}\alpha_t [^2] \\
&\le (1 + C_2)\mathbb{E}\alpha_t\alpha_c + 10\frac{\alpha_t u^2_{cD}}{l^2}\left(\frac{B^2}{N} + 2m^2 2^{2(t+2)}\right) + 16A_1\mathbb{E}\alpha_t [t]^2_{cD} \quad \text{(Lemma B.9)} \\
&= (1 + C_2)\mathbb{E}\alpha_t\alpha^2 + \alpha_c(t) + 16A^2\mathbb{E}\alpha_t [^2]; \quad (104)
\end{aligned}$$

where $\alpha_t(t) = 10\frac{\alpha_t u^2_{cD}}{l^2}\left(\frac{B^2}{N} + 2m^2 2^{2(t+2)}\right)$. Hence, for any $t \le t$, we have

$$\begin{aligned}
\mathbb{E}\alpha_{t+1}\alpha^2 &\le (1 + C_2)(1 + C_2)\mathbb{E}\alpha_t\alpha_c + \alpha_t(t-1) + 16A_1\mathbb{E}\alpha_t [t-1] \\
&\quad + \alpha_t(t) + 16A_1\mathbb{E}\alpha_t [t] \\
&\le (1 + C_2)^{t+1}\mathbb{E}\alpha_t\alpha^2 + \sum_{i=t}^{t} (1 + C_2)^{t-i}\alpha_t(i) + 16A^2\mathbb{E}\alpha_t [t] \\
&\quad + \sum_{i=t}^{t-1} (1 + C_2)^{t-i}\alpha_t(i) + 16A_1\mathbb{E}\alpha_t [i] \\
&\le (1 + C_2)^{t+1}\mathbb{E}\alpha_t\alpha^2 + \sum_{i=t}^{t} (1 + C_2)^{t-i}\alpha_t(i) + 16A_1\mathbb{E}\alpha_t [t] \\
&\quad + \sum_{i=t}^{t-1} (1 + C_2)^{t-i}\alpha_t(i) + 16A_1\mathbb{E}\alpha_t [i] : \quad (105)
\end{aligned}$$

T_4 T_5

For the term T_4 we have

$$\begin{aligned}
T_4 &= 10\frac{u^2_{cD}}{l^2}\sum_{i=0} (1 + C_2)^{-i}\left(\frac{B^2}{N} + 2m^2 2^{2i}\right)_{cD}^2 \\
&= 10\frac{u^2_{cD}B^2}{l^2_{cD}N}\frac{(1 + C_2)^{+1} - 1}{C_2} + 20\frac{u^2_D}{l^2_{cD}}m_2^2(1 + C_2)\sum_{i=0}\frac{i}{1+C_2} \\
&\le 10\frac{u^2_{cD}B^2}{l^2_{cD}N}\frac{(1 + C_2)^{+1} - 1}{C_2} + 20\frac{u^2_D}{l^2_{cD}}m_2^2(1 + C_2)\sum_{i=0} 2^i 10\frac{u^2_D B^2}{l^2_{cD}N}\frac{(1 + C_2)^{+1} - 1}{C_2} + 20\frac{u^2_D}{l^2_{cD}}m_2^2(1 + C_2)\frac{1}{1 - 2} : \quad (C_2 \ge 0) \quad (106)
\end{aligned}$$

By the same argument as in Lemma B.6, and by the assumption on , we have $(1 + C_2)^{+1} \le 1 + 2C_2(+ 1) \le 1 + 4C_2 \le 2$ and $(1 + C_2) \le 1 + 2C_2 \le 1 + 1=2 \le 2$. Hence, we have

$$T_4 \le 40\frac{u^2_{cD}}{l^2_{cD}}\left(\frac{B^2}{N} + \frac{m_2^2}{1 - 2}\right) : \quad (107)$$

Furthermore, for T_5, we have

$$T_5 = \sum_{i=0} (1 + C_2)^{-i}\mathbb{E}\alpha_t [t + i]^2$$

$$\sum_{i=0}^{\infty}(1+C_2)E_{t-2}[\cdot_{t+i}]\cdot_{i=0}^{2} \qquad (C_2>0)$$

$$\sum_{i=0}^{\infty}(1+2C_2)E_{t-2}[\cdot_{t+i}]\cdot_{i=0}^{2}$$

$$2\sum_{i=0}^{\infty}E_{t-2}[\cdot_{t-i}]:\cdot_{i=0}^{2} \qquad \text{(assumption on } \cdot\text{)}$$

(108)

Combining the bounds on T_4 (107) and T_5 (108) in (105) we have for any $t \geq t_0$,

$$E_{t-2}\|k_{t+1}-k_c\|^2 \leq E_{t-2}\|k_t-k_c\|^2 + 40\cdot\frac{2c_D}{l}\left(\frac{u^2}{2c_D}+\frac{B^2}{N}+m^2\right)\cdot^2 + \frac{1}{2}+32A_1\sum_{i=0}^{\infty}E_{t-2}[\cdot_{t-i}]:\cdot^2 \tag{109}$$

Furthermore, we have

$$\|k_t-k_c\|^2 \leq \left(\sum_{i=t_0}^{t-1}\|k_{i+1}-k_i\|+\|k_{t_0}-k_c\|\right)^2 \qquad \text{(triangle inequality)}$$

$$\leq 2\left(\sum_{i=t_0}^{t-1}\|k_{i+1}-k_i\|\right)^2+2\|k_{t_0}-k_c\|^2$$

$$\leq 2(t-t_0)\sum_{i=t_0}^{t-1}\|k_{i+1}-k_i\|^2+2\|k_{t_0}-k_c\|^2$$

$$\leq 2\cdot^2C^2\sum_{i=t_0}^{t-1}\|k_i-k_c\|^2+3\cdot^2\sum_{i=t_0}^{t-1}\|b(y_i)\|^2+3\cdot^2A^2\sum_{i=t_0}^{t-1}\|\cdot_{i-c}\|^2+2\|k_{t_0}-k_c\|^2 \qquad \text{(from (89))}$$

Taking expectation on both sides, and using the bounds in Lemma B.9 and (109), we get

$$E_{t-2}\|k_{t+1}-k_c\|^2 \leq 2\cdot^2C^2\sum_{i=t_0}^{t-1}\left[2E_{t-2}\|k_t-k_c\|^2+40\cdot\frac{2c_D}{l^2}\left(\frac{u^2}{2c_D}+\frac{B^2}{N}+m^2\right)\cdot^2+\frac{3}{2}+32A^2\sum_{i=0}^{\infty}E_{t-2}[\cdot_{t-i}]\cdot^2\right]$$

$$+3\cdot^2\sum_{i=t_0}^{t-1}\left(\frac{u^2_{c_D}}{2c_D}+\frac{B^2}{N}\right)\cdot^2+2m^2\cdot^{2(i-t_0+2)}+3\cdot^2A^2\sum_{i=t_0}^{t-1}E_{t-2}\|\cdot_{i-c}\|^2$$

$$\leq 2\cdot^2C^2\left[2E_{t-2}\|k_t-k_c\|^2+40\cdot\frac{2c_D}{l^2}\left(\frac{u^2_{c_D}}{2}+\frac{B^2}{N}+\frac{m^2\cdot^3}{1-\cdot^2}\right)+3\cdot^2\left(\frac{u^2_{c_D}}{2}+\frac{B^2}{N}+\frac{2m^2\cdot^2}{1-\cdot^2}\right)\right.$$

$$\left.+2\cdot^2A^2(3+32C_1)\sum_{i=0}^{\infty}E_{t-2}\|\cdot_{t-i}\|^2\right]. \qquad (*)$$

(110)

Furthermore, by triangle inequality, we have $\|k_t-k_c\|^2 \leq \|k_t-k_{t_0}+k_{t_0}-k_c\|^2 \leq (\|k_t-k_{t_0}\|+\|k_{t_0}-k_c\|)^2 \leq 2\|k_t-k_{t_0}\|^2+2\|k_{t_0}-k_c\|^2$. By assumption on $\cdot$, we have $2\cdot^2C^2\|k_t-k_{t_0}\|^2 \leq 4\cdot^2C^2\|k_{t-1}-k_{t_0}\|^2+4\cdot^2C^2\|k_t-k_{t-1}\|^2 \leq 4\cdot^2C^2\|k_{t-1}-k_{t_0}\|^2+0.5\|k_t-k_{t-1}\|^2 \leq 4\cdot^2C^2\|k_{t-1}-k_{t_0}\|^2+\|k_t-k_c\|^2+\|k_{t-1}-k_c\|^2$.

Proof of Lemma B.8. For $sK+1 \leq t \leq (s+1)K-1$, where $s=\lfloor t/K\rfloor$, □

$$\left\|\frac{1}{N}\sum_{i=1}^{N}k^i_t-k_t\right\|^2=\left\|\frac{1}{N}\sum_{i=1}^{N}k^i_t-k_c\right\|^2$$

$$=\left\|\frac{1}{N}\sum_{i=1}^{N}\left(k^i_{sK}+\sum_{t_0=sK}^{t-1}\left(G^i(\cdot^i_{t_0};y^i_{t_0})\cdot^i_{t_0}+b^i(y^i_{t_0})\right)\right)-\left(k_{sK}+\sum_{t_0=sK}^{t-1}\left(G(\cdot_{t_0};y_{t_0})\cdot_{t_0}+b(y_{t_0})\right)\right)\right\|^2$$

$$\leq\frac{2}{N}\sum_{i=1}^{N}\left\|k^i_{sK}-k_{sK}+\sum_{t_0=sK}^{t-1}\left(b^i(y^i_{t_0})-b(y_{t_0})\right)\right\|^2+\frac{2}{N}\sum_{i=1}^{N}\left\|\sum_{t_0=sK}^{t-1}\left(G^i(\cdot^i_{t_0};y^i_{t_0})\cdot^i_{t_0}-(G(\cdot_{t_0};y_{t_0})\cdot_{t_0}\right)\right\|^2$$

$$\leq 2\cdot^{2sK}\|k^i_{sK}-k_{sK}\|^2+2(t-sK)\sum_{t_0=sK}^{t-1}\|b^i(y^i_{t_0})-b(y_{t_0})\|^2+2(t-sK)\sum_{t_0=sK}^{t-1}\|G^i(\cdot^i_{t_0};y^i_{t_0})\cdot^i_{t_0}-(G(\cdot_{t_0};y_{t_0})\cdot_{t_0}\|^2$$

$$\leq 2\cdot^{2sK}\|k^i_{sK}-k_{sK}\|^2+4K\sum_{t_0=sK}^{t-1}\|b^i(y^i_{t_0})-b(y_{t_0})\|^2+4K\sum_{t_0=sK}^{t-1}\|G^i(\cdot^i_{t_0};y^i_{t_0})\cdot^i_{t_0}-(G(\cdot_{t_0};y_{t_0})\cdot_{t_0}\|^2$$

(* $\cdot^{sK}=\cdot_{sK}$ and Young's inequality)

$$\frac{2u_{c2}^2}{N}\sum_{i=1}^{N} \left[\sum_{t'=sK}^{t-1} b^i(y^i_{t'}) - b(y_{t'})^2 + 2 \sum_{\substack{t',t''=sK \\ t'\neq t''}}^{t-1} b^i(y^i_{t'}) - b(y_{t'}); b^i(y^i_{t''}) - b(y_{t''}) \right]$$

$$+ \frac{4^2}{N}(t - sK) \sum_{t'=sK}^{t-1}\sum_{i=1}^{N} \left[u_{c2}^2 \, G^i(\theta_{t'}; y^i_{t'}) - G(\theta_{t'}; y_{t'})^2 + \ldots \right]$$

(Triangle inequality, and $\|\cdot\|_c \leq u_{c2} \|\cdot\|_2$)

$$\frac{2u_{c2}^2}{N}\sum_{i=1}^{N} \left[\sum_{t'=sK}^{t-1} \frac{1}{l_{c2}^2} b^i(y^i_{t'})_c^2 + \sum_{\substack{t',t''=sK \\ t'\neq t''}}^{t-1} b^i(y^i_{t'}) - b(y_{t'}); b^i(y^i_{t''}) - b(y_{t''}) \right]$$

$$+ \frac{4^2}{N}(t - sK) \sum_{t'=sK}^{t-1}\sum_{i=1}^{N} \frac{u_{c2}^2}{l_{c2}^2} G^i(\theta_{t'}; y_{t'}) \ldots_c^2 : \qquad (l_{c2}\|\cdot\|_2 \leq \|\cdot\|_c \text{ and } Var(X) \leq E[X^2])$$

Taking expectation

$$E[\ldots_t] \leq \frac{2u_{c2}^2}{l_{c2}^2}(t - sK)B^2 + \frac{4u_{c2}^2}{N} \sum_{i=1}^{N} \sum_{\substack{t',t''=sK \\ t'<t''}}^{t-1} E\, b^i(y^i_{t'}) - b(y_{t'}); E_{t'}\, b^i(y^i_{t''}) - b(y_{t''})_c$$

$$+ \frac{4^2}{N}(t - sK) \sum_{t'=sK}^{t-1}\sum_{i=1}^{N} E\left\{ \frac{2u_{c2}^2}{l_{c2}^2} G^i(\theta_{t'}; y_{t'}) - G^i(\theta_{t'}; y^i_{t'}) + G(\theta_{t'}; y_{t'})_c^2 + \ldots_c^2 \right\}$$

(using Assumption 6.3)

$$\leq \frac{2^2u_{c2}^2}{l_{c2}^2}(t - sK)B^2 + \frac{4^2u_{c2}^2 N_c}{l_{c2}N} \sum_{i=1}^{N} \sum_{\substack{t',t''=sK \\ t'<t''}}^{t-1} E\, b^i(y^i_{t'}) - b(y_{t'}) \; E_{t'} b^i(y^i_{t''}) - b(y_{t''})_c$$

(using $l_{c2}\|\cdot\|_2 \leq \|\cdot\|_c$)

$$+ \frac{4^2(t - sK)}{N} \sum_{t'=sK}^{t-1}\sum_{i=1}^{N} E\left\{ \frac{2u_{c2}^2}{l_{c2}^2} A_1^2 \, \theta_{t'} - \theta_{t'}^i{}_c^2 + A_2^2 \|\theta_{t'}\|_c^2 + \ldots_c^2 \right\} \qquad \text{(Assumption 6.3)}$$

$$\overset{(a)}{\leq} \frac{2^2u_{c2}^2}{l_{c2}^2}(t - sK)B^2 + \frac{4^2u_{c2}^2 \, 2m_2 B}{l_{c2}N \; l_{c2}} \sum_{i=1}^{N} \sum_{t'=sK}^{t-1} \sum_{\substack{t''=sK \\ t'<t''}}^{t-1} E\left[\rho^{t''-t'}\right]$$

$$+ 4^2(t - sK) \sum_{t'=sK}^{t-1} \left(1 + \frac{2A_1^2u_{c2}^2}{l_{c2}^2}\right) \ldots + \frac{2A^2u^2}{l^2_{c2}} \|\theta_{t'}\|_c^2$$

$$\overset{(b)}{\leq} \frac{2^2u_{c2}^2}{l_{c2}^2}(t - sK)B^2 + \frac{4m_2B}{1-\rho} + 4^2(t - sK) \sum_{t'=sK}^{t-1} \left(1 + \frac{2A^2u_{c2}^2}{l_{c2}^2}\right) \ldots + \frac{2A^2u_{c2}^2}{l_{c2}^2} \|\theta_{t'}\|_c^2; \qquad (111)$$

where (a) follows since

$$E\, b^i(y^i_{t'}) - b(y_{t'})_2 \; E_{t'}\, b^i(y^i_{t''}) - b(y_{t''})_c$$

$$\leq E\left[b^i(y^i_{t'})_2^2 \right]^{1/2} \; E\left[\left\{ E_{t'} b^i(y^i_{t''})_c + E_{t'} \frac{1}{N}\sum_{j=1}^{N} b(y^j_{t''})_c \right\}^2 \right]^{1/2} \qquad (Var(X) \leq E[X^2], \text{ and Triangle inequality})$$

$$\leq E\left[\frac{1}{l_{c2}} b^i(y^i_{t'})_c^2 \right]^{1/2} \; E\left[\left\{ E_{t'}\, b^i(y^i_{t''})_c + \frac{1}{N}\sum_{j=1}^{N} E_{t'}\, b(y^j_{t''})_c \right\}^2 \right]^{1/2} \qquad \text{(Jensen's inequality)}$$

$$\leq \frac{B}{l_{c2}} 2m_2 E\left[\rho^{t''-t'}\right]: \qquad \text{(Assumption 6.1, 6.3)}$$

Finally, the inequality in (b) follows since

$$\sum_{t'=sK}^{t-1}\sum_{t''=sK}^{t-1} E\left[\|\theta_{t''}-\theta_{t'}\|\right] = \sum_{t'=sK}^{t-1}\sum_{t''=t'+1}^{t-1} E\left[\|\theta_{t''}-\theta_{t'}\|\right] = \sum_{t'=sK}^{t-1}\frac{t-t'}{1-\rho} = \frac{(t-sK)}{1-\rho}\left(t-sK-\frac{1-\rho^{t-sK}}{1-\rho}\right) \le \frac{(t-sK)^2}{1-\rho}.$$

For simplicity, we define $\tilde{A}_1 = \frac{2A_1^2u_{c2}^2}{l_{c2}^2}$; $\tilde{A}_2 = \frac{2A_2^2u_{c2}^2}{l_{c2}^2}$; $\tilde{B} = \frac{u_{c2}}{l_{c2}}B$. Hence, (111) simplifies to

$$E\left[\|\theta_t-\theta_{sK}\|^2\right] \le 2\alpha^2(t-sK)^2B^2\left(1+\frac{4m_2}{B(1-\rho)}\right) + 2\alpha\sum_{t'=sK}^{t-1}(1+\tilde{A}_1)E\left[\|\theta_{t'}-\theta_{sK}\|^2\right] + \tilde{A}_2E\left[\|\theta_{t'}-\theta_{sK}\|_c^2\right]. \quad (112)$$

Recursively applying (112), going back 2 steps, we see

$$\begin{aligned} E\left[\|\theta_t-\theta_{sK}\|^2\right] &\le 2\alpha^2B^2(t-sK)^2\left(1+\frac{4m_2}{B(1-\rho)}\right) + 4\alpha^2(t-sK)\tilde{A}_2\sum_{t'=sK}^{t-1}E\left[\|\theta_{t'}-\theta_{sK}\|_c^2\right] + 4\alpha^2(t-sK)(1+\tilde{A}_1)\sum_{t'=sK}^{t-1}E\left[\|\theta_{t'}-\theta_{sK}\|^2\right] \\ &\le 2\alpha^2B^2(t-sK)^2\left(1+\frac{4m_2}{B(1-\rho)}\right) + 4\alpha^2(t-sK)\tilde{A}_2\sum_{t'=sK}^{t-1}E\left[\|\theta_{t'}-\theta_{sK}\|_c^2\right] \\ &\quad + 4\alpha^2(t-sK)(1+\tilde{A}_1)\left[2\alpha^2B^2(t-sK-1)^2\left(1+\frac{4m_2}{B(1-\rho)}\right) + 2\alpha\sum_{t'=sK}^{t-2}(1+\tilde{A}_1)E\left[\|\theta_{t'}-\theta_{sK}\|^2\right] + \tilde{A}_2E\left[\|\theta_{t'}-\theta_{sK}\|_c^2\right]\right] \\ &= 2\alpha^2B^2(t-sK)^2\left(1+\frac{4m_2}{B(1-\rho)}\right)\left[1+4\alpha^2(1+\tilde{A}_1)(t-1-sK)\right] \\ &\quad + 4\alpha^2(t-sK)\tilde{A}_2\sum_{t'=sK}^{t-1}\left[\|\theta_{t'}-\theta_{sK}\|_c^2 + 4\alpha^2(1+\tilde{A}_1)(t-1-sK)\sum_{t'=sK}^{t-2}E\left[\|\theta_{t'}-\theta_{sK}\|_c^2\right]\right] \\ &\quad + 4\alpha^2(t-sK)(1+\tilde{A}_1)\left[1+4\alpha^2(1+\tilde{A}_1)(t-1-sK)\right]\sum_{t'=sK}^{t-2}E\left[\|\theta_{t'}-\theta_{sK}\|^2\right] \\ &\quad 4\alpha^2(t-sK)\left[1+4\alpha^2(1+\tilde{A}_1)(t-1-sK)\right]\left[\tilde{B}^2\left(1+\frac{4m_2}{B(1-\rho)}\right)+\tilde{A}_2\right]\sum_{t'=sK}^{t-1}E\left[\|\theta_{t'}-\theta_{sK}\|_c^2\right] \\ &\quad + (1+\tilde{A}_1)E\left[\|\theta_{t'}-\theta_{sK}\|^2\right]. \end{aligned} \quad (113)$$

To derive the bound for going back, in general, j steps (such that $t \ge j \ge sK$), we use an induction argument. Suppose for going back $k\,(<j)$ steps, the bound is

$$\begin{aligned} E\left[\|\theta_t-\theta_{sK}\|^2\right] &\le 4\alpha^2(t-sK)B^2\left(1+\frac{4m_2}{B(1-\rho)}\right)^k\prod_{\ell=1}^{k}\left[1+4\alpha^2(1+\tilde{A}_1)(t-\ell-sK)\right] \\ &\quad + 4\alpha^2(t-sK)\tilde{A}_2\sum_{\ell=1}^{k}\prod_{\ell'=1}^{\ell-1}\left[1+4\alpha^2(1+\tilde{A}_1)(t-\ell'-sK)\right]\sum_{t'=sK}^{t-1}E\left[\|\theta_{t'}-\theta_{sK}\|_c^2\right] \\ &\quad + 4\alpha^2(t-sK)(1+\tilde{A}_1)\sum_{\ell=1}^{k}\prod_{\ell'=1}^{\ell-1}\left[1+4\alpha^2(1+\tilde{A}_1)(t-\ell'-sK)\right]\sum_{t'=sK}^{t-\ell}E\left[\|\theta_{t'}-\theta_{sK}\|^2\right] \\ &\quad + (1+\tilde{A}_1)E\left[\|\theta_{t'}-\theta_{sK}\|^2\right]. \end{aligned} \quad (114)$$

We derive the bound for $k+1$ steps. For this, we further bound the last term in (114).

$$\sum_{\ell=1}^{k}\prod_{\ell'=1}^{\ell-1}\left[1+4\alpha^2(1+\tilde{A}_1)(t-\ell'-sK)\right]\sum_{t'=sK}^{t-\ell}E\left[\|\theta_{t'}-\theta_{sK}\|^2\right] \le \sum_{t'=sK}^{t-1}\sum_{\ell=1}^{k}\prod_{\ell'=1}^{\ell-1}\left[1+4\alpha^2(1+\tilde{A}_1)(t-\ell'-sK)\right]E\left[\|\theta_{t'}-\theta_{sK}\|^2\right]$$

$$4^2(t \quad sK)(1 + A_1) \qquad 1 + 4^2(1 + A_1)(t \quad ' \quad sK) \qquad E$$
$$t^0 + E$$
$$t \quad k\,' = 1 \qquad\qquad t^0 = sK$$

$$4^2(t \quad sK)(1+A_1)^{\sim} \prod_{t'=1}^{t \quad 1} \left[1+4^2(1+A_1)^{\sim}(t \quad ' \quad sK)\right] \sum_{t'=sK}^{t \quad 1} E$$
$$+ 4^2(t \quad sK)(1+A_1)^{\sim} \prod_{'=1}^{t \quad 1} \left[1+4^2(1+A_1)^{\sim}(t \quad ' \quad sK)\right]$$
$$\left[4^2(t \quad k \quad sK) \quad B^2 \, {}^{\sim}1+\frac{4m_2}{B(1 \quad)} + \sum_{t'=sK}^{t \quad k \quad 1} (1+A_1)E \right]^{\sim} {}^2$$
$$t' + A_2 E\|t'\|c \qquad \text{(using (112))}$$
$$4^2(t \quad sK)(1+A_1)^{\sim} \prod_{t'=1}^{t \quad 1} \left[1+4^2(1+A_1)^{\sim}(t \quad ' \quad sK)\right] \sum_{t'=sK}^{t \quad 1} E$$
$$+ 4^2(t \quad sK) \prod_{'=1}^{t \quad 1} \left[1+4^2(1+A_1)^{\sim}(t \quad ' \quad sK)\right]$$
$$\left[4^2(1+A_1)(t \quad k \quad sK) \quad B^2 \, {}^{\sim}1+\frac{4m_2}{B(1 \quad)} + \sum_{t'=sK}^{t \quad k \quad 1} (1+A_1)E \right]^{\sim} {}^2$$
$$t' + A_2 E\|t'\|c \qquad : \quad (115)$$

Substituting (115) into (114), we see that the induction hypothesis in (114) holds. We can go back as far as the last instant of synchronization, $j \quad t \quad sK$. For $j = t \quad sK$, we get

$$E \prod_{t'=sK}^{t \quad 1} \left[1+4^2(1+A_1)^{\sim}(t \quad ' \quad sK)\right]^{\sim} \left[4^2(t \quad sK) \prod_{'=1}^{t \quad sK} \left[1+4^2(1+A_1)^{\sim}(t \quad ' \quad sK)\right]\right]$$
$$B^{2\sim}\left[1+\frac{4m_2}{B(1 \quad)}\right] + A_2^{\sim} \sum_{t'=sK}^{t \quad 1} E\|t'\|c^2 + (1+A_1)^{\sim} E\|sK\|c^2 \qquad : \quad (116)$$

Next, using $1+x \quad e^x$ for $x \quad 0$, we get

$$\prod_{'=1}^{t \quad 1} \left[1+4^2(1+A_1)^{\sim}(t \quad ' \quad sK)\right] \quad \exp\left[\sum_{'=1}^{t \quad 1} 4^2(1+A_1)(t \quad ' \quad sK)\right]$$
$$\exp\left[2^2(1+A_1)(t \quad sK)^2\right]$$
$$\left[4^{\cdot}\right]^{5} \qquad (117)$$

if is small enough such that $2^2(1+A_1)(K \quad 1)^2 \quad \ln\frac{5}{4}$, which holds true by the assumption on the step size. Using (117) in (116), we get

$$E \quad {}^{\sim} \left[\frac{4m_2}{t \quad 5^2(t \quad sK)\,4B^2}\right]^{\sim} \left[1+\frac{B}{(1 \quad)}\right]^{c} + A_2 \sum_{t'=sK}^{t \quad 1} E\|t'\|c^2 +$$
$$^{\sim} \frac{4m_2}{(1+A)\,E} \quad {}^{\sim} \sum_{t'=sK}^{t \quad 1} {}_2$$
$$s \quad \}^{5}$$
$$= 5^2(t \quad sK)B^2 \left[1+\frac{1}{B(1 \quad)}\right] + 5^2(t \quad sK)A_2 \sum_{t'=sK}^{t \quad 1} E\|t'\|c ; \qquad (118)$$

which concludes the proof. □

Proof of Lemma B.9. We have

$$E_{t \quad r}[kb(Y_t)k_c] \quad u_{cD} E_{t \quad r}[kb(Y_t)k_D]$$
$$= u_{cD} E_{t \quad r} \quad q \frac{}{b(Y_t)^{>} D\, b(Y_t)}$$
$$q$$

$$u_{cD} \quad E_{t\ r}[b(Y_t)^{>} D\, b(Y_t)] \qquad \text{(concavity of square root)}$$

$$= \sqrt{\mathbb{E}_{t\sim r}\left[\left(\frac{1}{N}\sum_{i=1}^{N} b^i(y_t^i) - \frac{1}{N}\sum_{i=1}^{N} b^i(y_t^i)\right)^2\right]}$$

$$= \sqrt{\mathbb{E}_{t\sim r}\left[\frac{1}{N^2}\sum_{i=1}^{N} b^i(y_t^i)^\top b^i(y_t^i) + \frac{2}{N^2}\sum_{i<j} b^i(y_t^i)^\top b^j(y_t^j)\right]}$$

$$= \sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}_{t\sim r}\left[b^i(y_t^i)^2\right] + \frac{2}{N^2}\sum_{i<j}\mathbb{E}_{t\sim r}\left[b_i(y_t^i)^\top\right]\mathbb{E}_{t\sim r}\left[b^j(y_t^j)\right]} \quad \text{(Assumption 6.4)}$$

$$\le \sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}_{t\sim r}\left[b^i(y_t^i)^2\right]} + \sqrt{\frac{2}{N^2}\sum_{i<j}\mathbb{E}_{t\sim r}\left[b^i(y_t^i)^\top\right]\mathbb{E}_{t\sim r}\left[b^j(y_t^j)\right]}: \quad (119)$$

$$\underbrace{\qquad\qquad}_{T_1} \qquad\qquad \underbrace{\qquad\qquad}_{T_2}$$

For the term T_1 we have

$$T_1 \le \sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}_{t\sim r}\left[\|b^i(y_t^i)\|_2^2\right]} \le \sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}_{t\sim r}[B^2]} \quad \text{(Assumption 6.3)}$$

$$= \frac{B}{N}\sqrt{N} = \frac{1}{\sqrt{N}}: \quad (120)$$

For the term T_2 we have

$$T_2 \le \frac{2}{N}\sum_{i<j}\sqrt{\|\mathbb{E}_{t\sim r}\left[b^i(y_t^i)\right]\|_2\,\|\mathbb{E}_{t\sim r}\left[b^j(y_t^j)\right]\|_2} \quad \text{(Cauchy–Schwarz)}$$

$$\le \frac{2}{N}\sum_{i<j}\sqrt{\|c_D^2\|\mathbb{E}_{t\sim r}\left[b^i(y_t^i)\right]\|_2\,\|\mathbb{E}_{t\sim r}\left[b^j(y_t^j)\right]\|_2}$$

$$\le \frac{2}{N}\sum_{i<j}\sqrt{\|c_D^2\|m_{2r}\,m_{2r}} \quad \text{(Assumption 6.1)}$$

$$\le \frac{2m_{2r}}{\|c_D\|}: \quad (121)$$

Substituting (120) and (121) in (119), we get the result in (56).

The proof of (57) follows analogously. □

Proof of Lemma B.10. By definition, we have

$$\|G(\theta_t; Y_t) - G(\theta_t; Y_t)\|_c^2 = \frac{1}{N}\sum_{i=1}^{N}\left(G^i(\theta^i; y_t^i) - G^i(\theta_t; y^i)\right)_{i=1}^{c}{}^2$$

$$\le \frac{1}{N}\sum_{i=1}^{N}\|G^i(\theta^i; y_t^i) - G^i(\theta_t; y^i)\|_c^2 \quad \text{(convexity of norm)}$$

$$\le \frac{1}{N}\sum_{i=1}^{N}\left(A_1\|\theta^i - \theta_t\|_c\right)^2 \quad \text{(Assumption 6.3)}$$

$$= (A_1\tau_t)^2\left(\frac{1}{N}\sum_{i=1}^{N} A_1\|\theta^i - \theta_t\|_c\right)^2 \quad \text{(convexity of square)}$$

$= A_1^2$ $_t$: (By definition of $_t$)

Furthermore, by the convexity of $()^2$, we have

$$_t^2 \quad \frac{1}{N}\sum_{i=1}^{N} {}_t^{(i)} \Big)^2 \quad \frac{1}{N}\sum_{i=1}^{N} ({}_t^{(i)})^2 = {}_t \qquad \square$$

Proof of Lemma B.11. Since $M_f^{;g}()$ is convex, and there exists a norm, $k\ k_m$, such that $M_f^{;g}(x) = \frac{1}{2}kxk_m^2$ (see Proposition B.1), using the chain rule of subdifferential calculus,

$$r M_f^{;g}(x) = kxk_m u_x;$$

where $u_x \ 2\ @kxk_m$ is a subgradient of kxk_m at x. Hence,

$$r M_f^{;g}(x)^? = kxk_m ku_xk_m^?;$$

where $kk_m^?$ is the dual norm of kk_m. Since kk_m is convex and, as a function of x, is 1-Lipschitz w.r.t. kk_m, we have $ku_xk_m^? \ 1$ (see Lemma 2.6 in (Shalev-Shwartz et al., 2012)).

Further, by convexity of kk_m norm, $k0k_m \ kxk_m + hu_x;\ xi$. Therefore,

$$D \qquad E \qquad r M_f^{;g}(x); x = kxk_m hu_x; xi \ kxk_m^2 = 2M_f^{;g}(x): \qquad \square$$

C. Federated TD-learning

C.1. On-policy Function Approximation

Proposition C.1. On-policy TD-learning with linear function approximation Algorithm 1 satisfies the following:

1. $^i = v^i$ $_t$ v
2. $S_t = (S_t^1; : : : ; S_t^N)$ and $A_t = (A_t^1; : : : ; A_t^N)$
3. $y_t^i = (S_t^i; A_t^i; : : : ; S_{t+n\ 1}^i; A_{t+n\ 1}^i; S_{t+n}^i)$ and $Y_t = (S_t; A_t; : : : ; S_{t+n\ 1}; A_{t+n\ 1}; S_{t+n})$
4. $$: Stationary distribution of the policy $$.

Furthermore, choose some arbitrary positive constant $\ > 0$. The corresponding $G^i(^i; y_t^i)$ and $b^i(y^i)$ in Algorithm 1 for On-policy TD-learning with linear function approximation is as follows

1. $G^i(^i; {}_t y^i)_t = {}^i + {}_t {}^1(S^i) \sum_{l=t}^{t+n\ 1} {}^{l\ t} (S^i{}_{1})^{>i} \quad (S^i)^{>i}{}_t \qquad {}_l \qquad {}_t$
2. $b^i(y_t^i) = {}^1(S^i)_t \sum_{l=t}^{t+n\ 1} {}^{l\ t} R(S^i; A^i_l) + {}_l(S^{l+}{}_{1})^{>i} v \quad (S^i)^{>} v \qquad {}_l$

where v solves the projected bellman equation $v = ((T)^n v)$. Furthermore, the corresponding step size $$ in Algorithm 4 is $$.

Lemma C.1. Consider the federated on-policy TD-learning Algorithm 1 as a special case of FedSAM Algorithm 4 (see Proposition C.1). Suppose the trajectory $fS_t^i g_{t=0;1;:::}$ converges geometrically fast to its stationary distribution as follows $d_{TV}(P(S_t^i = jS_0^i = j^i())\ m$ for all $i = 1; 2; : : : ; N$. The corresponding $G^i()$ in Assumption 6.1 for the federated TD-learning is as follows

$$G^i(\theta) = \theta + \alpha\left(\frac{1}{n}(P)^n \theta\right); \tag{122}$$

where $\alpha > 0$ is an arbitrary constant introduced in Proposition C.1. In addition, (6) holds. Furthermore, for $t \geq n+1$, we have $m = \max\{\frac{2A_2 m}{n}, 2B\alpha\}$, where A_2 and B are specified in Lemma C.3 and $\alpha = \alpha$.

Lemma C.2. Consider the federated on-policy TD-learning 1 as a special case of FedSAM (as specified in Proposition C.1). Consider the $|S| \times |S|$ matrix $U = \alpha(P)^n - I)$ with eigenvalues $\{\lambda_1, \ldots, \lambda_{|S|}\}$. Define $\lambda_{\max} = \max_i |\lambda_i|$ and $\omega = -\max_i \mathrm{Re}[\lambda_i] > 0$, where $\mathrm{Re}[\cdot]$ evaluates the real part. By choosing α large enough in the linear function (122), there exist a weighted 2-norm $\|\cdot\| = \sqrt{\cdot^\top \cdot}$, such that $G^i(\theta)$ is a contraction with respect to this norm, i.e., $\|G^i(\theta_1) - G^i(\theta_2)\| \leq c\|\theta_1 - \theta_2\|$ for $c = 1 - \frac{\omega^2}{8\lambda_{\max}^2}$.

Lemma C.3. Consider the federated on-policy TD-learning Algorithm 1 as a special case of FedSAM (as specified in Proposition C.1). There exist some constants A_1, A_2, and B such that the properties of Assumption 6.3 are satisfied.

Lemma C.4. Consider the federated on-policy TD-learning Algorithm 1 as a special case of FedSAM (as specified in Proposition C.1). Assumption 6.4 holds for this algorithm.

C.1.1. PROOFS

Proof of Proposition C.1. Items 1-4 are by definition. Subtracting v from both sides of the update of the TD-learning, we have

$$\begin{aligned}
\underbrace{v_{t+1}^i - v}_{\theta_{t+1}^i} &= \underbrace{v_t^i - v}_{\theta_t^i} + \alpha(S_t)\sum_{l=t}^{t+n-1} R(S_l^i, A_l^i) + (S_{l+1})^\top v_t^i - (S_t^i)^\top v_t^i \\
&= \theta_t^i + \alpha(S_t)\sum_{l=t}^{t+n-1} R(S_l^i, A_l^i) + (S_{l+1})^\top (\underbrace{v_t^i - v}_{\theta_t^i}) - (S_t^i)^\top (\underbrace{v_t^i - v}_{\theta_t^i}) \\
&= \theta_t^i + \underbrace{\alpha(S_t)\sum_{l=t}^{t+n-1} (S_{l+1})^\top \theta_t^i - (S_t^i)^\top \theta_t^i}_{G^i(\theta_t^i, y_t^i)} \\
&\quad + \alpha(S_t)\sum_{l=t}^{t+n-1} R(S_l^i, A_l^i) + (S_{l+1})^\top v - (S_t^i)^\top v \Big) \\
&\quad \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{b^i(y_t^i)} .
\end{aligned}$$

which proves items 1 and 2. Furthermore, for the synchronization part of TD-learning, we have

$$\begin{aligned}
v_t^i - \frac{1}{N}\sum_{j=1}^N v_t^j &= \underbrace{v_t^i - v}_{\theta_t^i} - \frac{1}{N}\sum_{j=1}^N (\underbrace{v_t^j - v}_{\theta_t^j}), \\
&= \theta_t^i - \frac{1}{N}\sum_{j=1}^N \theta_t^j,
\end{aligned}$$

which is equivalent to the synchronization step in FedSAM Algorithm 4. Notice that here we used the fact that all agents have the same fixed point v.

Proof of Lemma C.1. It is easy to observe that □

$$G^i(\theta; y_t)_i = \theta + \alpha(S_t)\left(\sum_{l=t}^{t+n-1}(S_{l+1})^\top \theta - (S_t)^\top \theta\right)_i.$$

Taking expectation with respect to the stationary distribution, we have

$$G^i(\theta) = E_{S_t^i}\Big[\frac{1}{n}\sum_{t}^{t+n}(S_t^i)^\top \theta - \sum_{t}^{t+n}(S_t^i)^\top \theta\Big]$$

$$= E_{S_t^i}\Big[E\Big[\frac{1}{n}\sum_{t}^{t+n}(S_t^i)^\top \theta - \sum_{t}^{t+n}(S_t^i)^\top \theta \,\Big|\, S_t^i\Big]\Big] \quad \text{(tower property of expectation)}$$

$$= E_{S_t^i}\Big[\sum_{t}^{t+n}(S_t^i)^\top \theta - \sum_{t}^{t+n}\sum_{s'}^{S}\sum_{s}^{S} P(S_{t+n}^i = s' \mid S_t^i = s)\,(s)^\top \theta\Big]$$

$$= E_{S_t^i}\Big[\sum_{t}^{t+n}(S_t^i)^\top \theta - \sum_{t}^{t+n}(S_t^i)^\top (P^n)^\top \theta\Big]$$

$$= \sum_{t}^{t+n} E_{S_t^i}\big[(S_t^i)^\top \theta - (S_t^i)^\top (P^n)^\top \theta\big] = \sum_{t}^{t+n} E_{S_t^i}\big[(S_t^i)^\top (I - (P^n)^\top)\theta\big].$$

where P is the transition probability matrix corresponding to the policy θ, and Π is a diagonal matrix with diagonal entries corresponding to μ elements of μ.

As explained in (Tsitsiklis & Van Roy, 1997), the projection operator Π is a linear operator and can be written as $\Pi = (\Phi^\top \Pi \Phi)^{-1}\Phi^\top \Pi$, where Π is a diagonal matrix with diagonal entries corresponding to the stationary distribution of the policy θ. Hence, the fixed point equation is as follows $\Phi v = \Phi(\Phi^\top \Pi \Phi)^{-1}\Phi^\top \Pi (T^\theta)^n(\Phi v)$. Since Φ is a full column matrix, we can eliminate it from both sides of the equality, and further multiply both sides with $\Phi^\top \Pi$. We have $\Phi^\top \Pi \Phi v = \Phi^\top \Pi (T^\theta)^n(\Phi v)$, and hence $\Phi^\top \Pi((T^\theta)^n \Phi v - \Phi v) = 0$, which is equivalent to $E_S[\phi^\top(S)((T^\theta)^n(\Phi v)(S) - (\Phi v)(S)] = 0$. By expanding $(T^\theta)^n$, we have $E_{S_0^i}[\phi^\top(S_0^i)\big(\sum_{l=0}^{n-1}(R(S_l^i, A_l^i) + (\Phi v)(S_{l+1}^i) - (\Phi v)(S_l^i))\big)] = 0$, which means

$$E_y\, b^i(y) = 0, \tag{123}$$

and proves (6).

Moreover, we have

$$\|G^i(\theta) - E[G^i(\theta; y^i)]\|_c = \big\|E_{y^i}[G^i(\theta; y^i)] - E[G^i(\theta; y^i)]\big\|_c = \Big\|\sum_{y_t^i}\sum_{t}^{t+n} P(y_t^i = y_t^i \mid y_0^i) G^i(\theta; y^i) - \sum_{t}^{t+n}\sum_{y^i} \pi(y^i) P_t(y_t^i = y_t^i \mid y_0^i) G^i(\theta; y^i)\Big\|_c$$

$$\le \sum_{t}^{t+n}\sum_{y^i} \big|\pi(y^i) - P_t(y_t^i = y_t^i \mid y_0^i)\big| \cdot \|G^i(\theta; y^i)\|_c \le \sum_{t}^{t+n}\sum_{y^i} \big|\pi(y^i) - P_t(y_t^i = y_t^i \mid y_0^i)\big| \cdot A_2 \|\theta\|_c \quad (\|ax\|_c = |a|\|x\|_c)$$

$$\le \sum_{t}^{t+n} A_2 \|\theta\|_c \cdot \max_{y_0^i} \sum_{y^i} \big|\pi(y^i) - P_t(y_t^i = y_t^i \mid y_0^i)\big| \le \sum_{t}^{t+n} A_2 \|\theta\|_c \cdot A_1 \rho^t \quad \text{(Assumption 6.3)}$$

For brevity, we denote $P(S_t^i = s_t^i) = P(s_t^i)$. We have

$$\sum_{y^i} \pi(y^i) P(y_t^i = y^i \mid y_0^i)$$

$$= \sum_{s_t^i, a_t^i, \ldots, s_{t+n}^i} \pi(s_t, a_t^i, \ldots, s_{t+n}^i) P(s_t, a_t^i, \ldots, s_{t+n}^i \mid y_0^i)$$

$$= \sum_{s_t, a_t^i, \ldots, s_{t+n}} \pi(a_t \mid s_t) P(s_{t+1} \mid s_t, a_t) \cdots P(s_{t+n} \mid s_{t+n-1}, a_{t+n-1}) P(s_t \mid S_0^i) P(a_t \mid s_t) P(s_{t+1} \mid s_t, a_t) \cdots P(s_{t+n} \mid s_{t+n-1}, a_{t+n-1}) \quad (t \ge n+1)$$

$$= \sum_{s_t^i, a_t^i, \ldots, s_{t+n}^i} \pi(s_t) \prod_{i=t}^{t+n-1} P(s_i \mid S_0^i) \pi(a_t \mid s_t) P(s_{t+1} \mid s_t, a_t) \cdots P(s_{t+n} \mid s_{t+n-1}, a_{t+n-1})$$

$$= \sum_{s^i} \pi(s^i) \, {}_t P(s^i | S_t^i) \, s_h$$
$$= 2 d_{TV}(P(S_t^i = \cdot | S_n^i = \cdot), \pi(\cdot)) \, 2m$$
$$\leq (2m^n) \cdot {}^t$$

$$\mathbb{E}[b^i(y^i)]^c = \mathbb{E}[b^i(y^i)] \quad \mathbb{E}_{y_t}[b^i(y^i)]_c \qquad \text{(By (123))}$$
$$\leq \sum_{t} \sum_{t} \pi(y^i) \, P(y^i = y^i | y^i) \, b^i(y^i) =$$
$$\sum_{y_t^i} \pi(y^i) \, P(y^i = y^i | y^i) b^i(y^i)_c$$
$$\sum_{y_t^i} \pi(y^i) \, P_t(y_t = y_t^i | y_0) B^i$$
$$2Bm^t \qquad \text{(Assumption 6.3)}$$

Proof of Lemma C.2. Consider the $|S||S|$ matrix $U = {}^{>}(^n(P)^n \quad I)$ with eigenvalues $\{_1; \cdots; _{|S|}\}$. As shown in (Tsitsiklis & Van Roy, 1997), since is a full rank matrix, the real part of $_i$ is strictly negative for all $i = 1; \cdots; |S|$. □

Furthermore, define $_{max} = \max_i |_i|$ and $= \max_i \mathrm{Re}[_i] > 0$, where $\mathrm{Re}[\cdot]$ evaluates the real part. Consider the matrix $U^0 = I + \frac{1}{2 \, ^2_{max} =} U$. It is easy to show that the eigenvalues of U^0 are $\{1 + \frac{1}{2 \, ^2_{max} =}; \cdots; 1 + \frac{|S|}{2 \, ^2_{max} =}\}$. For an arbitrary i, the norm of the i'th eigenvalue satisfies

$$1 + \frac{2}{2_{max}=} \quad {}^{i} \quad {}^{2} = 1 + \frac{\mathrm{Re}[_i]}{2_{max}=} \quad {}^{2} + \frac{\mathrm{Im}[_i]}{2_{max}=} \quad {}^{2}$$
$$1 + \frac{\mathrm{Re}[_i]}{2_{max}=} \quad {}^{2} + \frac{\mathrm{Im}[_i]}{2_{max}=} \quad {}^{2} \qquad (\mathrm{Re}[_i] < 0)$$
$$1 + \frac{2}{2_{max}=} + \frac{\quad {}^{2}}{2_{max}=}$$
$$= 1 \quad \frac{2}{4 \, ^2_{max}}:$$

Hence, all the eigenvalues of U^0 are in the unit circle. By (Bertsekas et al., 1995), Page 46 footnote, we can find a weighted 2-norm as $k \cdot k_{\quad} = \sqrt{\quad {}^{>}}$ such that U^0 is contraction with respect to this norm with some contraction factor $_c$. In particular, there exist a choice of such that we have $= _c 1 \quad \frac{2}{8 \, ^2_{max}}$.

Proof of Lemma C.3. The existence of A_1 and A_2 immediately follows after observing that $G^i(^i; y^i)_t$ is a linear function of ^i_t. Furthermore, the result on B follows due to v being bounded as shown in (Chen et al., 2021b). □

Proof of Lemma C.4. For the sake of brevity, we write $S_t^i = s_t^i$ simply as s_t^i, and similarly for other random variables. We have

$$\mathbb{E}_{t \quad r}[f(y_t^i) \quad g(y_t)]$$

$$= \sum_{s_t^i, a_t^i, \ldots, s_{t+n-1}^i, a_{t+n-1}^i, s_{t+n}^i, s_t^j, a_t^j, \ldots, s_{t+n-1}^j, a_{t+n-1}^j, s_{t+n}^j} P(s_t^i, a_t^i, \ldots, s_{t+n-1}^i, a_{t+n-1}^i, s_{t+n}^i, s_t^j, a_t^j, \ldots, s_{t+n-1}^j, a_{t+n-1}^j, s_{t+n}^j | s_{t-r}^i, s_{t-r}^j) f(y_t^i) g(y_t^j)$$

$$= \sum_{s_{t-r}^i, \ldots, s_{t+n-1}^i, a_{t-r}^i, \ldots, a_{t+n-1}^i, s_{t+n}^i} \sum_{s_{t-r}^j, \ldots, s_{t+n-1}^j, a_{t-r}^j, \ldots, a_{t+n-1}^j, s_{t+n}^j} P(s_{t-r}^i, \ldots, s_{t+n-1}^i, a_{t+n-1}^i, s_{t+n}^i | s_{t-r}^i) P(s_{t-r}^j, \ldots, s_{t+n-1}^j, a_{t+n-1}^j, s_{t+n}^j | s_{t-r}^j) f(y_t^i) g(y_t^j)$$

$$= \sum_{s_{t-r}^i, \ldots, s_{t+n-1}^i, a_{t-r}^i, \ldots, a_{t+n-1}^i, s_{t+n}^i} P(s_{t-r}^i, \ldots, s_{t+n-1}^i, a_{t+n-1}^i, s_{t+n}^i | s_{t-r}^i) f(y_t^i) \sum_{s_{t-r}^j, \ldots, s_{t+n-1}^j, a_{t+n-1}^j, s_{t+n}^j} P(s_{t-r}^j, \ldots, s_{t+n-1}^j, a_{t+n-1}^j, s_{t+n}^j | s_{t-r}^j) g(y_t^j)$$

$$= \mathbb{E}_{t-r}[f(y_t^i)] \, \mathbb{E}_{t-r}[g(y_t^j)].$$

□

Proof of Theorem 4.1. By Proposition C.1 and Lemmas C.1, C.2, C.3, and C.4, it is clear that the federated TD-learning with linear function approximation Algorithm 1 satisfies all the Assumptions 6.1, 6.2, 6.3, and 6.4 on the FedSAM Algorithm 4. Furthermore, by the proof of Theorem B.1, we have $w_t = (1 - \frac{\omega_2'}{2})^t$, and the constant c_{TDL} in the sampling distribution $q_T^{c_{TDL}}$ in Algorithm 1 is $c_{TDL} = (1 - \frac{\omega_2'}{2})^{-1}$. Furthermore, by choosing the step size small enough, we can satisfy the requirements in (21), (23), (32), (35). By choosing K large enough, we can satisfy $K > \cdot$. Hence, the result of Theorem B.1 holds for this algorithm with some $c_{TDL} > 1$. Also, it is easy to see that $(1 - \frac{\omega_2'}{2}) = O(1)$, which is a constant that can be absorbed in C_1^{TDL}. Finally, for the sample complexity result, we simply employ Corollary B.1.1.

Next, we derive the constant c_{TDL}. Since $\|\cdot\| \le c\|\cdot\|$, which is smooth, we choose $g(\cdot) = \frac{1}{2}\|\cdot\|_2^2$. By taking $= 1$, we have $l_{cs} = u_{cs} = 1$. Therefore, we have $\omega_1' = 1$, and $\omega_2' = 1 - c$, and $c_{TDL} = \frac{1}{2(1-c)}$, where c is defined in Lemma C.2.

□

C.2. Off-policy Tabular Setting

In this subsection, we verify that the Off-policy federated TD-learning Algorithm 2 satisfies the properties of the FedSAM Algorithm 4. In the following, V is the solution to the Bellman equation (124).

$$V(s) = \sum_a \pi(a|s) \left[R(s,a) + \sum_{s'} P(s'|s,a) V(s') \right] \tag{124}$$

Note that V is independent of the sampling policy of the agent. Furthermore, we take $\|\cdot\| = \|\cdot\|_1$.

Proposition C.2. Off-policy n-step federated TD-learning is equivalent to the FedSAM Algorithm 4 with the following parameters.

1. $\Delta^i = V^i - V$
2. $S_t = (S_t^1, \ldots, S_t^N)$ and $A_t = (A_t^1, \ldots, A_t^N)$
3. $y_t^i = (S_t^i, A_t^i, \ldots, S_{t+n-1}^i, A_{t+n-1}^i, S_{t+n}^i)$ and $Y_t = (S_t, A_t, \ldots, S_{t+n-1}, A_{t+n-1}, S_{t+n})$
4. π^i : Stationary distribution of the sampling policy of the i-th agent.
5. $G^i(\Delta^i; y_t^i)(s) = \Delta^i(s) + \mathbb{1}_{\{s=S_t\}} \sum_{l=t}^{t+n-1} \mathbb{1}_{\{t=l\}} \pi^{(i)}(S^i, A^i) \Delta^i(S_{j-1}^i) - \Delta^i(S_t^i)$
6. $b^i(y_t^i)(s) = \mathbb{1}_{\{s=S_t^i\}} \sum_{l=t}^{t+n-1} \mathbb{1}_{\{t=l\}} \pi^{(i)}(S^i, A^i) R(S^i, A^i) + \gamma^l V(S_{t+l}^i) - V(S_t^i)$

Lemma C.5. Consider the federated off-policy TD-learning Algorithm 2 as a special case of FedSAM (as specified in Proposition C.2). Suppose the trajectory $\{S_t^i\}_{t=0,1,\ldots}$ converges geometrically fast to its stationary distribution as follows

$d_{TV}(P(S_t^i = \cdot | S_0^i = j) \| \mu^i(\cdot)) \le m\rho^t$ for all $i = 1, 2, \dots, N$. The corresponding $G^i(\cdot)$ in Assumption 6.1 for the federated TD-learning is as follows

$$G^i(\cdot)_s = (s) + {}^{n\,i}(P)^n \quad {}^i(s).$$

Furthermore, for $t \ge n + 1$, we have $m_1 = \frac{2A}{n}\frac{m_2}{\rho}$ where A_2 is the constant specified in Assumption 6.3, $m_2 = 0$, and $= .$

Lemma C.6. Consider the federated off-policy TD-learning 2 as a special case of FedSAM (as specified in Proposition C.2). The corresponding contraction factor $_c$ in Assumption 6.2 for this algorithm is $_c = 1 \quad _{min}(1 \quad ^{n+1})$, where $_{min} = \min_{s,i} {}^i(s)$.

Lemma C.7. Consider the federated off-policy TD-learning 2 as a special case of FedSAM (as specified in Proposition C.2). The constants A_1, A_2, and B in Assumption 6.3 can be chosen as follows: $A_1 = A_2 = 1 + (1 +) \begin{cases} n & \text{if } _{max} = 1 \\ \frac{1}{1 \ _{max}^{\frac{1}{n}(o.w.)}} & \text{o.w.} \end{cases}$; and $B = \frac{2_{max}}{1} \begin{cases} n & \text{if } _{max} = 1 \\ \frac{1}{1 \ (1 \ _{max})^n} & \text{o.w.} \end{cases}$, where $_{max} = \max_{s^i,a^i,i} {}^{(i)}(s^i, a^i)$.

Lemma C.8. Consider the federated off-policy TD-learning 2 as a special case of FedSAM (as specified in Proposition C.2). Assumption 6.4 holds for this algorithm.

C.2.1. PROOFS

Proof of Proposition C.2. Items 1-4 are by definition. Furthermore, by the update of the TD-learning, and subtracting V from both sides, we have

$$\underbrace{V_{t+1}^i \quad V^*(s)}_{{}_{t+1}^i(s)} = \underbrace{V_t^i \quad V^*(s)}_{{}_t^i(s)}$$

$$+ 1_{\{s=S_t\}} \sum_{l=t}^{t+n-1} \left[{}^{(i)}(S_j^i, A_j^i) \right]_{j=t}^{l} \Big[R(S_l^i, A_l^i) + V_t(S_{l+1}^i) \quad V_t^i(S_l) \Big]$$

$$= {}_t^i(s) + 1_{\{s=S_t\}} \sum_{l=t}^{t+n-1} \left[{}^{(i)}(S_j^i, A_j^i) \right]_{j=t}^{l} \Big[R(S_l^i, A_l^i) + \underbrace{V_t^i(S_{l+1}^i) \quad V^*(S_{l+1}^i)}_{{}_t^i(S_{l+1}^i)} + V^*(S_{l+1}) \quad \underbrace{V_t^i(S_l) \quad V^*(S_l)}_{{}_t^i(S_l)} \Big]$$

$$= {}_t^i(s) + {}^i(s) \quad {}_t^i(s) + 1_{\{s=S_t\}} \sum_{l=t}^{t+n-1} \left[{}^{(i)}(S_j^i, A_j^i) \right]_{j=t}^{l} \Big[{}_t^i(S_{l+1}^i) \quad {}_t^i(S_l^i) \Big]$$

$$+ 1_{\{s=S_t\}} \sum_{l=t}^{t+n-1} \left[{}^{(i)}(S_j^i, A_j^i) \right]_{j=t}^{l} \Big[\underbrace{R(S_l^i, A_l^i) + V^*(S_{l+1}) \quad V^*(S_l)}_{b^i(y_t)_s} \Big],$$

which proves items 5 and 6. Furthermore, for the synchronization part of TD-learning, if $t \mod K = 0$,

$$V_t^i \quad \frac{1}{N} \sum_{j=1}^{N} V_t^j =) \underbrace{V_t^i \quad V^*}_{{}_t^i} \quad \frac{1}{N} \sum_{j=1}^{N} (\underbrace{V_t^j \quad V^*}_{{}_t^j});$$

which is equivalent to the synchronization step in FedSAM Algorithm 4. Notice that here we used the fact that all agents have the same fixed point V. □

Proof of Lemma C.5. By taking expectation of $G^i(\theta^i; y_t)$ w.r.t. π_s, we have

$$G^i(\theta)(s) = \mathbb{E}_{S^i_t}\Big[\mathbb{1}_{\{s=S^i_t\}}\sum_{l=t}^{t+n-1} l_t - l_{=t} J^{(i)}(S^i_j; A^i_j)_j \quad (S^i_{l+1}) \quad (S^i)_{l=t}\Big]$$
$$= (s) + \sum_{l=t}^{t+n-1} \underbrace{\mathbb{E}_{S^i_t}\Big[\mathbb{1}_{\{s=S_t\}} {}_{i\ t\ l=t} I^{(i)}{}_j(S^i; A^i)_j \quad {}_j \quad (S^i_{l+1}) \quad (S^i)_{l}\Big]}_{T_l}.$$

Denote $\mathbb{E}^i_k[\cdot] = \mathbb{E}[\cdot | \{S^i_\tau, A^i_\tau\}_{\tau=k-1}; S^i_k]$. For T_l, we have

$$T_l = \mathbb{E}_{S^i_t}\Big[\mathbb{E}^i_l\Big[\mathbb{1}_{\{s=S^i_t\}} {}_{j=t} I^{(i)}(S^i_j; A^i_j) \quad (S^i_{l+1}) \quad (S^i_l)\Big]\Big]$$
$$= \mathbb{E}_{S^i_t}\Big[\mathbb{1}_{\{s=S_t\}} {}_{j=t} I^{(i)}(S^i; A^i)_j \quad {}_j \quad \mathbb{E}^i_l I^{(i)}(S^i; A^i)_l (S^i_{l+1}) \quad (S^i_l)\Big].$$

Here,

$$\mathbb{E}^i_l\big[I^{(i)}(S^i_l; A^i_l)(S^i_{l+1})\big] = \sum_{s;a} P(S^i_{l+1} = s; A^i_l = a | S^i_l) I^{(i)}(S^i_l; A^i_l = a)(s)$$
$$= \sum_{s;a} P(S^i_{l+1} = s; A^i_l = a | S^i_l) \frac{(a|S^i_l)}{{}^i(a|S^i_l)} (s)$$
$$= \sum_{s;a} {}^i(a|S^i_l) P(s|S^i_l; a) \frac{(a|S^i_l)}{{}^i(a|S^i_l)}(s)$$
$$= \sum_{s;a} P(s|S^i_l; a)(a|S^i_l)(s) = [P](S^i_l); \quad {}_{s;a}$$

where $[P]_{s_0;s_1} = \sum_a P(s_1|s_0; a)(a|s_0)$. Hence, we have

$$T_l = \mathbb{E}_{S^i_t}\Big[\mathbb{1}_{\{s=S^i_t\}} {}_{t\ j=t} I^{(i)}(S^i_j; A^i_j) \mathbb{E}^i_{l-1} I^{(i)}(S^i_{l-1}; A^i_{l-1}) (P)(S^i_l) \quad (S^i_t)\Big]$$
$$= \mathbb{E}_{S^i_t}\Big[\mathbb{1}_{\{s=S_t\}} {}_{j=t} I^{(i)}(S^i; A^i)_j ((P)^2)(S^i_{l-1}) \quad (P_l)(S_{l-1})\Big]$$
$$= \cdots$$
$$= \mathbb{E}_{S^i_t}\Big[\mathbb{1}_{\{s=S_t\}} {}_{j=t} I^{(i)}(S^i; A^i)_j ((P)^{l-t+1})(S^i_t) \quad ((P)^{l-t})(S^i_t)\Big] = {}^i(s)^{l-t}{}_t \quad {}^i$$
$$((P)^{l-t+1})(s) \quad ((P)^{l-t})(s)$$
$$= {}^{l-t}({}^i(P)^{l-t+1})(s) \quad ({}^i(P)^{l-t})(s) =$$
$${}^{t+1}{}^i(P)^{l-t+1} \quad {}^{l-t}{}^i(P)^{l-t}(s);$$

where we denote ^{i} as diagonal matrix with diagonal entries corresponding to the stationary distribution ^{i}. Hence, in total we have

$$G^i(\theta)(s) = (s) + \sum_{l=t}^{t+n-1} {}^{l-t+1}{}^i(P)^{l-t+1} \quad {}^{l-t}{}^i(P)^{l-t}(s)$$
$$= (s) + {}^{n\ i}(P)^n \quad {}^i(s).$$

Furthermore, by the same argument as in the proof of Lemma C.1, we have

$$\|G^i(\theta) \quad \mathbb{E}[G^i(\theta; y_t)]\|_c \quad {}^i \sum_{y_t} {}^i(y_t) \quad P(y_t = y_t | y_0): A^i_2 \|\theta\|_c; \quad {}_{y_t}$$

and

$$\sum_{y^i} \pi^i(y^i) \, t \, P(y^i = y^i | y) \, (2m) \, ; \quad y_t \quad t$$

which proves that $m_1 = 2mA_2^{\,n}$ constant. In addition, we have

$$\begin{aligned}
E[b^i(y^i)]^c &= E\Big[1_{\{s=S^i\}_t} \sum_{l=t}^{t+n-1} \pi^{(i)}(S_j; A_j) \, R(S_l; A_l) + {}^i V(S_{l+1}) - V(S_l) \Big]^c \\
&= E\Big[1_{\{s=S^i\}_t} \sum_{l=t}^{t+n-1} \Big[l \, t E_l \, {}_{l=t} \pi^{(i)}(S^i; A_j^i) \, {}_j R(S^i; A_l^i) + V(S^i_{\,1}) \, {}_{l+} V(S^i_{\,l}) \Big] \Big]^c \\
&= E\Big[1_{\{s=S^i\}_t} \sum_{l=t}^{t+n-1} \Big[l \, t \sum_{j=t}^{l} \pi^{(i)}(S^i; A_j^i) \, E_l \, \pi^{(i)}(S^i; A_l^i) \, {}_l R(S^i; A_l^i) + V(S^i_{\,1}) \, {}_{l+} V(S^i_{\,l}) \Big] \Big]^c \\
&\qquad \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{T}
\end{aligned}$$

For the term T, we have

$$\begin{aligned}
T &= \sum_{a} \pi^i(a|S^i_l) \cdot \frac{(a|S_l)^{\,i}}{(i a|S)^{\,i}} R(S^i; a) + \sum_{s'} P(s'|S^i_l; a) V(s') - V(S^i_{\,l}) \Big]_a \\
&= \sum_{s'} (a|S_l)^{\,i} \, R(S^i; a) + \sum_{s'} P(s'|S^i_l; a) V(s') - V(S^i_{\,l}) \Big]_a \\
&= 0 ;
\end{aligned}$$

which shows that $m_2 = 0$.

□

Proof of Lemma C.6.

$$\begin{aligned}
\|G^i(\theta_1) - G^i(\theta_2)\|_c &= {}_1 + {}^{n+1} {}^i(P)^{n+1} \, {}_1 \, {}^i{}_1 - {}_2 + {}^{n+1} {}^i(P)^{n+1} \, {}_2 \, {}^i{}_2 \, {}_1 = I - {}^i(I - {}^{n+1}(P)^{n+1}) \, (\theta_1 - \theta_2)_1 \\
&\le I - {}^i(I - {}^{n+1}(P)^{n+1}) \, \|\theta_1 - \theta_2\|_1 \, k_1 \quad \text{(definition of matrix norm)}
\end{aligned}$$

Since the elements of the matrix $I - {}^i(I - {}^{n+1}(P)^{n+1})$ is all positive, we have $I - {}^i(I - {}^{n+1}(P)^{n+1}) = (I - {}^i(I - {}^{n+1}(P)^{n+1}))1^1 = k1 - {}^i(1 - {}^{n+1}1)k_1 = 1 - {}^i{}_n(1 - {}^{n+1}) \, 1 - {}_{\min}(1 - {}^{n+1})$.

□

Proof of Lemma C.7.

$$\begin{aligned}
&\|G^i(\theta_1; y) - G^i(\theta_2; y)\|_c \\
&= \max_s |\theta_1(s) - \theta_2(s)| + 1_{\{s=S^i\}_t} \Big| \sum_{l=t}^{t+n-1} \Big[l \, t \, {}_{l=t} \pi^{(i)}(S^i; A_j^i) \, {}_j \theta_1(S_{l+1}^i) - \theta_1(S_l) \, {}_{l=t}^{i} \Big] \Big| \\
&\quad - 1_{\{s=S^i_t\}} \Big| \sum_{l=t}^{t+n-1} \Big[l \, t \, {}_{l=t} \pi^{(i)}(S^i; A^i) \, {}_j \theta_2(S^i_{\,1}) - \theta_2(S_l) \, {}_i \Big] \Big| \qquad (*\, c = 1) \\
&\le \max_s |\theta_1(s) - \theta_2(s)| + \sum_{l=t}^{t+n-1} \Big| l \, t \, {}_{l=t} \pi^{(i)}(S^i; A^i) \, {}_j \theta_1(S_{l+1}) - {}_i \theta_2(S_{l+1}) - {}_i \theta_1(S_l) - \theta_2(S_l) \Big|_i \\
&\qquad \text{(triangle inequality)} \\
&\le \max_s |\theta_1(s) - \theta_2(s)| + \sum_{l=t}^{t+n-1} \Big| l \, t \sum_{j=t}^{l} \pi^{(i)}(S^i; A_j^i) \, {}_j \theta_1(S^i_{\,1}) - {}_l \theta_2(S^i_{\,1}) + {}_l \theta_2(S^i) - \theta_2(S^i) \Big|_l \\
&\qquad \text{(triangle inequality)}
\end{aligned}$$

$$\max_s \left| \left[\mathcal{T}_1(s) - \mathcal{T}_2(s) \right] + \sum_{l=t}^{t+n-1} \mathbb{I}^{(i)}(S_l^i; A^i) \left[k_{l-t} \mathcal{T}_2 k_1 + k_{l-t+1} \mathcal{T}_2 k_1 \right] \right| \qquad \text{(definition of } k \text{ } k_1\text{)}$$

$$\max_s \left| \left[\mathcal{T}_1(s) - \mathcal{T}_2(s) \right] + \sum_{l=t}^{t+n-1} \mathbb{I}_{\max}^{l-t+1} \left[k_{l-t} \mathcal{T}_2 k_1 + k_{l-t+1} \mathcal{T}_2 k_1 \right] \right| \qquad \text{(definition of } \mathbb{I}_{\max}\text{)}$$

$$= \left\| k_1 \mathcal{T}_2 k_1 + \left[k_1 \mathcal{T}_2 k_1 + k_1 \mathcal{T}_2 k_1 \right] \mathbb{I}_{\max} \sum_{l=0}^{n-1} (\mathbb{I}_{\max})^l \right\|$$

$$= \left\| k_1 \mathcal{T}_2 k_1 + \left[k_1 \mathcal{T}_2 k_1 + k_1 \mathcal{T}_2 k_1 \right] \mathbb{I}_{\max} \begin{cases} n & \text{if } \mathbb{I}_{\max} = 1 \\ \frac{1 - (\mathbb{I}_{\max})^{n-1}}{1 - \mathbb{I}_{\max}} & \text{o:w:} \end{cases} \right\|$$

Furthermore, we have

$$\begin{aligned}
\left\| b^i(y_t^i) - b_c \right\| &= \max_{S_t^i, A_t^i, \ldots, S_{t+n}^i} \mathbb{1}_{\{s = S_t^i\}} \left| \sum_{l=t}^{t+n-1} \left[\prod_{j=t}^{l-1} \mathbb{I}^{(i)}(S_j^i; A_j^i) \right] \left[R(S_l^i; A_l^i) + V(S_{l+1}^i) - V(S_l) \right] \right| \\
&\leq \max_{S_t^i, A_t^i, \ldots, S_{t+n}^i} \sum_{l=t}^{t+n-1} \left| \prod_{j=t}^{l-1} \mathbb{I}^{(i)}(S_j^i; A_j^i) \left[R(S_l^i; A_l^i) + V(S_{l+1}^i) - V(S_l) \right] \right| \qquad \text{(triangle inequality)} \\
&\leq \max_{S_t^i, A_t^i, \ldots, S_{t+n}^i} \sum_{l=t}^{t+n-1} \mathbb{I}_{\max}^{l-t+1} \left| R(S_l^i; A_l^i) + V(S_{l+1}^i) - V(S_l) \right| \\
&\leq \max_{S_t^i, A_t^i, \ldots, S_{t+n}^i} \sum_{l=t}^{t+n-1} \mathbb{I}_{\max}^{l-t+1} \left[\left| R(S_l^i; A_l^i) \right| + \left| V(S_{l+1}^i) \right| + \left| V(S_l^i) \right| \right] \qquad \text{(triangle inequality)} \\
&\leq \max_{S_t^i, A_t^i, \ldots, S_{t+n}^i} \sum_{l=t}^{t+n-1} \mathbb{I}_{\max}^{l-t+1} \left[1 + \frac{1}{1-\gamma} + \frac{1}{1-\gamma} \right] \\
&= \frac{2}{1-\gamma} \sum_{l=t}^{t+n-1} (\mathbb{I}_{\max})^{l-t} \\
&= \frac{2}{1-\gamma} \begin{cases} n & \text{if } \mathbb{I}_{\max} = 1 \\ \frac{1 - (\mathbb{I}_{\max})^n}{1 - \mathbb{I}_{\max}} & \text{o:w:} \end{cases}
\end{aligned}$$

□

Proof of Lemma C.8. The proof follows similar to Lemma C.4. □

Proof of Theorem 5.1. By Proposition C.2 and Lemmas C.5, C.6, C.7, and C.8, it is clear that the federated off-policy TD-learning Algorithm 2 satisfies all the Assumptions 6.1, 6.2, 6.3, and 6.4 of the FedSAM Algorithm 4. Furthermore, by the proof of Theorem B.1, we have $w_t = (1 - \frac{\gamma_2}{2})^{-t}$, and the constant c in the sampling distribution q_T^c in Algorithm 1 is $c = (1 - \frac{\gamma_2}{2})^{-1}$. In equation (125) we evaluate the exact value of w_t.

Furthermore, by choosing step size small enough, we can satisfy the requirements in (21), (23), (32), (35). By choosing K large enough, we can satisfy $K > $, and by choosing T large enough we can satisfy $T > K + $. Hence, the result of Theorem B.1 holds for this algorithm.

Next, we derive the constants involved in Theorem B.1 step by step. After deriving the constants C_1, C_2, C_3, and C_4 in Theorem B.1, we can directly get the constants $C_i^{T D_T}$ for $i = 1; 2; 3; 4$.

In this analysis we only consider the terms involving $|S|, |A|, \frac{1}{1-\gamma}, \mathbb{I}_{\max}$, and $\min$. Since $\| \cdot \|_c = \| \cdot \|_1$, we choose $g(\cdot) = \frac{1}{2}\| \cdot \|_p^2$, i.e. the p-norm with $p = 2\log(|S|)$. It is known that $g(\cdot)$ is $(p-1)$ smooth with respect to $\| \cdot \|_p$ norm (Beck, 2017), and hence $L = (\log(|S|))$. Hence, we have $l_{cs} = |S|^{1-1/p} = \frac{1}{e} = (1)$ and $u_{cs} = \frac{p-1}{e} = (1)$. Therefore, we have $\gamma_1 = \frac{1+\gamma}{1+\gamma - u_{cs}^2 \gamma_{cs}^2} = \frac{1+\gamma}{1+\gamma - \frac{\gamma_{cs}^2}{e}} \leq 1 + $. By choosing $= (\frac{1+c}{2c})^2 \quad 1 = \frac{1+2\gamma - 3c + c^2}{4c^2}(1 - c) = \min(1 - \gamma^{n+1}) =$

$(_{min}(1\quad))$, which is $\quad = O(1)$, we have $'_1 = \frac{1+}{1+\frac{}{e}} = e^{\frac{(\frac{1+c}{2c})^2}{e+(\frac{1+c}{2c})^2\ 1}} = O(1)$, and

$$'_2 = 1 \quad c\ \sqrt{\frac{1+}{1+\frac{}{e}}} = 1 \quad c\ \sqrt{\frac{(\frac{1+c}{2c})^2}{1+\frac{(\frac{1+c}{2c})^2}{e}\ 1}} = 1 \quad \sqrt{\frac{0{:}5(1+\ c)e^{1=4}}{e\quad 1+\frac{1+c}{2c}}}^2 = 1 \quad \sqrt{\frac{0{:}5e^{1=4}(2\quad _{min}(1\quad ^{n+1}))}{e\quad 1+\frac{2\quad _{min}(1\quad ^{n+1})}{2\ 2_{min}(1\quad ^{n+1})}}}^2$$

$$>1 \quad c\ 1+\ = 1 \quad c\ \frac{1+\ c}{2\ c} = 1 \quad \frac{c}{2} = 0{:}5_{min}(1\quad ^{n+1}) =$$

$$(_{min}(1\quad))^2$$

$$'_3 = \frac{}{L(1+\ '^2 u_{cs})} = O\frac{}{\log(jSj)(1+\ _{cs})} \quad O\frac{}{\log(jSj)} = O\frac{}{\log(jSj)\ 1} :$$

Using $'_2$, we have

$$w_t = 1 \quad \frac{'^t_2}{2} = \begin{pmatrix} 1 & 0 \\ B & B \\ @ & =2+ \end{pmatrix}^t \frac{0{:}25e^{1=4}(2\quad _{min}\ 1\quad ^{n+1}))}{e\quad 1+\frac{2\quad _{min}(1\quad ^{n+1})}{2\ 2_{min}(1\quad ^{n+1})}}^2 C \quad C \quad A : \tag{125}$$

Further, we have

$$l_{cm} = (1+\ l^2_{cs})^{1=2} = (1)$$
$$u_{cm} = (1+\ u^2_{cs})^{1=2} = (1)$$

Since TV-divergence is upper bounded with 1, we have $m = O(1)$. By Lemma C.7, we have

$$A_1 = A_2 = 1+(1+\ ) \begin{cases} \frac{1(l_{ma})^n}{l_{max}} & \text{if } l_{max} = 1 \\ & \text{o.w.} \end{cases} = O(l^n_{max})^{1}_{1}$$

and $A_1 = A_2 =$ (1),

$$B = \frac{2l_{max}}{1} \begin{cases} \frac{1\ (l_{max})\ 1^n}{l_{max}} & \text{if } l_{max} = 1 \\ & \text{o:w:} \end{cases} = O\frac{\frac{max}{1}}{1} ;$$

and $B =$ $\quad ^2 \quad n\ 1$ (1). Hence $m_1 = \frac{2A\ m}{n} = O(l_{max})$. Also, we have $m_2 = 0$.

We choose the D-norm in Lemma B.9 as the 2-norm $k\ k_2$. Hence, by primary norm equivalence, we have $l_{cD} = \frac{1}{j\ Sj}$, and $u_{cD} = 1$, and hence $\frac{u_{cD}}{l_{cD}} = \sqrt{jSj}$.

We can evaluate the rest of the constants as follows

$$_1 = _4 = _6 = \ ,\ _2 = \sqrt{10} = \frac{}{(\sqrt{\ _{min}(1\ -\ )})}$$

$$_2 = \sqrt{\frac{}{10\ 2LA^2}} \ ,\ l^2_{cs} = \sqrt{\frac{}{\ _{min}(1\quad)}} \ \sqrt{\frac{}{\ _{in}(1\quad)}} \ _{min}\ 1\quad): \frac{}{\log(jSj)} = \frac{(\quad m\quad)}{\log(jSj)} ;$$

and similarly

$$_3 = \sqrt{\frac{}{'^2_{10}\ L(A_1+\ 1)}} \ \frac{l^2_{cs}}{} = \sqrt{\frac{}{\ _{min}(1\quad)}} \ \frac{}{\log(jSj)} :$$

$$C_1 = 3 \sqrt{A_2^2 + 1} = O\ I_{max}^{n}{}^{1};$$

and

$$C_1 =$$
(1);

$$C_2 = 3C_1 + 8 = O\left(\log_{\gamma_{\min}}\right);$$

$$C_3 = \frac{m_2 L}{l_{cs}^2} = 0;$$

$$C_4 = \frac{L^2}{2\,l_{cs}^4} + \frac{2LA_2}{l_{cs}^2}\left(\frac{1}{2} + \frac{u_{cm}^2}{2}\right) + \frac{L(A_1+1)}{l_{cs}^2}\left(\frac{u_{cm}^2}{2} + \frac{3}{2}\right) + \frac{3m_1L^2}{l_{cs}^2}$$
$$= O\left(\frac{\log^2(\epsilon)}{m_2^2 n(1-\gamma_{\min})^2} + \frac{\log(\epsilon)\log(\epsilon\gamma_{\min}^{n-1})}{m_2 n(1-\gamma_{\min})^2} + \frac{\log(\epsilon)\log(\epsilon\gamma_{\min}^{n-1})}{m_2 n(1-\gamma_{\min})^2} + \frac{\log(\epsilon)}{n(1-\gamma_{\min})}\log(\epsilon)\right)$$
$$= O\left(\frac{\log_2(\epsilon)\log(\epsilon\gamma_{\min}^{n-1})}{2\,m_2 n(1-\gamma_{\min})^2}\right);$$

$$C_5 = \frac{3u_{2m}^2}{2} + 2\,\frac{3L(A_1+1)}{l_{cs}^2} + \frac{LA_2}{l_{cs}^2} + \frac{L^2u_{2m}^2}{2l_{cs}^2} + \frac{3A^2L^2}{2l_{cs}^2}$$
$$= O\left(\frac{\log(\epsilon)}{m_3 n(1-\gamma_{\min})^2}\frac{\log(\epsilon)\log(\epsilon\gamma_{\min}^{n-1})}{(1-\gamma_{\min})} + \frac{\log(\epsilon)\log^4(\epsilon\gamma_{\min}^{n-1})}{m_3 n(1-\gamma_{\min})^3} + \frac{\log^2(\epsilon)}{n(1-\gamma_{\min})}\log(\epsilon)\right)$$
$$= O\left(\frac{\log^2(\epsilon)(1-\gamma_{\max})^2}{m_3 n(1-\gamma_{\min})^3}\right);$$

$$C_6 = \frac{3u_{cm}^2}{2^2} + 2\,\frac{3L(A_1+1)}{l_{cs}^2} + \frac{m_1 L}{2l_{cs}^2} = O\left(\frac{\log(\epsilon)}{m_2 n(1-\gamma_{\min})^2}\frac{\log(\epsilon)\log(\epsilon\gamma_{\min}^{n-1})}{(1-\gamma_{\min})}\log(\epsilon)\right)$$
$$= O\left(\frac{\log_2^3(\epsilon)\log(\epsilon\gamma_{\min}^{n-1})}{3\,m_3 n(1-\gamma_{\min})^3}\right);$$

$$C_7 = \frac{m_2^2 u_{cm}^2 L^2}{2(1-\gamma_{cs}^4)^2} = 0;$$

$$C_{mix} = (1-\gamma)\frac{1}{2}\frac{3L}{l_{cs}}\log(\epsilon) = O\left(\frac{\tau^2 l_{cs}^2}{\epsilon^8}\right);$$

$$C_9 = \frac{8u_{cD}B}{l_{cD}} = O\left(\sqrt{|S||A|}\,\log_{\gamma_{\max}}\frac{1}{n}\right)$$

$$C_{10} = \frac{8m_2 u_{cD}}{l_{cD}(1-\gamma)\sqrt{n}} = 0;$$

$$C_{11} = \frac{8C_1^2 C_3^2 u_{cm}^2}{6} = 0;$$

$$C_{12} = \frac{8C_4 u_{2D}B^2}{l_{cD}^2} = O\left(\frac{\log_2(\epsilon)\log(\epsilon\gamma_{\max}^{n-1})}{2}\frac{1}{m_3 n(1-\gamma_{\min})^2}\right) = O\left(\frac{|S||A|\log_2(\epsilon)\log(\epsilon\gamma_{\max}^{n-1})}{(1-\gamma)^2}\right);$$

$$C_{13} = \frac{14C_4 u_{cD}^2 m_2^2}{l_{cD}^2(1-\gamma^2)} = 0.$$

$$C_{14}() = C_7 + C_{11} + 0{:}5C_3^2C_9^2 + C_3C_{10} + 2C_1C_3C_{10} + 3A_1C_3 + C_{13} + C_8\frac{u_{cD}^2}{l^2}2m_2^2{}_2 = 0;_{cD}$$

$$C_{16}() = C_8\ u\ {}_{l_c}\frac{B^2}{{}_{cD}\ 2} + 2^{+1}C_{12}{}^2 = O\ \frac{jSj\log(jSj)l^{2n}{}_x + 1 + ma}{(1\quad)^3{}_{min}}\ \frac{jSj\log_(jSj)l^{3n}\ {}_{22}^{1}}{{}_{min}^{2}(1\quad)^4}{}^{max}$$

$$= O\ \frac{jSj\log_2(jSj)l_{max1}{}^{2}}{(1\quad)^4{}_mi_n{}^2};$$

$$C_{17} = (3A_1C_3 + 8A_1^2C_4 + C_5 + C_6)$$

$$= O\ \ 0 + l_{max} : {}^{2}\frac{n\ {}^{2}{}_{2}}{(jSj)^2l_{max}{}_mi_n(1\quad)^2}\log_(jSj)l_{max}{}^{2}\frac{{}^{2n}\log_{}^{2}(jSj)l_{max}{}^{2}}{{}_mi_n(1\quad)^3} + \frac{{}^{n}\log^{1}}{{}_mi_n(1\quad)^3}$$

$$= O\ \frac{l_{max}^{3n}{}^3\log^2(jSj)}{{}_m^3i_n(1\quad)^3} :$$

Similar to $k\ k_D = k\ k_2$, we have $l_{c2} = \rho_j\ \frac{1}{S^j}$ and $u_{c2} = 1$.

$$\tilde{A}_1 = \frac{2A_1^2u_{c2}^2}{l_{c2}^2} = O\ l_{max}^{2n}{}^2jSj;$$

$$A_2 = \frac{2A_{2\ c2}}{{}^{2}{}_{c2}^{2}u^2} = O\ l^{2n}{}^2jSj;$$

$$\tilde{} \quad \frac{}{l^2}\ {}^{max}\ B$$

$$= \tilde{}\ \frac{u_{c2}^2}{lc_2{}^2B^2} = O\ \frac{jSjl_{max}^{2n}}{(1\quad)^{-2}};$$

$$M_0 = \frac{}{l_{cm}^3}\ \frac{1}{C_1^2}B + (A_2 + 1)k_0k_c + {}^{2}\frac{B}{2C_1} + k_0k^2 = O\ {}_c\ \frac{!}{1^{-}l_{max}^{n} + (l_{max}) : 1^{-}l_{max}^{n}{}^2 + 1}\ !$$

$$= O\ \frac{}{(1\ {}^{4n})^2}{}^{max}$$

$$C_1 = 16u_{cm}^2M_0\ \log\ {}_{(l}\frac{1}{e^{+}} , \frac{1}{2}) = O\ \frac{l_{max}^{4n}{}^2}{(1\quad)^{-2}} : \frac{1}{{}_{min}(1\quad)} = O\ \frac{}{(1\quad)_{min}^{-x}}\ \frac{l_{ma}^{4n}{}^2}{{}^{3}}$$

$$C_2 = \frac{8u_{2m}\ C_8 + {}^{1} + C_{12}}{{}_{min}^{2}(1\quad)\ {}_{mn}(0}\ \frac{1}{\quad)}\ \frac{\log(jSj)}{{}_mi_n(1\quad)^4} + 1 + \frac{jSj\log_2(jSj)l^{3n}{}^{1}{}'_2}{2}\ \frac{}{{}^{max}}$$

$$= O\ \frac{jSj\log_2(jSj)l^{3n}{}^1}{{}_{m2n}i_n(1\quad)^4{}^{max}}$$

$$C_3 = \frac{80\tilde{B}^2C_1{}_7u_{2m}1 + {}^{B(1}{}_{4m_2})}{{}_{min}(1\quad)\ (1\quad)^4}\ \frac{1}{\quad)}\ \frac{jSj^2l^{4n}{}_x\ l^{3n}{}^3\log_(jSj)\ {}'_2}{{}_2}$$

$$= O\ \frac{l^{7n}{}^3jSj^2\log_(jSj)}{{}_{max}i_n(1\quad)^8}\ \frac{}{{}_4}\ ;$$

$$C_4 = 8u_{2m}\ C_7 + C_{11} + 0{:}5C_3{}_2C_9{}_2 + C_3C_{10} + 2C_1C_3C_{10} + 3A_1C_3 + C_{13} = '_2 = 0:$$

Finally, for the sample complexity result, we simply employ Corollary B.1.1.

□

D. Federated Q-learning

In this section, we verify that the federated Q-learning algorithm 3 satisfies the properties of the FedSAM Algorithm 4. In the following, Q^* is the solution to the Bellman optimality equation (126)

$$Q(s,a) = R(s,a) + E_{s' \sim P(\cdot|s,a)} \max_{a'} Q(s',a') \quad . \tag{126}$$

Note that Q^* is independent of the sampling policy of the agent. Furthermore, $\|Q\|_c = \|Q\|_1$.

Proposition D.1. Federated Q-learning algorithm 3 is equivalent to the FedSAM Algorithm 4 with the following parameters.

1. $\theta_t^i = Q_t^i - Q^*$
2. $S_t = (S_t^1, \dots, S_t^N)$ and $A_t = (A_t^1, \dots, A_t^N)$
3. $y_t^i = (S_t^i, A_t^i, S_{t+1}^i, A_{t+1}^i)$ and $Y_t = (S_t, A_t, S_{t+1}, A_{t+1})$
4. π^i : Stationary distribution of the sampling policy of the i'th agent.
5. $G^i(\theta^i, y^i)_{(s,a)} = \theta^i(s,a) + 1_{\{S_t^i = s, A_t^i = a\}} \max_{a'} \theta^i(S_{t+1}, a') - \theta^i(S_t^i, A_t^i) - \max_{a'} Q(S_{t+1}, a') - Q(S_t^i, A_t^i)$

$b^i(y^i)_{(s,a)} = 1_{\{S_t^i = s, A_t^i = a\}} R(S_t^i, A_t^i) + \max_{a'} Q(S_{t+1}, a') - Q(S_t^i, A_t^i)$

where 1_A is the indicator function corresponding to set A, such that $1_A = 1$ is A is true, and 0 otherwise.

Lemma D.1. Consider the federated Q-learning Algorithm 3 as a special case of FedSAM (as specified in Proposition D.1). Suppose the trajectory $\{S_t^i, A_t^i\}_{t=0,1,\dots}$ converges geometrically fast to its stationary distribution as follows $d_{TV}(P(S_t^i = \cdot, A_t^i = \cdot | S_0^i, A_0^i) \| \pi^i(\cdot,\cdot)) \le m\rho^t$ for all $i = 1, 2, \dots, N$. The corresponding $G^i(\cdot)$ in Assumption 6.1 for the federated Q-learning is as follows

$$G^i(\theta)_{(s,a)} = \theta(s,a) + \pi^i(s,a)\left[E_{s' \sim P(\cdot|s,a)} \max_{a'}(\theta + Q)(s',a')) - (\theta + Q)(s,a) - \max_{a'} Q(s',a')\right].$$

Furthermore, we have $m_1 = 2A_2 m$, where A_2 is specified in Lemma D.3, $m_2 = 0$, and $\rho = \rho$.

Lemma D.2. Consider the federated Q-learning as a special case of FedSAM (as specified in Proposition D.1). The corresponding contraction factor κ_c in Assumption 6.2 for this algorithm is $\kappa_c = (1 - (1-\gamma)^{\pi_{min}})$, where $\pi_{min} = \min_{s,a,i} \pi^i(s,a)$

Lemma D.3. Consider the federated Q-learning as a special case of FedSAM (as specified in Proposition D.1). The constants A_1, A_2, and B in Assumption 6.2 are as follows: $A_1 = A_2 = 2$ and $B = \frac{2}{1-\gamma}$.

Lemma D.4. Consider the federated Q-learning as a special case of FedSAM (as specified in Proposition D.1). Assumption 6.4 holds for this algorithm.

D.1. Proofs

Proof of Proposition D.1. Items 1-4 are by definition. Furthermore, by the update of the Q-learning, and subtracting Q^* from both sides, we have

$$\underbrace{\theta_{t+1}^i(s,a)}_{\theta_{t+1}^i(s,a)} - Q^*(s,a) = \underbrace{\theta_t^i(s,a)}_{\theta_t^i(s,a)} - Q^*(s,a)$$

$$+ 1_{\{(s,a)=(S_t^i,A_t^i)\}} \left[R(S_t^i, A_t^i) + \max_a Q_t(S_{t+1}^i, a) - Q_t(S_t^i, A_t^i)\right]$$

$$= \theta_t^i(s,a)$$

$$+ 1_{\{(s,a)=(S_t,A_t)\}} \Big[R(S_t^i, A_t^i) + \max_a \big[Q_t(S_{t+1}^i, a) - Q^*(S_{t+1}^i, a)\big] + Q^*(S_{t+1}^i, a^*) - Q^*(S_t^i, A_t^i)\Big]$$

$$\underbrace{\qquad\qquad}_{\theta_t^i(S_t^i,A_t^i)}$$

$$\left(Q^i(S^i,A^i) - Q(S^i,A^i) + Q(S^i,A^i) - Q(S^i,A^i)\right)_t^2 \quad \text{(addition and subtraction)}$$

$$= {}_t(s,a) + {}_t(s,a) + 1_{\{S_t^i=s,A_t^i=a\}}\Big[\max_{a'} {}_t + Q(S_{t+1},a') - {}_t(S_t,A_t) - \max_{a'} Q(S_{t+1},a') - {}^i(s,a)\Big]_t^2$$

$$+ 1_{\{S_t^i=s,A_t^i=a\}}\Big[R(S_t^i,A_t^i) + \max_{a'} Q(S_{t+1}^i,a') - Q(S^i,A_t^i) - {}_t\Big]^2;$$

which proves items 5 and 6. Furthermore, for the synchronization part of Q-learning, we have

$$Q_t^i - \frac{1}{N}\sum_{j=1}^N Q_t^j = \Big(Q_t^i - Q_\cdot\Big) - \frac{1}{N}\sum_{j=1}^N \big(Q_t^j - Q_\cdot\big)$$

which is equivalent to the synchronization step in FedSAM Algorithm 4. Notice that here we used the fact that all agents have the same fixed point Q. □

Proof of Lemma D.1. $G(\cdot)_{(s,a)}$ can be found by simply taking expectation of $G^i(\cdot;y_t^i)_{(s,a)}$ defined in Proposition D.1, with respect to the stationary distribution $\cdot^i$. Furthermore, we have

$$\|G^i(\cdot) - E[G^i(\cdot;y^i)]\|_c = \|E_{y_t^i}[G^i(\cdot;y^i)] - E[G^i(\cdot;y^i)]\|_c$$

$$= \Big\|\sum_{s,a,s',a'} {}^i(s,a,s',a')\,P(S^i=s,A^i=a,S_{t+1}=s',A_{t+1}=a'|S^i,A^i)\,G^i(\cdot;y^i)\Big\|_c$$

$$= \Big\|\sum_{s,a,s',a'} {}^i(s,a,s',a')\,P(S^i=s,A^i=a,S_{t+1}=s',A_{t+1}=a'|S^i,A^i)\cdot G^i(\cdot;y^i)\Big\|_c \quad (\|ax\|_c = |a|\|x\|_c)$$

$$\le \sum_{s,a,s',a'} {}^i(s,a,s',a')\,P(S^i=s,A^i=a,S_{t+1}=s',A_{t+1}=a'|S^i,A^i)\cdot\|G^i(\cdot;y^i)\|_c \quad \text{(Assumption D.3)}$$

$$= \sum_{s,a,s',a'} {}^i(s,a)P(s'|s,a)^i(a'|s')\,P(S^i=s,A^i=a|S^i,A_0^i)P(s'|s,a)^i(a'|s')\cdot A_2\|\cdot\|_c$$

(definition of transition probability)

$$= \sum_{s,a} {}^i(s,a)\,P(S_t=s,A_t=a|S_0,A_0)\cdot A_2\|\cdot\|_c$$

$$= 2d_{TV}({}^i(\cdot,\cdot),P(S^i=\cdot,A^i=\cdot|S^i,A_0^i))\cdot A_2\|\cdot\|_c$$

$$\le 2A_2\|\cdot\|_c m_t.$$

In addition, we have

$$\|E[b^i(y_t^i)]\|_c = \max_{s,a} P(S_t^i=s,A_t^i=a|S_0^i,A_0^i)\Big[R(s,a) + E_{S'\sim P(\cdot|s,a)}[\max_{a'} Q(S',a')] - Q(s,a)\Big]$$

$$= 0. \quad \text{(Bellman optimality equation (126))}$$

□

Proof of Lemma D.2.

$$\begin{aligned}
&\|G^i(\theta_1) - G^i(\theta_2)\|_c = \max_{s,a} |G^i(\theta_1)_{(s,a)} - G^i(\theta_2)_{(s,a)}| \\
&= \max_{s,a} \theta_1(s,a) + \theta^i(s,a)\left[E_{S' \sim P(\cdot|s,a)} \max_{a'}(\theta_1 + Q(S',a')) - \theta_1(s,a) - \max_{a'} Q(S',a') \right] \\
&\quad - \theta_2(s,a) - \theta^i(s,a)\left[E_{S' \sim P(\cdot|s,a)} \max_{a'}(\theta_2 + Q(S',a')) - \theta_2(s,a) - \max_{a'} Q(S',a') \right] \\
&= \max_{s,a} (1-\theta^i(s,a))(\theta_1(s,a) - \theta_2(s,a)) \\
&\quad + \theta^i(s,a) E_{S' \sim P(\cdot|s,a)} \left[\max_{a'}(\theta_1 + Q(S',a')) - \max_{a'}(\theta_2 + Q(S',a')) \right] \\
&\le \max_{s,a} (1-\theta^i(s,a))(\theta_1(s,a) - \theta_2(s,a)) \\
&\quad + \theta^i(s,a) E_{S' \sim P(\cdot|s,a)} \left| \max_{a'}(\theta_1 + Q(S',a')) - \max_{a'}(\theta_2 + Q(S',a')) \right| \quad \text{(triangle inequality)} \\
&\le \max_{s,a} (1-\theta^i(s,a)) \|\theta_1 - \theta_2\|_1 \quad \text{(definition of } \|\cdot\|_1) \\
&\quad + \theta^i(s,a) E_{S' \sim P(\cdot|s,a)} \left| \max_{a'}(\theta_1 + Q(S',a')) - \max_{a'}(\theta_2 + Q(S',a')) \right| \quad (\theta^i(s,a) \ge 0) \\
&\le \max_{s,a} (1-\theta^i(s,a)) \|\theta_1 - \theta_2\|_1 \\
&\quad + \theta^i(s,a) E_{S' \sim P(\cdot|s,a)} \max_{a'} |\theta_1 + Q(S',a') - \theta_2 - Q(S',a')| \quad \text{(Jensen's inequality)}
\end{aligned}$$

Next, we note that for any functions $f(\cdot)$ and $g(\cdot)$, we have

$$|\max_x f(x) - \max_x g(x)| \le \max_x |f(x) - g(x)|. \tag{127}$$

The reason is as follows. We have $\max_x f(x) = \max_x f(x) - g(x) + g(x) \le (\max_x f(x) - g(x)) + (\max_x g(x))$. Hence, $(\max_x f(x)) - (\max_x g(x)) \le \max_x f(x) - g(x) \le \max_x |f(x) - g(x)|$. Now suppose $\max_x f(x) \ge \max_x g(x)$. Then we can apply absolute value to the left hand side of the inequality, and we get the bound. By a similar argument for the case $\max_x f(x) \le \max_x g(x)$, we get the bound in (127). Hence, we have $|\max_{a'} (\theta_1 + Q(S',a')) - \max_{a'} (\theta_2 + Q(S',a'))| \le \max_{a'} |\theta_1(S',a') - \theta_2(S',a')| \le \|\theta_1 - \theta_2\|_1$. As a result, we have

$$\begin{aligned}
\|G^i(\theta_1) - G^i(\theta_2)\|_c &\le \max_{s,a} (1-\theta^i(s,a)) \|\theta_1 - \theta_2\|_1 + \theta^i(s,a) [\|\theta_1 - \theta_2\|_1] = \max_{s,a} (1 - (1-\theta)\theta^i(s,a)) \|\theta_1 - \theta_2\|_1 \\
&\le (1 - (1-\theta)\min_{s,a} \theta^i(s,a)) \|\theta_1 - \theta_2\|_1 \\
&\le (1 - (1-\theta)\min_i \min_{s,a} \theta^i(s,a)) \|\theta_1 - \theta_2\|_1 = (1 - (1-\theta)\theta_{\min}) \|\theta_1 - \theta_2\|_1 \\
&\le (1 - (1-\theta)\theta_{\min}) \|\theta_1 - \theta_2\|_c.
\end{aligned}$$

Proof of Lemma D.3. First, for A_1, we have

$$\|G^i(\theta_1; y) - G^i(\theta_2; y)\|_c$$

□

$$= \max_{s,a} |\mathcal{T}_1(s,a) + \mathbb{1}_{\{S=s,A=a\}}(\max_{a'}(\mathcal{T}_1 + Q(S',a')) - \mathcal{T}_1(S,A)) - \max_{a'} Q(S',a')|$$
$$|\mathcal{T}_2(s,a) + \mathbb{1}_{\{S=s,A=a\}}(\max_{a'}(\mathcal{T}_2 + Q(S',a')) - \mathcal{T}_2(S,A)) - \max_{a'} Q(S',a')|$$
$$\max_{s,a} |(1 - \mathbb{1}_{\{S=s,A=a\}})(\mathcal{T}_1(s,a) - \mathcal{T}_2(s,a))|$$
$$+ \mathbb{1}_{\{S=s,A=a\}} |\max_{a'}(\mathcal{T}_1 + Q(S',a')) - \max_{a'}(\mathcal{T}_2 + Q(S',a'))| \quad \text{(triangle inequality)}$$
$$\max_{s,a} \|\mathcal{T}_1 - \mathcal{T}_2\| + \mathbb{1}_{\{S=s,A=a\}} |\max_{a'}(\mathcal{T}_1 + Q(S',a')) - \max_{a'}(\mathcal{T}_2 + Q(S',a'))| \quad \text{(definition of } \|\mathcal{T}_1 - \mathcal{T}_2\|)$$
$$\max_{s,a} \|\mathcal{T}_1 - \mathcal{T}_2\| + \mathbb{1}_{\{S=s,A=a\}} \|\mathcal{T}_1 - \mathcal{T}_2\| \quad \text{(By (127))}$$
$$2\|\mathcal{T}_1 - \mathcal{T}_2\|$$
$$= 2\|\mathcal{T}_1 - \mathcal{T}_2\|_c$$

Second, for $\mathcal{A}_2$, we have

$$\|G^i(\mathcal{T},y) - G^i(\mathcal{T}',y)\|_c = \max_{s,a} |\mathcal{T}(s,a) + \mathbb{1}_{\{S=s,A=a\}}(\max_{a'}(\mathcal{T} + Q(S',a')) - \mathcal{T}(S,A)) - \max_{a'} Q(S',a') - \max_{a'}(\mathcal{T}' + Q(S',a')) + \max_{a'} Q(S',a')|$$
$$\max_{s,a} |(1 - \mathbb{1}_{\{S=s,A=a\}})(\mathcal{T}(s,a) - \mathcal{T}'(s,a)) + \mathbb{1}_{\{S=s,A=a\}}(\max_{a'}(\mathcal{T} + Q(S',a')) - \max_{a'}(\mathcal{T}' + Q(S',a')))| \quad \text{(triangle inequality)}$$
$$\max_{s,a} \|\mathcal{T} - \mathcal{T}'\| + \mathbb{1}_{\{S=s,A=a\}} |\max_{a'}(\mathcal{T} + Q(S',a')) - \max_{a'}(\mathcal{T}' + Q(S',a'))| \quad \text{(definition of } \|\mathcal{T} - \mathcal{T}'\|)$$
$$\max_{s,a} \|\mathcal{T} - \mathcal{T}'\| + \mathbb{1}_{\{S=s,A=a\}} \|\mathcal{T} - \mathcal{T}'\| \quad \text{(By (127))}$$
$$2\|\mathcal{T} - \mathcal{T}'\|$$
$$= 2\|\mathcal{T} - \mathcal{T}'\|_c.$$

Lastly, for $\mathcal{B}$, we have

$$\|b^i(y^i) - b^i(y^i)\|_c = \max_{s,a} \mathbb{1}_{\{S=s,A=a\}} |R(S,A) + \max_{a'} Q(S',a') - Q(S,A)|$$
$$\max_{s,a} \mathbb{1}_{\{S=s,A=a\}} |R(S,A)| + \max_{a'} |Q(S',a')| + |Q(S,A)| \quad \text{(triangle inequality)}$$
$$\max_{s,a} 1 + 1 \quad \frac{1}{1} + \frac{1}{1}$$
$$= \frac{2}{1}.$$

Proof of Lemma D.4. The proof follows similar to Lemma C.4. □

□

Proof of Theorem 5.2. By Proposition D.1 and Lemmas D.1, D.2, D.3, and D.4, it is clear that the federated Q Algorithm 3 satisfies all the Assumptions 6.1, 6.2, 6.3, and 6.4 of the FedSAM Algorithm 4. Furthermore, by the proof of Theorem B.1, we have $w_t = (1 - \frac{'_2}{2})^t$, and the constant c in the sampling distribution q^c_T in Algorithm 1 is $c = (1 - \frac{'_2}{2})^{-1}$. In equation (128) we evaluate the exact value of w_t.

Furthermore, by choosing step size α small enough, we can satisfy the requirements in (21), (23), (32), (35). By choosing K large enough, we can satisfy $K > \tau$, and by choosing T large enough we can satisfy $T > K + \tau$. Hence, the result of Theorem B.1 holds for this algorithm.

Next, we derive the constants involved in Theorem B.1 step by step. In this analysis we only consider the terms involving $|\mathcal{S}|$, $|\mathcal{A}|$, $\frac{1}{1-\gamma}$, $l_{\max}$, and $\mu_{\min}$. Since $\|\mathcal{T} - \mathcal{T}'\|_c = \|\mathcal{T} - \mathcal{T}'\|_1$, we choose $g(\mathcal{T}) = \frac{1}{2}\|\mathcal{T}\|_p^2$, i.e. the p-norm with $p = 2\log(|\mathcal{S}|)$.

It is known that $g()$ is $(p \quad 1)$ smooth with respect to $k \quad k_p$ norm (Beck, 2017), and hence $L = (\log(jSj))$. Hence, we have $l_{cs} = jSj^{1=p} \frac{p=}{e} \frac{1}{} = (1)$ and $u_{cs} = 1$. Therefore, we have $'_1 = \frac{1+}{1} \frac{u_{cs}^2}{2} = \frac{1+}{1+\frac{p}{e}} \quad 1 + \quad$. By choosing $= (\frac{1+c}{2c})^2 \quad 1 = \frac{1+2 \quad 3^2}{4 \quad c^2} (1 \quad c) = \min(1 \quad) = (\min(1 \quad))$, which is $\frac{(\frac{c}{2c})^2}{1+}$ $= O(1)$, we have $'_1 = \frac{1+}{1+\frac{p}{e}} = e^{p} \quad e^{+(\frac{1+c}{2c})^2 \quad 1} = O(1)$, and

$$'_2 = 1 \quad {}_c \sqrt{\frac{1+}{1+\frac{p}{e}}} = 1 \quad {}_c \sqrt{\frac{(\frac{1}{2}+\frac{c}{2})^2}{1+\frac{(\frac{c}{2c})^2}{ep}}} = 1 \quad \sqrt{\frac{0:5(1+{}_c)e^{1=4}}{p\overline{e} \quad 1+\frac{1+c}{2c}}} = 1 \quad \sqrt{\frac{0:5e^{1=4}(2 \quad \min(1 \quad))}{p\overline{e} \quad 1+\frac{2 \quad \min(1 \quad)}{2 \quad 2\min(1 \quad)}}}$$

$$> 1 \quad {}_c \sqrt{1+\frac{}{}} = 1 \quad {}_c \frac{1+{}_c}{2} = 1 \quad \frac{c}{2} = 0:5\min(1 \quad) = (\min(1 \quad)) {}_c$$

$$'_3 = \frac{L(1+\frac{u_{cs}^2}{'_2})}{} = O\frac{\log(jSj)(1+{}_{cs} \quad)}{} \quad O\frac{\log(jSj)}{{}_c} = O_{\min(1 \quad)}\frac{\log(jSj) \quad 1}{} :$$

Using $'_2$, we have

$$w_t = 1 \quad \frac{'^{t}_2}{2} = \begin{pmatrix} 0 & & & 1 \\ B & & & t \\ 1 & B=2+ & \sqrt{\frac{0:25e^{1=4}(2 \quad \min 1 \quad))}{p\overline{e} \quad 1+\frac{2 \quad \min(1 \quad)}{2 \quad 2\min(1 \quad)}}} & C \\ @ & & & A \end{pmatrix} : \tag{128}$$

Further, we have

$$l_{cm} = (1+ \quad l_{cs}^2)^{1=2} = (1)$$
$$u_{cm} = (1+ \quad u_{cs}^2)^{1=2} = (1)$$

Since TV-divergence is upper bounded with 1, we have $m = O(1)$. By Lemma D.3, we have

$$A_1 = A_2 = 2 = O(1)$$

and $A_1 = A_2 =$

$$(1), \quad \frac{2}{} \quad \frac{1}{}$$
$$B = {}_1$$
$$= O \quad {}_1 :$$

Hence $m_1 = 2A_2 m = O(1)$. Also, we have $m_2 = 0$.

We choose the D-norm in Lemma B.9 as the 2-norm $k \quad k_2$. Hence, by primary norm equivalence, we have $l_{cD} = \frac{p}{j} \frac{1}{Sj}$, and $u_{cD} = 1$, and hence $\frac{u_{cD}}{l_{cD}} = \sqrt{jSj}$. The rest of the proof is similar to the proof of Theorem 5.1 where l_{max} is substituted with 1. The sample complexity can also be derived using Corollary B.1.1.

□