Federated Natural Policy Gradient and Actor Critic Methods for Multi-task Reinforcement Learning

Tong Yang* CMU
Shicong Cen† CMU
Yuting Wei‡ UPenn
Yuxin Chen§ UPenn
Yuejie Chi¶ CMU

October 2023; revised August 2024

Abstract

Federated reinforcement learning (RL) enables collaborative decision making of multiple distributed agents without sharing local data trajectories. In this work, we consider a multi-task setting, in which each agent has its own private reward function corresponding to different tasks, while sharing the same transition kernel of the environment. Focusing on infinite-horizon Markov decision processes, the goal is to learn a globally optimal policy that maximizes the sum of the discounted total rewards of all the agents in a decentralized manner, where each agent only communicates with its neighbors over some prescribed graph topology.

We develop federated vanilla and entropy-regularized natural policy gradient (NPG) methods in the tabular setting under softmax parameterization, where gradient tracking is applied to estimate the global Q-function to mitigate the impact of imperfect information sharing. We establish non-asymptotic global convergence guarantees under exact policy evaluation, where the rates are nearly independent of the size of the state-action space and illuminate the impacts of network size and connectivity. To the best of our knowledge, this is the first time that near dimension-free global convergence is established for federated multi-task RL using policy optimization. We further go beyond the tabular setting by proposing a federated natural actor critic (NAC) method for multi-task RL with function approximation, and establish its finite-time sample complexity taking the errors of function approximation into account.

Keywords: federated reinforcement learning, multi-task reinforcement learning, natural policy gradient methods, entropy regularization, dimension-free global convergence

Contents

1 Introduction **2**
1.1 Our contributions . . . 3
1.2 Related work . . . 4

2 Model and backgrounds **6**
2.1 Markov decision processes . . . 6
2.2 Entropy-regularized RL . . . 6
2.3 Natural policy gradient methods . . . 7

*Department of Electrical and Computer Engineering, Carnegie Mellon University; email: tongyang@andrew.cmu.edu.
†Department of Electrical and Computer Engineering, Carnegie Mellon University; email: shicongc@andrew.cmu.edu.
‡Department of Statistics and Data Science, Wharton School, University of Pennsylvania; email: ytwei@wharton.upenn.edu.
§Department of Statistics and Data Science, Wharton School, University of Pennsylvania; email: yuxinc@wharton.upenn.edu.
¶Department of Electrical and Computer Engineering, Carnegie Mellon University; email: yuejiechi@cmu.edu.

3 Federated NPG methods for multi-task RL **8**
3.1 Federated multi-task RL . . . 8
3.2 Proposed federated NPG algorithms . . . 9
3.3 Global convergence of FedNPG . . . 11
3.4 Global convergence of FedNPG with entropy regularization . . . 12

4 Federated NAC methods for multi-task RL **13**
4.1 Algorithm design . . . 14
4.2 Theoretical guarantees . . . 15

5 Conclusions **16**

A Convergence analysis of FedNPG **21**
A.1 Analysis of entropy-regularized FedNPG with exact policy evaluation . . . 21
A.2 Analysis of entropy-regularized FedNPG with inexact policy evaluation . . . 23
A.3 Analysis of FedNPG with exact policy evaluation . . . 25
A.4 Analysis of FedNPG with inexact policy evaluation . . . 28

B Convergence analysis of FedNAC **30**
B.1 Proof of Theorem 11 . . . 34
B.2 Proof of Theorem 10 . . . 37

C Proof of key lemmas for FedNPG **39**
C.1 Proof of Lemma 1 . . . 39
C.2 Proof of Lemma 2 . . . 44
C.3 Proof of Lemma 3 . . . 46
C.4 Proof of Lemma 4 . . . 48
C.5 Proof of Lemma 5 . . . 50
C.6 Proof of Lemma 13 . . . 52
C.7 Proof of Lemma 15 . . . 53
C.8 Proof of Lemma 16 . . . 56
C.9 Proof of Lemma 18 . . . 57

D Proof of key lemmas for FedNAC **58**
D.1 Proof of Lemma 10 . . . 58
D.2 Proof of Lemma 11 . . . 60
D.3 Proof of Lemma 12 . . . 64

1 Introduction

Federated reinforcement learning (FRL) is an emerging paradigm that combines the advantages of federated learning (FL) and reinforcement learning (RL) (Qi et al., 2021; Zhuo et al., 2019), allowing multiple agents to learn a shared policy from local experiences, without exposing their private data to a central server nor other agents. FRL is poised to enable collaborative and efficient decision making in scenarios where data is distributed, heterogeneous, and sensitive, which arise frequently in applications such as edge computing, smart cities, and healthcare (Wang et al., 2023, 2020; Zhuo et al., 2019), to name just a few. As has been observed (Lian et al., 2017), decentralized training can lead to performance improvements in FL by avoiding communication congestions at busy nodes such as the server, especially under high-latency scenarios. This motivates us to design algorithms for the fully decentralized setting, a scenario where the agents can only communicate with their local neighbors over a prescribed network topology.[1]

In this work, we study the problem of *federated multi-task reinforcement learning* (Anwar and Raychowdhury, 2021; Qi et al., 2021; Yu et al., 2020), where each agent collects its own reward — possibly unknown

[1] Our work seamlessly handles the server-client setting as a special case, by assuming the network topology as a fully connected network.

to other agents — corresponding to the local task at hand, while having access to the same dynamics (i.e., transition kernel) of the environment. The collective goal is to learn a shared policy that maximizes the total rewards accumulated from all the agents; in other words, one seeks a policy that performs well in terms of overall benefits, rather than biasing towards any individual task, achieving the Pareto frontier in a multi-objective context. There is no shortage of application scenarios where federated multi-task RL becomes highly relevant. For instance, in healthcare (Zerka et al., 2020), different hospitals may be interested in finding an optimal treatment for all patients without disclosing private data, where the effectiveness of the treatment can vary across different hospitals due to demographical differences. As another potential application, to enhance ChatGPT's performance across different tasks or domains (M Alshater, 2022; Rahman et al., 2023), one might consult domain experts to chat and rate ChatGPT's outputs for solving different tasks, and train ChatGPT in a federated manner without exposing private data or feedback of each expert.

Nonetheless, despite the promise, provably efficient algorithms for federated multi-task RL remain substantially under-explored, especially in the fully decentralized setting. The heterogeneity of local tasks leads to a higher degree of disagreements between the global value function and local value functions of individual agents. Due to the lack of global information sharing, care needs to be taken to judiciously balance the use of neighboring information (to facilitate consensus) and local data (to facilitate learning) when updating the policy. To the best of our knowledge, very limited algorithms are currently available to find the global optimal policy with non-asymptotic convergence guarantees even for tabular infinite-horizon Markov decision processes.

Motivated by the connection with decentralized optimization, it is tempting to take a policy optimization perspective to tackle this challenge. Policy gradient (PG) methods, which seek to learn the policy of interest via first-order optimization methods, play an eminent role in RL due to their simplicity and scalability. In particular, natural policy gradient (NPG) methods (Amari, 1998; Kakade, 2001) are among the most popular variants of PG methods, underpinning default methods used in practice such as trust region policy optimization (TRPO) (Schulman et al., 2015) and proximal policy optimization (PPO) (Schulman et al., 2017). On the theoretical side, it has also been established recently that the NPG method enjoys fast global convergence to the optimal policy in an almost dimension-free manner (Agarwal et al., 2021; Cen et al., 2021), where the iteration complexity is nearly independent of the size of the state-action space. These benefits can be translated to their sample-based counterparts such as the natural actor critic (NAC) method (Xu et al., 2020), where the value functions are learned via temporal difference learning. Inspired by the efficacy of NPG methods, it is natural to ask:

*Can we develop **federated** variants of NPG and NAC methods in the fully decentralized setting, that come with **non-asymptotic and finite-sample global convergence** guarantees for multi-task RL?*

1.1 Our contributions

Focusing on infinite-horizon Markov decision processes (MDPs), we provide an affirmative answer to the above question, by developing federated NPG (FedNPG) methods for solving both the vanilla and entropy-regularized multi-task RL problems with finite-time global convergence guarantees. While entropy regularization is often incorporated as an effective strategy to encourage exploration during policy learning, solving the entropy-regularized RL problem is of interest in its own right, as the optimal regularized policy possesses desirable robust properties with respect to reward perturbations (Eysenbach and Levine, 2021; McKelvey and Palfrey, 1995).

Due to the multiplicative update nature of NPG methods under softmax parameterization, it is more convenient to work with the logarithms of local policies in the decentralized setting. In each iteration of the proposed FedNPG method, the logarithms of local policies are updated by a weighted linear combination of two terms (up to normalization): a gossip mixing (Nedic and Ozdaglar, 2009) of the logarithms of neighboring local policies, and a local estimate of the global Q-function tracked via the technique of dynamic average consensus (Zhu and Martínez, 2010), a prevalent idea in decentralized optimization that allows for the use of large constant learning rates (Di Lorenzo and Scutari, 2016; Nedic et al., 2017; Qu and Li, 2017) to accelerate convergence. We further develop sample-efficient federated NAC (FedNAC) methods that allow for both stochastic updates and function approximation. Our contributions are as follows.

- We propose FedNPG methods for both the vanilla and entropy-regularized multi-task RL problems,

setting	algorithms	iteration complexity	optimality criteria
unregularized	NPG (Agarwal et al., 2021)	$\mathcal{O}\left(\frac{1}{(1-\gamma)^2\varepsilon}+\frac{\log\lvert\mathcal{A}\rvert}{\eta\varepsilon}\right)$	$V^\star - V^{\pi^{(t)}} \le \varepsilon$
	FedNPG (ours)	$\mathcal{O}\left(\frac{\sigma\sqrt{N}\log\lvert\mathcal{A}\rvert}{(1-\gamma)^{\frac{9}{2}}(1-\sigma)\varepsilon^{\frac{3}{2}}}+\frac{1}{(1-\gamma)^2\varepsilon}\right)$	$\frac{1}{T}\sum_{t=0}^{T-1}\left(V^\star - V^{\bar{\pi}^{(t)}}\right) \le \varepsilon$
regularized	NPG (Cen et al., 2021)	$\mathcal{O}\left(\frac{1}{\tau\eta}\log\left(\frac{1}{\varepsilon}\right)\right)$	$V_\tau^\star - V_\tau^{\pi^{(t)}} \le \varepsilon$
	FedNPG (ours)	$\mathcal{O}\left(\max\left\{\frac{1}{\tau\eta},\frac{1}{1-\sigma}\right\}\log\left(\frac{1}{\varepsilon}\right)\right)$	$V_\tau^\star - V_\tau^{\bar{\pi}^{(t)}} \le \varepsilon$

Table 1: Iteration complexities of NPG and FedNPG (ours) methods to reach ε-accuracy of the vanilla and entropy-regularized problems, where we assume exact gradient evaluation, and only keep the dominant terms w.r.t. ε. The policy estimates in the t-iteration are $\pi^{(t)}$ and $\bar{\pi}^{(t)}$ for NPG and FedNPG, respectively, where T is the number of iterations. Here, N is the number of agents, $\tau \le 1$ is the regularization parameter, $\sigma \in [0,1]$ is the spectral radius of the network, $\gamma \in [0,1)$ is the discount factor, $|\mathcal{A}|$ is the size of the action space, and $\eta > 0$ is the learning rate. The iteration complexities of FedNPG reduce to their centralized counterparts when $\sigma = 0$. For vanilla FedNPG, the learning rate is set as $\eta = \eta_1 = \mathcal{O}\left(\frac{(1-\gamma)^9(1-\sigma)^2\log|\mathcal{A}|}{TN\sigma}\right)^{1/3}$; for entropy-regularized FedNPG, the learning rate satisfies $0 < \eta < \eta_0 = \mathcal{O}\left(\frac{(1-\gamma)^7(1-\sigma)^2\tau}{\sigma N}\right)$.

where each agent only communicates with its neighbors and performs local computation using its own reward or task information.

- Assuming access to exact policy evaluation, we establish that the average iterate of vanilla FedNPG converges globally at a rate of $\mathcal{O}(1/T^{2/3})$ in terms of the sub-optimality gap for the multi-task RL problem, and that the last iterate of entropy-regularized FedNPG converges globally at a linear rate to the regularized optimal policy. Our convergence theory highlights the impacts of all salient problem parameters (see Table 1 for details), such as the size and connectivity of the communication network. In particular, the iteration complexities of FedNPG are again almost independent of the size of the state-action space, which recover prior results on the centralized NPG methods when the network is fully connected.

- We further demonstrate the stability of the proposed FedNPG methods when policy evaluations are only available in an inexact manner. To be specific, we prove that their convergence rates remain unchanged as long as the approximation errors are sufficiently small in the ℓ_∞ sense.

- We go beyond the tabular setting by proposing FedNAC— a federated actor critic method for multi-task RL with function approximation — and establish a finite-sample sample complexity on the order of $\mathcal{O}(1/\varepsilon^{7/2})$ for each agent in terms of the expected sub-optimality gap.

To the best of our knowledge, the proposed federated NPG and NAC methods are the first policy optimization methods for multi-task RL that achieve near dimension-free global convergence guarantees in terms of iteration and sample complexities, allowing for fully decentralized communication without any need to share local reward/task information.

1.2 Related work

Global convergence of NPG methods for tabular MDPs. Agarwal et al. (2021) first establishes a $\mathcal{O}(1/T)$ last-iterate convergence rate of the NPG method under softmax parameterization with constant step size, assuming access to exact policy evaluation. When entropy regularization is in place, Cen et al. (2021) establishes a global linear convergence to the optimal regularized policy for the entire range of admissible constant learning rates using softmax parameterization and exact policy evaluation, which is further shown to be stable in the presence of ℓ_∞ policy evaluation errors. The iteration complexity of NPG methods is

nearly independent with the size of the state-action space, which is in sharp contrast to softmax policy gradient methods that may take exponential time to converge (Li et al., 2023c; Mei et al., 2020). Lan (2023) proposed a more general framework through the lens of mirror descent for regularized RL with global linear convergence guarantees, which is further generalized in Zhan et al. (2023); Lan et al. (2023a). Earlier analysis of regularized MDPs can be found in Shani et al. (2020). Besides, Xiao (2022) proves that vanilla NPG also achieves linear convergence when geometrically increasing learning rates are used; see also Khodadadian et al. (2021); Bhandari and Russo (2021). Zhou et al. (2022) developed an anchor-changing NPG method for multi-task RL under various optimality criteria in the centralized setting.

Convergence and sample complexity bounds of NAC. The convergence and sample complexity of a variety of natural actor-critic methods (NACs) are extensively studied in the literature (Bhatnagar et al., 2009; Wang et al., 2019; Khodadadian et al., 2022a; Agarwal et al., 2021; Yuan et al., 2022). More pertinent to our work, Agarwal et al. (2021) introduced Q-NPG — a sample version of the NPG method with function approximation under softmax parameterization — and obtained a convergence rate of $\mathcal{O}(1/\sqrt{T})$. Yuan et al. (2022) weakens some of its assumptions and improves the convergence rate to $\mathcal{O}(1/T)$ and gives the $\widetilde{\mathcal{O}}(1/\varepsilon^3)$ sample complexity using a constant actor learning rate. The FedNAC method we propose in this paper can be seen as a decentralized version of Q-NPG, and in the server-client setting where the network is fully connected, our convergence rate and sample complexity match those in Yuan et al. (2022).

Distributed and federated RL. There have been a variety of settings being set forth for distributed and federated RL. Mnih et al. (2016); Espeholt et al. (2018); Assran et al. (2019); Khodadadian et al. (2022b); Woo et al. (2023) focused on developing federated versions of RL algorithms to accelerate training, assuming all agents share the same transition kernel and reward function; in particular, Khodadadian et al. (2022b); Woo et al. (2023, 2024) established the provable benefits of federated learning in terms of linear speedup. More pertinent to our work, Zhao et al. (2023); Anwar and Raychowdhury (2021) considered the federated multi-task framework, allowing different agents having private reward functions. Zhao et al. (2023) proposed an empirically probabilistic algorithm that can seek an optimal policy under the server-client setting, while Anwar and Raychowdhury (2021) developed new attack methods in the presence of adversarial agents. Recently Lan et al. (2023b) discussed how to avoid transmitting the Hessian matrix during communication in the server-client setting where all agents share the same reward function. Different from the FRL framework, Chen et al. (2021a,b); Omidshafiei et al. (2017); Kar et al. (2012); Chen et al. (2022); Zeng et al. (2021) considered the distributed multi-agent RL setting where the agents interact with a dynamic environment through a multi-agent Markov decision process, where each agent can have their own state or action spaces. Zeng et al. (2021) developed a decentralized policy gradient method where different agents have different MDPs, where a special case of their setting recovers ours. However, the convergence rate developed in Zeng et al. (2021) has rather pessimistic dependencies with the size of the state-action space, together with other parameters, without leveraging natural policy gradients and gradient tracking techniques.

Decentralized first-order optimization algorithms. Early work of consensus-based first-order optimization algorithms for the fully decentralized setting include but are not limited to Lobel and Ozdaglar (2008); Nedic and Ozdaglar (2009); Duchi et al. (2011). Gradient tracking, which leverages the idea of dynamic average consensus (Zhu and Martínez, 2010) to track the gradient of the global objective function, is a popular method to improve the convergence speed (Qu and Li, 2017; Nedic et al., 2017; Di Lorenzo and Scutari, 2016; Pu and Nedić, 2021; Li et al., 2020).

Notation. Boldface small and capital letters denote vectors and matrices, respectively. Sets are denoted with curly capital letters, e.g., $\mathcal{S}, \mathcal{A}$. We let $(\mathbb{R}^d, \|\cdot\|)$ denote the d-dimensional real coordinate space equipped with norm $\|\cdot\|$. The ℓ^p-norm of $\boldsymbol{v}$ is denoted by $\|\boldsymbol{v}\|_p$, where $1 \leq p \leq \infty$, and the spectral norm and the Frobenius norm of a matrix $\boldsymbol{M}$ are denoted by $\|\boldsymbol{M}\|_2$ and $\|\boldsymbol{M}\|_F$, resp. We let $[N]$ denote $\{1, \ldots, N\}$, use $\mathbf{1}_N$ to represent the all-one vector of length N, and denote by $\mathbf{0}$ a vector or a matrix consisting of all 0's. We allow the application of functions such as $\log(\cdot)$ and $\exp(\cdot)$ to vectors or matrices, with the understanding that they are applied in an element-wise manner.

2 Model and backgrounds

2.1 Markov decision processes

Markov decision processes. We consider an infinite-horizon discounted Markov decision process (MDP) denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state space and the action space, respectively, $\gamma \in [0, 1)$ indicates the discount factor, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ stands for the reward function. To be more specific, for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ and any state $s' \in \mathcal{S}$, we denote by $P(s'|s, a)$ the transition probability from state s to state s' when action a is taken, and $r(s, a)$ the instantaneous reward received in state s when action a is taken. Furthermore, a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ specifies an action selection rule, where $\pi(a|s)$ specifies the probability of taking action a in state s for each $(s, a) \in \mathcal{S} \times \mathcal{A}$.

For any given policy π, we denote by $V^\pi : \mathcal{S} \mapsto \mathbb{R}$ the corresponding value function, which is the expected discounted cumulative reward with an initial state $s_0 = s$, given by

$$\forall s \in \mathcal{S} : \quad V^\pi(s) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s\right], \tag{1}$$

where the randomness is over the trajectory generated following the policy $a_t \sim \pi(\cdot|s_t)$ and the MDP dynamic $s_{t+1} \sim P(\cdot|s_t, a_t)$. We also overload the notation $V^\pi(\rho)$ to indicate the expected value function of policy π when the initial state follows a distribution ρ over $\mathcal{S}$, namely, $V^\pi(\rho) \coloneqq \mathbb{E}_{s\sim\rho}[V^\pi(s)]$. Similarly, the Q-function $Q^\pi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ of policy π is defined by

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad Q^\pi(s, a) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s, a_0 = a\right], \tag{2}$$

which measures the expected discounted cumulative reward with an initial state $s_0 = s$ and an initial action $a_0 = a$, with expectation taken over the randomness of the trajectory. The optimal policy $\pi^\star$ refers to the policy that maximizes the value function $V^\pi(s)$ for all states $s \in \mathcal{S}$, which is guaranteed to exist (Puterman, 2014). The corresponding optimal value function and Q-function are denoted as $V^\star$ and $Q^\star$, respectively.

2.2 Entropy-regularized RL

Entropy regularization (Williams and Peng, 1991; Ahmed et al., 2019) is a popular technique in practice that encourages stochasticity of the policy to promote exploration, as well as robustness against reward uncertainties. Mathematically, this can be viewed as adjusting the instantaneous reward based the current policy in use as

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad r_\tau(s, a) \coloneqq r(s, a) - \tau \log \pi(a|s), \tag{3}$$

where $\tau \geq 0$ denotes the regularization parameter. Typically, τ should not be too large to outweigh the actual rewards; for ease of presentation, we assume $\tau \leq \min\left\{1, \frac{1}{\log|\mathcal{A}|}\right\}$ (Cen et al., 2022b). Equivalently, this amounts to the entropy-regularized (also known as "soft") value function, defined as

$$\forall s \in \mathcal{S} : \quad V_\tau^\pi(s) \coloneqq V^\pi(s) + \tau\mathcal{H}(s, \pi). \tag{4}$$

Here, we define

$$\mathcal{H}(s, \pi) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t)\big|s_0 = s\right] = \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^\pi}\left[-\sum_{a\in\mathcal{A}} \pi(a|s') \log \pi(a|s')\right], \tag{5}$$

where $d_{s_0}^\pi$ is the discounted state visitation distribution of policy π given an initial state $s_0 \in \mathcal{S}$, denoted by

$$\forall s \in \mathcal{S} : \quad d_{s_0}^\pi(s) \coloneqq (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s|s_0), \tag{6}$$

with the trajectory generated by following policy π in the MDP $\mathcal{M}$ starting from state s_0. Analogously, the regularized (or soft) Q-function Q_τ^π of policy π is related to the soft value function $V_\tau^\pi(s)$ as

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q_\tau^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \in P(\cdot|s,a)} \left[V_\tau^\pi(s') \right], \tag{7a}$$

$$\forall s \in \mathcal{S}: \quad V_\tau^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[-\tau \pi(a|s) + Q_\tau^\pi(s,a) \right]. \tag{7b}$$

The optimal regularized policy, the optimal regularized value function, and the Q-function are denoted by $\pi_\tau^\star$, $V_\tau^\star$, and $Q_\tau^\star$, respectively.

For a distribution $\rho \in \Delta(\mathcal{S})$, we define $d_\rho^\pi(s) = \mathbb{E}_{s_0 \sim \rho}[d_{s_0}^\pi(s)]$. We also define the *state-action visitation distribution* $\bar{d}_\rho^\pi$ as

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \bar{d}_\rho^\pi(s,a) \coloneqq d_\rho^\pi(s)\pi(a|s) = (1-\gamma)\mathbb{E}_{s_0 \sim \rho} \left[\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a | s_0) \right]. \tag{8}$$

Furthermore, we define the state-action visitation distribution induced by an initial state-action distribution $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$, i.e.,

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \tilde{d}_\nu^\pi(s,a) \coloneqq (1-\gamma)\mathbb{E}_{(s_0,a_0) \sim \nu} \left[\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a | s_0, a_0) \right]. \tag{9}$$

The following simple fact holds for all $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$d_\rho^\pi(s) \geq (1-\gamma)\rho(s), \quad \bar{d}_\rho^\pi(s,a) \geq (1-\gamma)\rho(s)\pi(a|s), \quad \tilde{d}_\nu^\pi(s,a) \geq (1-\gamma)\nu(s,a). \tag{10}$$

2.3 Natural policy gradient methods

Natural policy gradient (NPG) methods lie at the heart of policy optimization, serving as the backbone of popular heuristics such as TRPO (Schulman et al., 2015) and PPO (Schulman et al., 2017). Instead of directly optimizing the policy over the probability simplex, one often adopts the softmax parameterization, which parameterizes the policy as

$$\pi_\theta \coloneqq \mathrm{softmax}(\theta) \quad \text{or} \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \pi_\theta(a|s) \coloneqq \frac{\exp \theta(s,a)}{\sum_{a' \in \mathcal{A}} \exp \theta(s,a')} \tag{11}$$

for any θ: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

Vanilla NPG method. In the tabular setting, the update rule of vanilla NPG at the t-th iteration can be concisely represented as

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \pi^{(t+1)}(a|s) \propto \pi^{(t)}(a|s) \exp\left(\frac{\eta Q^{(t)}(s,a)}{1-\gamma} \right), \tag{12}$$

where $\eta > 0$ denotes the learning rate, and $Q^{(t)} = Q^{\pi^{(t)}}$ is the Q-function under policy $\pi^{(t)}$. Agarwal et al. (2021) shows that: in order to find an ε-optimal policy, NPG takes at most $\mathcal{O}\left(\frac{1}{(1-\gamma)^2 \varepsilon}\right)$ iterations, assuming exact policy evaluation.

Entropy-regularized NPG method. Turning to the regularized problem, we note that the update rule of entropy-regularized NPG becomes

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \pi^{(t+1)}(a|s) \propto (\pi^{(t)}(a|s))^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta Q_\tau^{(t)}(s,a)}{1-\gamma} \right), \tag{13}$$

where $\eta \in (0, \frac{1-\gamma}{\tau}]$ is the learning rate, and $Q_\tau^{(t)} = Q_\tau^{\pi^{(t)}}$ is the soft Q-function of policy $\pi^{(t)}$. Cen et al. (2022a) proves that entropy-regularized NPG enjoys fast global linear convergence to the optimal regularized policy: to find an ε-optimal regularized policy, entropy-regularized NPG takes no more than $\mathcal{O}\left(\frac{1}{\eta\tau} \log\left(\frac{1}{\varepsilon}\right)\right)$ iterations.

3 Federated NPG methods for multi-task RL

3.1 Federated multi-task RL

In this paper, we consider the federated multi-task RL setting, where a set of agents learn collaboratively a single policy that maximizes its average performance over all the tasks using only local computation and communication.

Multi-task RL. Each agent $n \in [N]$ has its own private reward function $r_n(s,a)$ — corresponding to different tasks — while sharing the same transition kernel of the environment. The goal is to collectively learn a single policy π that maximizes the global value function given by

$$V^\pi(s) = \frac{1}{N}\sum_{n=1}^{N} V_n^\pi(s), \tag{14}$$

where V_n^π is the value function of agent $n \in [N]$, defined by

$$\forall s \in \mathcal{S}: \quad V_n^\pi(s) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_n(s_t,a_t)|s_0 = s\right]. \tag{15}$$

Clearly, the global value function (14) corresponds to using the average reward of all agents

$$r(s,a) = \frac{1}{N}\sum_{n=1}^{N} r_n(s,a). \tag{16}$$

The global Q-function $Q^\pi(s,a)$ and the agent Q-functions $Q_n^\pi(s,a)$ can be defined in a similar manner obeying $Q^\pi(s,a) = \frac{1}{N}\sum_{n=1}^{N} Q_n^\pi(s,a)$.

In parallel, we are interested in the entropy-regularized setting, where each agent $n \in [N]$ is equipped with a regularized reward function given by

$$r_{\tau,n}(s,a) \coloneqq r_n(s,a) - \tau \log \pi(a|s), \tag{17}$$

and we define similarly the regularized value function and the global regularized value function as

$$\forall s \in \mathcal{S}: \quad V_{\tau,n}^\pi(s) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_{\tau,n}(s_t,a_t)|s_0 = s\right], \quad \text{and} \quad V_\tau^\pi(s) = \frac{1}{N}\sum_{n=1}^{N} V_{\tau,n}^\pi(s). \tag{18}$$

The soft Q-function of agent n is given by

$$Q_{\tau,n}^\pi(s,a) = r_n(s,a) + \gamma \mathbb{E}_{s' \in P(\cdot|s,a)}\left[V_{\tau,n}^\pi(s')\right], \tag{19}$$

and the global soft Q-function is given by $Q_\tau^\pi(s,a) = \frac{1}{N}\sum_{n=1}^{N} Q_{\tau,n}^\pi(s,a)$.

Federated policy optimization in the fully decentralized setting. We consider a federated setting with fully decentralized communication, that is, all the agents are synchronized to perform information exchange over some prescribed network topology denoted by an undirected weighted graph $\mathcal{G}([N],E)$. Here, E stands for the edge set of the graph with N nodes — each corresponding to an agent — and two agents can communicate with each other if and only if there is an edge connecting them. The information sharing over the graph is best described by a mixing matrix (Nedic and Ozdaglar, 2009), denoted by $\boldsymbol{W} = [w_{ij}] \in [0,1]^{N\times N}$, where w_{ij} is a positive number if $(i,j) \in E$ and 0 otherwise. We also make the following standard assumptions on the mixing matrix.

Assumption 1 (double stochasticity)**.** *The mixing matrix $\boldsymbol{W} = [w_{ij}] \in [0,1]^{N\times N}$ is symmetric (i.e., $\boldsymbol{W}^\top = \boldsymbol{W}$) and doubly stochastic (i.e., $\boldsymbol{W}\mathbf{1}_N = \mathbf{1}_N$, $\mathbf{1}_N^\top \boldsymbol{W} = \mathbf{1}_N^\top$).*

The following standard metric measures how fast information propagates over the graph.

Definition 1 (spectral radius)**.** *The spectral radius of* $\boldsymbol{W}$ *is defined as*

$$\sigma \coloneqq \left\| \boldsymbol{W} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top \right\|_2 \in [0,1). \tag{20}$$

The spectral radius σ determines how fast information propagate over the network. For instance, in a fully-connected network, we can achieve $\sigma = 0$ by setting $\boldsymbol{W} = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$. For control of $1/(1-\sigma)$ regarding different graphs, we refer the readers to paper Nedić et al. (2018). In an Erdös-Rényi random graph, as long as the graph is connected, one has with high probability $\sigma \asymp 1$. Another immediate consequence is that for any $\boldsymbol{x} \in \mathbb{R}^N$, letting $\overline{x} = \frac{1}{N}\mathbf{1}_N^\top \boldsymbol{x}$ be its average, we have

$$\|\boldsymbol{W}\boldsymbol{x} - \overline{x}\mathbf{1}_N\|_2 \le \sigma \|\boldsymbol{x} - \overline{x}\mathbf{1}_N\|_2 \ , \tag{21}$$

where the consensus error contracts by a factor of σ.

3.2 Proposed federated NPG algorithms

Assuming softmax parameterization, the problem can be formulated as decentralized optimization,

$$\text{(unregularized)} \qquad \max_\theta \ V^{\pi_\theta}(s) = \frac{1}{N}\sum_{n=1}^N V_n^{\pi_\theta}(s), \tag{22}$$

$$\text{(regularized)} \qquad \max_\theta \ V_\tau^{\pi_\theta}(s) = \frac{1}{N}\sum_{n=1}^N V_{\tau,n}^{\pi_\theta}(s), \tag{23}$$

where $\pi_\theta \coloneqq \mathrm{softmax}(\theta)$ subject to communication constraints. Motivated by the success of NPG methods, we aim to develop federated NPG methods to achieve our goal. For notational convenience, let $\boldsymbol{\pi}^{(t)} \coloneqq \left(\pi_1^{(t)}, \cdots, \pi_N^{(t)}\right)^\top$ be the collection of policy estimates at all agents in the t-th iteration. Let

$$\overline{\pi}^{(t)} \coloneqq \mathrm{softmax}\left(\frac{1}{N}\sum_{n=1}^N \log \pi_n^{(t)}\right), \tag{24}$$

which satisfies that $\overline{\pi}^{(t)}(a|s) \propto \left(\prod_{n=1}^N \pi_n^{(t)}(a|s)\right)^{1/N}$ for each $(s,a) \in \mathcal{S}\times\mathcal{A}$. Therefore, $\overline{\pi}^{(t)}$ could be seen as the normalized geometric mean of $\{\pi_n^{(t)}\}_{n\in[N]}$. Define the collection of Q-function estimates as

$$\boldsymbol{Q}^{(t)} \coloneqq \left(Q_1^{\pi_1^{(t)}}, \cdots, Q_N^{\pi_N^{(t)}}\right)^\top, \qquad \boldsymbol{Q}_\tau^{(t)} \coloneqq \left(Q_{\tau,1}^{\pi_1^{(t)}}, \cdots, Q_{\tau,N}^{\pi_N^{(t)}}\right)^\top.$$

We shall often abuse the notation and treat $\boldsymbol{\pi}^{(t)}$, $\boldsymbol{Q}_\tau^{(t)}$ as matrices in $\mathbb{R}^{N\times|\mathcal{S}||\mathcal{A}|}$, and treat $\boldsymbol{\pi}^{(t)}(a|s)$, $\boldsymbol{Q}_\tau^{(t)}(a|s)$ as vectors in $\mathbb{R}^N$, for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Vanilla federated NPG methods. To motivate the algorithm development, observe that the NPG method (cf. (12)) applied to (22) adopts the update rule

$$\pi^{(t+1)}(a|s) \propto \pi^{(t)}(a|s)\exp\left(\frac{\eta Q^{\pi^{(t)}}(s,a)}{1-\gamma}\right) = \pi^{(t)}(a|s)\exp\left(\frac{\eta\sum_{n=1}^N Q_n^{\pi^{(t)}}(s,a)}{N(1-\gamma)}\right)$$

for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Two challenges arise when executing this update rule: the policy estimates are maintained locally without consensus, and the global Q-function are unavailable in the decentralized setting. To address these challenges, we apply the idea of dynamic average consensus (Zhu and Martínez, 2010), where each agent maintains its own estimate $T_n^{(t)}(s,a)$ of the global Q-function, which are collected as vector

$$\boldsymbol{T}^{(t)} = \left(T_1^{(t)}, \cdots, T_N^{(t)}\right)^\top.$$

Algorithm 1 Federated NPG (FedNPG)

Input: learning rate $\eta > 0$, iteration number $T \in \mathbb{N}_+$, mixing matrix $\boldsymbol{W} \in \mathbb{R}^{N \times N}$.
Initialize: $\boldsymbol{\pi}^{(0)}$, $\boldsymbol{T}^{(0)} = \boldsymbol{Q}^{(0)}$.
for $t = 0, 1, \cdots T-1$ **do**
  Update the policy for each $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\log \boldsymbol{\pi}^{(t+1)}(a|s) = \boldsymbol{W}\left(\log \boldsymbol{\pi}^{(t)}(a|s) + \frac{\eta}{1-\gamma}\boldsymbol{T}^{(t)}(s,a)\right) - \log \boldsymbol{z}^{(t)}(s)\,, \tag{25}$$

  where $\boldsymbol{z}^{(t)}(s) = \sum_{a' \in \mathcal{A}} \exp\left\{\boldsymbol{W}\left(\log \boldsymbol{\pi}^{(t)}(a'|s) + \frac{\eta}{1-\gamma}\boldsymbol{T}^{(t)}(s,a')\right)\right\}$.
  Evaluate $\boldsymbol{Q}^{(t+1)}$.
  Update the global Q-function estimate for each $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\boldsymbol{T}^{(t+1)}(s,a) = \boldsymbol{W}\Big(\boldsymbol{T}^{(t)}(s,a) + \underbrace{\boldsymbol{Q}^{(t+1)}(s,a) - \boldsymbol{Q}^{(t)}(s,a)}_{\text{Q-tracking}}\Big)\,. \tag{26}$$

end for

At each iteration, each agent updates its policy estimates based on its neighbors' information via gossip mixing, in addition to a correction term that tracks the difference $Q_n^{\pi_n^{(t+1)}}(s,a) - Q_n^{\pi_n^{(t)}}(s,a)$ of the local Q-functions between consecutive policy updates. Note that the mixing is applied linearly to the logarithms of local policies, which translates into a multiplicative mixing of the local policies. Algorithm 1 summarizes the detailed procedure of the proposed algorithm written in a compact matrix form, which we dub as federated NPG (FedNPG). Note that the agents do not need to share their reward functions with others, and agent $n \in [N]$ will only be responsible to evaluate the local policy $\pi_n^{(t)}$ using the local reward r_n.

Entropy-regularized federated NPG methods. Moving onto the entropy regularized case, we adopt similar algorithmic ideas to decentralize (13), and propose the federated NPG (FedNPG) method with entropy regularization, summarized in Algorithm 2. Clearly, the entropy-regularized FedNPG method reduces to the vanilla FedNPG in the absence of the regularization (i.e., when $\tau = 0$).

Algorithm 2 Federated NPG (FedNPG) with entropy regularization

Input: learning rate $\eta > 0$, iteration number $T \in \mathbb{N}_+$, mixing matrix $\boldsymbol{W} \in \mathbb{R}^{N \times N}$, regularization coefficient $\tau > 0$.
Initialize: $\boldsymbol{\pi}^{(0)}$, $\boldsymbol{T}^{(0)} = \boldsymbol{Q}_\tau^{(0)}$.
for $t = 0, 1, \cdots$ **do**
  Update the policy for each $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\log \boldsymbol{\pi}^{(t+1)}(a|s) = \boldsymbol{W}\left(\left(1 - \frac{\eta\tau}{1-\gamma}\right)\log \boldsymbol{\pi}^{(t)}(a|s) + \frac{\eta}{1-\gamma}\boldsymbol{T}^{(t)}(s,a)\right) - \log \boldsymbol{z}^{(t)}(s)\,, \tag{27}$$

  where $\boldsymbol{z}^{(t)}(s) = \sum_{a' \in \mathcal{A}} \exp\left\{\boldsymbol{W}\left(\left(1 - \frac{\eta\tau}{1-\gamma}\right)\log \boldsymbol{\pi}^{(t)}(a'|s) + \frac{\eta}{1-\gamma}\boldsymbol{T}^{(t)}(s,a')\right)\right\}$.
  Evaluate $\boldsymbol{Q}_\tau^{(t+1)}$.
  Update the global Q-function estimate for each $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\boldsymbol{T}^{(t+1)}(s,a) = \boldsymbol{W}\Big(\boldsymbol{T}^{(t)}(s,a) + \underbrace{\boldsymbol{Q}_\tau^{(t+1)}(s,a) - \boldsymbol{Q}_\tau^{(t)}(s,a)}_{\text{Q-tracking}}\Big)\,. \tag{28}$$

end for

3.3 Global convergence of FedNPG

Convergence with exact policy evaluation. We begin with the global convergence of FedNPG (cf. Algorithm 1), stated in the following theorem. The formal statement and proof of this result can be found in Appendix A.3.

Theorem 1 (Global sublinear convergence of exact FedNPG (informal))**.** *Suppose* $\pi_n^{(0)}, n \in [N]$ *are set as the uniform distribution. Then for* $0 < \eta \le \eta_1 := \frac{(1-\sigma)^2(1-\gamma)^3}{16\sqrt{N}\sigma}$*, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(V^\star(\rho) - V^{\overline{\pi}^{(t)}}(\rho)\right) \le \frac{V^\star(d_\rho^{\pi^\star})}{(1-\gamma)T} + \frac{\log|\mathcal{A}|}{\eta T} + \frac{32N\sigma^2\eta^2}{(1-\gamma)^9(1-\sigma)^2}\,. \tag{29}$$

Furthermore, the consensus error satisfies

$$\forall n \in [N]: \quad \left\|\log \pi_n^{(t)} - \log \overline{\pi}^{(t)}\right\|_\infty \le \frac{32N\sigma}{3(1-\gamma)^4(1-\sigma)}\eta\,. \tag{30}$$

Theorem 1 characterizes the average-iterate convergence of the average policy $\overline{\pi}^{(t)}$ (cf. (24)) across the agents, which depends logarithmically on the size of the action space, and independently on the size of the state space. In addition, the consensus error of the local policies $\pi_n^{(t)}$ towards the average policy $\overline{\pi}^{(t)}$ is characterized in (30). When $T \ge \frac{128\sqrt{N}\log|\mathcal{A}|\sigma^4}{(1-\sigma)^4}$, by optimizing the learning rate $\eta = \left(\frac{(1-\gamma)^9(1-\sigma)^2\log|\mathcal{A}|}{32TN\sigma^2}\right)^{1/3}$ to balance the latter two terms, we arrive at

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(V^\star(\rho) - V^{\overline{\pi}^{(t)}}(\rho)\right) \lesssim \frac{V^\star(d_\rho^{\pi^\star})}{(1-\gamma)T} + \frac{N^{1/3}\sigma^{2/3}}{(1-\gamma)^3(1-\sigma)^{2/3}}\left(\frac{\log|\mathcal{A}|}{T}\right)^{2/3}. \tag{31a}$$

$$\left\|\log \pi_n^{(t)} - \log \overline{\pi}^{(t)}\right\|_\infty \lesssim \frac{N^{2/3}\sigma^{1/3}}{(1-\gamma)(1-\sigma)^{1/3}}\left(\frac{\log|\mathcal{A}|}{T}\right)^{1/3}. \tag{31b}$$

A few comments are in order.

- *Server-client setting.* When the network is fully connected, i.e., $\sigma = 0$, the convergence rate of FedNPG recovers the $\mathcal{O}(1/T)$ rate, matching that of the centralized NPG established in Agarwal et al. (2021).
- *Well-connected networks.* When the network is relatively well-connected in the sense of $\frac{\sigma^2}{(1-\sigma)^2} \lesssim \frac{1-\gamma}{N^{1/2}}$, FedNPG first converges at the rate of $\mathcal{O}(1/T)$, and then at the slower $\mathcal{O}(1/T^{2/3})$ rate after $T \gtrsim \frac{(1-\gamma)^3(1-\sigma)^2}{N\sigma^2}$.
- *Poorly-connected networks.* In addition, when the network is poorly connected in the sense of $\frac{\sigma^2}{(1-\sigma)^2} \gtrsim \frac{1-\gamma}{N^{1/2}}$, we see that FedNPG converges at the slower $\mathcal{O}(1/T^{2/3})$ rate.

We state the iteration complexity in Corollary 1.

Corollary 1 (Iteration complexity of exact FedNPG)**.** *To reach*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(V^\star(\rho) - V^{\overline{\pi}^{(t)}}(\rho)\right) \le \varepsilon,$$

the iteration complexity of FedNPG is at most $\mathcal{O}\left(\left(\frac{\sigma}{(1-\gamma)^{9/2}(1-\sigma)\varepsilon^{3/2}} + \frac{\sigma^2}{(1-\sigma)^4}\right)\sqrt{N}\log|\mathcal{A}| + \frac{1}{\varepsilon(1-\gamma)^2}\right)$.

Convergence with inexact policy evaluation. In practice, the policies need to be evaluated using samples collected by the agents, where the Q-functions are only estimated approximately. We are interested in gauging how the approximation error impacts the performance of FedNPG, as demonstrated in the following theorem.

Theorem 2 (Global sublinear convergence of inexact FedNPG (informal))**.** *Suppose that an estimate $q_n^{\pi_n^{(t)}}$ are used in replace of $Q_n^{\pi_n^{(t)}}$ in Algorithm 1. Under the assumptions of Theorem 1, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(V^\star(\rho)-V^{\overline{\pi}^{(t)}}(\rho)\right)\le\frac{V^\star(d_\rho^{\pi^\star})}{(1-\gamma)T}+\frac{\log|\mathcal{A}|}{\eta T}+\frac{32N\sigma^2\eta^2}{(1-\gamma)^9(1-\sigma)^2}+C_3\max_{n\in[N],t\in[T]}\left\|Q_n^{\pi_n^{(t)}}-q_n^{\pi_n^{(t)}}\right\|_\infty, \quad (32)$$

where $C_3 \coloneqq \frac{32\sqrt{N}\sigma\eta}{(1-\gamma)^5(1-\sigma)^2}\left(\frac{\eta\sqrt{N}}{(1-\gamma)^3}+1\right)+\frac{2}{(1-\gamma)^2}$.

The formal statement and proof of this result is given in Appendix A.4.

As long as $\max_{n\in[N],t\in[T]}\left\|Q_n^{\pi_n^{(t)}}-q_n^{\pi_n^{(t)}}\right\|_\infty\le\frac{\varepsilon}{C_3}$, inexact FedNPG reaches $\frac{1}{T}\sum_{t=0}^{T-1}\left(V^\star(\rho)-V^{\overline{\pi}^{(t)}}(\rho)\right)\le 2\varepsilon$ at the same iteration complexity as predicted in Corollary 1. Equipped with existing sample complexity bounds on policy evaluation, e.g. using a simulator as in Li et al. (2023b) and Li et al. (2023a), this immediate leads to a sample complexity bound for a federated actor-critic type algorithm for multi-task RL. We detail this in the following remark.

Remark 1 (sample complexity bound of inexact FedNPG)**.** *Recall that Li et al. (2023b) shows that for any fixed policy π, model-based policy evaluation achieves $\|q_\tau^\pi-Q_\tau^\pi\|_\infty\le\varepsilon_{\mathsf{eval}}$ with high probability if the number of samples per state-action pair exceeds the order of $\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^3\varepsilon_{\mathsf{eval}}^2}\right)$. When $T\gtrsim\frac{\sqrt{N}\log|\mathcal{A}|\sigma^4}{(1-\sigma)^4}$ and $\eta=\left(\frac{(1-\gamma)^9(1-\sigma)^2\log|\mathcal{A}|}{32TN\sigma^2}\right)^{1/3}$, we have $C_3\asymp 1/(1-\gamma)^2$. By employing fresh samples for the policy evaluation of each agent at every iteration, we can set $\varepsilon_{\mathsf{eval}}\coloneqq\max_{n\in[N],t\in[T]}\left\|Q_n^{\pi_n^{(t)}}-q_n^{\pi_n^{(t)}}\right\|_\infty\asymp\frac{\varepsilon}{C_3}\asymp(1-\gamma)^2\varepsilon$, and invoke the union bound over all iterations to give a loose upper bound of sample complexity of FedNPG per state-action pair at each agent as follows:*

$$\underbrace{\widetilde{\mathcal{O}}\left(\left(\frac{\sigma}{(1-\gamma)^{9/2}(1-\sigma)\varepsilon^{3/2}}+\frac{\sigma^2}{(1-\sigma)^4}\right)\sqrt{N}+\frac{1}{\varepsilon(1-\gamma)^2}\right)}_{\text{iteration complexity}}\cdot\underbrace{\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^7\varepsilon^2}\right)}_{\text{sample complexity per iteration}}$$
$$=\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^7\varepsilon^2}\cdot\left[\left(\frac{\sigma}{(1-\gamma)^{9/2}(1-\sigma)\varepsilon^{3/2}}+\frac{\sigma^2}{(1-\sigma)^4}\right)\sqrt{N}+\frac{1}{\varepsilon(1-\gamma)^2}\right]\right).$$

Hence, the total sample complexity scales linearly with respect to the size of the state-action space up to logarithmic factors. When σ is close to 1, which corresponds to the case where the network exhibits a high degree of locality, the above sample complexity becomes

$$\widetilde{\mathcal{O}}\left(\frac{\sqrt{N}}{(1-\gamma)^7\varepsilon^2}\cdot\left[\left(\frac{1}{(1-\gamma)^{9/2}(1-\sigma)\varepsilon^{3/2}}+\frac{1}{(1-\sigma)^4}\right)\right]\right),$$

which further simplifies to

$$\widetilde{\mathcal{O}}\left(\frac{\sqrt{N}}{(1-\gamma)^{11.5}(1-\sigma)\varepsilon^{3.5}}\right) \quad (33)$$

for sufficiently small ε.

3.4 Global convergence of FedNPG with entropy regularization

Convergence with exact policy evaluation. Next, we present our global convergence guarantee of entropy-regularized FedNPG with exact policy evaluation (cf. Algorithm 2).

Theorem 3 (Global linear convergence of exact entropy-regularized FedNPG (informal))**.** *For any $\gamma\in(0,1)$ and $0<\tau\le 1$, there exists $\eta_0=\min\left\{\frac{1-\gamma}{\tau},\mathcal{O}\left(\frac{(1-\gamma)^7(1-\sigma)^2\tau}{\sigma^2N}\right)\right\}$, such that if $0<\eta\le\eta_0$, then we have*

$$\left\|\overline{Q}_\tau^{(t)}-Q_\tau^\star\right\|_\infty\le 2\gamma C_1\rho(\eta)^t, \quad \left\|\log\pi_\tau^\star-\log\overline{\pi}^{(t)}\right\|_\infty\le\frac{2C_1}{\tau}\rho(\eta)^t, \quad (34)$$

where $\overline{Q}_\tau^{(t)} := Q_\tau^{\overline{\pi}^{(t)}}$, $\rho(\eta) \le \max\{1-\frac{\tau\eta}{2}, \frac{3+\sigma}{4}\} < 1$, *and* C_1 *is some problem-dependent constant. Furthermore, the consensus error satisfies*

$$\forall n \in [N]: \quad \left\| \log \pi_n^{(t)} - \log \overline{\pi}^{(t)} \right\|_\infty \le 2C_1 \rho(\eta)^t. \tag{35}$$

The exact expressions of C_1 and η_0 are specified in Appendix A.1. Theorem 3 confirms that entropy-regularized FedNPG converges at a linear rate to the optimal regularized policy, which is almost independent of the size of the state-action space, highlighting the positive role of entropy regularization in federated policy optimization. When the network is fully connected, i.e. $\sigma = 0$, the iteration complexity of entropy-regularized FedNPG reduces to $\mathcal{O}\left(\frac{1}{\eta\tau}\log\frac{1}{\varepsilon}\right)$, matching that of the centralized entropy-regularized NPG established in Cen et al. (2021). When the network is less connected, one needs to be more conservative in the choice of learning rates, leading to a higher iteration complexity, as described in the following corollary.

Corollary 2 (Iteration complexity of exact entropy-regularized FedNPG)**.** *To reach* $\left\|\log \pi_\tau^\star - \log \overline{\pi}^{(t)}\right\|_\infty \le \varepsilon$, *the iteration complexity of entropy-regularized FedNPG is at most*

$$\widetilde{\mathcal{O}}\left(\max\left\{\frac{2}{\tau\eta}, \frac{4}{1-\sigma}\right\}\log\frac{1}{\varepsilon}\right) \tag{36}$$

up to logarithmic factors. Especially, when $\eta = \eta_0$, *the best iteration complexity becomes*

$$\widetilde{\mathcal{O}}\left(\left(\frac{N\sigma^2}{(1-\gamma)^7(1-\sigma)^2\tau^2} + \frac{1}{1-\gamma}\right)\log\frac{1}{\tau\varepsilon}\right).$$

Convergence with inexact policy evaluation. Last but not the least, we present the informal convergence results of entropy-regularized FedNPG with inexact policy evaluation, whose formal version can be found in Appendix A.2.

Theorem 4 (Global linear convergence of inexact entropy-regularized FedNPG (informal))**.** *Suppose that an estimate* $q_{\tau,n}^{\pi_n^{(t)}}$ *are used in replace of* $Q_{\tau,n}^{\pi_n^{(t)}}$ *in Algorithm 2. Under the assumptions of Theorem 3, we have*

$$\begin{aligned} \left\|\overline{Q}_\tau^{(t)} - Q_\tau^\star\right\|_\infty &\le 2\gamma\Big(C_1\rho(\eta)^t + C_2 \max_{n\in[N],t\in[T]} \left\|Q_{\tau,n}^{\pi_n^{(t)}} - q_{\tau,n}^{\pi_n^{(t)}}\right\|_\infty\Big), \\ \left\|\log \pi_\tau^\star - \log \overline{\pi}^{(t)}\right\|_\infty &\le \frac{2}{\tau}\Big(C_1\rho(\eta)^t + C_2 \max_{n\in[N],t\in[T]} \left\|Q_{\tau,n}^{\pi_n^{(t)}} - q_{\tau,n}^{\pi_n^{(t)}}\right\|_\infty\Big), \end{aligned} \tag{37}$$

where $\overline{Q}_\tau^{(t)} := Q_\tau^{\overline{\pi}^{(t)}}$, $\rho(\eta) \le \max\{1-\frac{\tau\eta}{2}, \frac{3+\sigma}{4}\} < 1$, *and* C_1, C_2 *are problem-dependent constants.*

4 Federated NAC methods for multi-task RL

In this section, motivated by the design and analysis of FedNPG, we go beyond the tabular setting and exact policy evaluation, by proposing a federated natural actor-critic (FedNAC) method with function approximation and stochastic policy evaluation. Specifically, we consider the policy with function approximation under softmax parameterization of the following form:

$$f_\xi(a|s) = \frac{\exp(\phi^\top(s,a)\boldsymbol{\xi})}{\sum_{a'\in\mathcal{A}}\exp(\phi^\top(s,a')\boldsymbol{\xi})}, \tag{38}$$

for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $\boldsymbol{\xi} \in \mathbb{R}^p$, where $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^p$ is a known feature map. We assume ϕ is bounded over $\mathcal{S}\times\mathcal{A}$, i.e., there exists $C_\phi > 0$ such that

$$\|\phi(s,a)\|_2 \le C_\phi$$

holds for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Following Agarwal et al. (2021); Yuan et al. (2022), given any $\boldsymbol{w} \in \mathbb{R}^p$, $Q : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ and probability distribution $\zeta \in \Delta(\mathcal{S}\times\mathcal{A})$ over the state-action space, we define the *function approximation error* $\ell(\boldsymbol{w}, Q, \zeta)$ as follows:

$$\ell(\boldsymbol{w}, Q, \zeta) := \mathbb{E}_{(s,a)\sim\zeta}\left[\left(\boldsymbol{w}^\top\phi(s,a) - Q(s,a)\right)^2\right]. \tag{39}$$

Algorithm 3 Federated Natural Actor-Critic (FedNAC)

Input: number of actor iterations T, number of critic iterations K, actor learning rate α, critic learning rate β, discounted factor $\gamma \in [0, 1)$

Initialization: initial state-action distribution ν, actor parameter $\boldsymbol{\xi}^{(0)} = (\boldsymbol{\xi}_1^{(0)\top}, \cdots, \boldsymbol{\xi}_N^{(0)\top})^\top \in \mathbb{R}^{N\times p}$, $\boldsymbol{h}^{(-1)} = \boldsymbol{w}^{(-1)} = \mathbf{0} \in \mathbb{R}^{N\times p}$

for $t = 0, \cdots, T-1$ **do**

Critic update: $\boldsymbol{w}_n^{(t)} =$ Critic(K, ν, $\boldsymbol{\xi}_n^{(t)}$, γ, β, r_n), $n \in [N]$ (Algorithm 4)

Update the critic parameter for estimating the global Q-function:

$$\boldsymbol{h}^{(t)} = \boldsymbol{W}\left(\boldsymbol{h}^{(t-1)} + \boldsymbol{w}^{(t)} - \boldsymbol{w}^{(t-1)}\right) \tag{40}$$

Actor update:

$$\boldsymbol{\xi}^{(t+1)} = \boldsymbol{W}\left(\boldsymbol{\xi}^{(t)} + \alpha \boldsymbol{h}^{(t)}\right) \tag{41}$$

end for

By searching for $\boldsymbol{w}$ that minimizes $\ell(\boldsymbol{w}, Q, \zeta)$, it approximates $Q(s, a)$ using the feature map ϕ with respect to the distribution ζ.

4.1 Algorithm design

Our proposed federated NAC method FedNAC could be seen as a decentralized version of Q-NPG method (Agarwal et al., 2021; Yuan et al., 2022), which we briefly review as follows.

Q-NPG method. Q-NPG is a sample version of NPG with function approximation which is suitable for the case where $\mathcal{S}$ or $\mathcal{A}$ is large or infinite. We consider the policy with function approximation under softmax parameterization (38).

Given an approximate solution $\boldsymbol{w}^{(t)}$ for minimizing the function approximation error $\ell(\boldsymbol{w}, Q^{f_{\xi^{(t)}}}, \tilde{d}_\nu^{f_{\xi^{(t)}}})$ (see (39)), the Q-NPG update rule $\boldsymbol{\xi}^{(t+1)} = \boldsymbol{\xi}^{(t)} + \alpha \boldsymbol{w}^{(t)}$, when plugged in parameterization (38), results in the following policy update rule when we set $\alpha = \eta/(1-\gamma)$:

$$f^{(t+1)}(a|s) \propto f^{(t)}(a|s) \exp\left(\frac{\eta \phi^\top(s,a)\boldsymbol{w}^{(t)}}{1-\gamma}\right), \tag{42}$$

which could be seen as the function approximation version of the update rule (12) of vanilla NPG method.

Federated NAC method. Let us now discuss the high-level design of FedNAC, which is presented in Algorithm 3. At the t-th iteration ($t = 0, \ldots, T-1$), denote the actor (concerning the policies) parameters of all agents as $\boldsymbol{\xi}^{(t)} = (\boldsymbol{\xi}_1^{(t)}, \ldots, \boldsymbol{\xi}_N^{(t)})^\top \in \mathbb{R}^{N\times p}$, and the critic parameters of all agents as $\boldsymbol{w}^{(t)} = (\boldsymbol{w}_1^{(t)}, \ldots, \boldsymbol{w}_N^{(t)})^\top \in \mathbb{R}^{N\times p}$ (concerning the local Q-values) and $\boldsymbol{h}^{(t)} = (\boldsymbol{h}_1^{(t)}, \ldots, \boldsymbol{h}_N^{(t)})^\top \in \mathbb{R}^{N\times p}$ (concerning the global Q-values).

- First, the critic parameter $\boldsymbol{w}_n^{(t)}$ is locally updated at each agent by aiming to minimize $\ell(\boldsymbol{w}, Q_n^{(t)}, \tilde{d}_n^{(t)})$ (cf. (39)) with gradient descent, where $Q_n^{(t)}$ is the local Q-function of the local policy $f_{\xi_n^{(t)}}$, and $\tilde{d}_n^{(t)}$ is the state-action visitation distribution induced by the local policy $f_{\xi_n^{(t)}}$ and an initial state-action distribution ν (determined from the data sampling mechanism, cf. (9)). However, since $Q_n^{(t)}$ is not directly available, it needs to be estimated from samples. Therefore, the critic update takes K steps of stochastic gradient descent with critic learning rate β, given by

$$\widetilde{\boldsymbol{w}}_{k+1} = \widetilde{\boldsymbol{w}}_k - \beta\big(\widetilde{\boldsymbol{w}}_k^\top \phi(s_k, a_k) - \widehat{Q}_\xi(s_k, a_k)\big)\phi(s_k, a_k),$$

  for $k = 0, \ldots, K-1$, where (s_k, a_k) is sampled on the local policy $f_{\xi_n^{(t)}}$, and $\widehat{Q}_\xi(s_k, a_k)$ is a careful estimate of the Q-value using a trajectory with expected length $1/(1-\gamma)$ (see Algorithm 5 adopted from

Yuan et al. (2022, Lemma 4)), and $\widetilde{\boldsymbol{w}}_0 = \mathbf{0}$ for simplicity. As a consequence, in line 4 of Algorithm 4, we have

$$\mathbb{E}\left[\widehat{\nabla}_w \ell(\widetilde{\boldsymbol{w}}_k, \widehat{Q}^\pi, \tilde{d}^{f_\xi})\right] = \nabla_w \ell(\widetilde{\boldsymbol{w}}_k, \widehat{Q}^\pi, \tilde{d}^{f_\xi}) \,. \tag{43}$$

The final critic is updated as $\boldsymbol{w}_n^{(t)} = \frac{1}{K}\sum_{k=1}^K \widetilde{\boldsymbol{w}}_k$. The total sample complexity of the critic update per iteration is then on the order of $K/(1-\gamma)$.

- Next, the critic parameter $\boldsymbol{h}_n^{(t)}$ for estimating the global Q-function can then be estimated by averaging with the neighbors with the Q-tracking term, given by $\boldsymbol{h}^{(t)} = \boldsymbol{W}\left(\boldsymbol{h}^{(t-1)} + \boldsymbol{w}^{(t)} - \boldsymbol{w}^{(t-1)}\right)$.
- Finally, the actor parameter $\boldsymbol{\xi}_n^{(t)}$ can be updated via averaging with the neighbors along with the policy gradient informed by $\boldsymbol{h}_n^{(t)}$, given by $\boldsymbol{\xi}^{(t+1)} = \boldsymbol{W}\left(\boldsymbol{\xi}^{(t)} + \alpha \boldsymbol{h}^{(t)}\right)$, where α is the learning rate of the actor.

Note that the sample complexity of FedNAC is on the order of $KT/(1-\gamma)$. An important aspect of the FedNAC method is that the policy is updated using trajectory data collected via executing the learned policy, which is closer to practice and more challenging to learn than using the generative model.

Algorithm 4 Critic($K, \nu, \boldsymbol{\xi}, \gamma, \beta, r$): sample-based regression solver to minimize $\ell(\boldsymbol{w}, Q_n^{(t)}, \tilde{d}_n^{(t)})$

1: **Initialize:** critic parameter $w_0 \in \mathbb{R}^p$
2: **for** $k = 0, \cdots, K-1$ **do**
3: Sampling: $(s_k, a_k), \widehat{Q}^\pi(s_k, a_k)$ =Q-Sampler(ν, f_ξ, γ, r) (Algorithm 5)
4: Compute the stochastic gradient estimator of L_Q:

$$\widehat{\nabla}_w \ell(\widetilde{\boldsymbol{w}}_k, \widehat{Q}^\pi, \tilde{d}^{f_\xi}) = 2\left(\widetilde{\boldsymbol{w}}_k^\top \phi(s_k, a_k) - \widehat{Q}^\pi(s_k, a_k)\right)\phi(s_k, a_k) \tag{44}$$

5: Critic Update: $\widetilde{\boldsymbol{w}}_{k+1} = \widetilde{\boldsymbol{w}}_k - \beta \widehat{\nabla}_w \ell(\widetilde{\boldsymbol{w}}_k, \widehat{Q}^\pi, \tilde{d}^{f_\xi})$
6: **end for**
7: **Output:** $\boldsymbol{w}_{\text{out}} = \frac{1}{K}\sum_{k=1}^K \widetilde{\boldsymbol{w}}_k$

Algorithm 5 Q-Sampler(ν, π, γ, r)

1: **Initialize:** $(s_0, a_0) \sim \nu$, time step $h, t = 0$, variable $X \sim$ Bernoulli(γ)
2: **while** $X = 1$ **do**
3: Sample $s_{h+1} \sim P(\cdot|s_h, a_h)$
4: Sample $a_{h+1} \sim \pi(\cdot|s_{h+1})$
5: $h \leftarrow h+1$
6: $X \sim$ Bernoulli(γ)
7: **end while**
8: Set $\widehat{Q}^\pi(s_h, a_h) = r(s_h, a_h)$, $X \sim$ Bernoulli(γ), $t = h$
9: **while** $X = 1$ **do**
10: Sample $s_{t+1} \sim P(\cdot|s_t, a_t)$
11: Sample $a_{t+1} \sim \pi(\cdot|s_{t+1})$
12: $\widehat{Q}^\pi(s_h, a_h) \leftarrow \widehat{Q}^\pi(s_h, a_h) + r(s_{t+1}, a_{t+1})$
13: $t \leftarrow t+1$
14: $X \sim$ Bernoulli(γ)
15: **end while**
16: **Output:** (s_h, a_h) and $\widehat{Q}^\pi(s_h, a_h)$

4.2 Theoretical guarantees

We first state the assumptions that are needed to guarantee the convergence of Algorithm 3, which are all commonly used in the literature, e.g., Yuan et al. (2022); Agarwal et al. (2021). To begin, we require the

covariance matrix of the feature map induced by the initial state-action distribution ν satisfies the following assumption to guarantee the convergence of the critic.

Assumption 2 (PSD of the covariance matrix of the feature map)**.** *There exists $\mu > 0$ such that*

$$\mathbb{E}_{(s,a)\sim\nu}\left[\phi(s,a)\phi^\top(s,a)\right] = \Sigma_\nu \geq \mu \boldsymbol{I}. \tag{45}$$

We also need to ensure that the Q-values can be well approximated by the linear function approximation using feature map $\phi(s,a)$, which is captured next.

Assumption 3 (Bounded approximation error)**.** *For each $n \in [N]$, there exists $\varepsilon^n_{approx} \geq 0$ such that for all $t \in \mathbb{N}$, it holds that $\mathbb{E}\left[\ell\left(\boldsymbol{w}^{(t)}_{\star,n}, Q^{(t)}_n, \tilde{d}^{(t)}_n\right)\right] \leq \varepsilon^n_{approx}$, where $\boldsymbol{w}^{(t)}_{\star,n} := \arg\min_{\boldsymbol{w}} \ell\left(\boldsymbol{w}^{(t)}_{\star,n}, Q^{(t)}_n, \tilde{d}^{(t)}_n\right)$.*

We denote the average approximation error as $\bar{\varepsilon}_{\text{approx}} = \frac{1}{N}\sum_{n=1}^N \varepsilon^n_{\text{approx}}$. Similar as Yuan et al. (2022), we need the following assumption that bounds the transfer errors due to distribution shifts.

Assumption 4 (Bounded transfer error)**.** *There exists $C_\nu > 0$ such that for all $n \in [N]$ and $t \in \mathbb{N}$, it holds that $\mathbb{E}_{(s,a)\sim\tilde{d}^{(t)}_n}\left[\left(\frac{h^\pi(s,a)}{\tilde{d}^{(t)}_n(s,a)}\right)^2\right] \leq C_\nu$, where $h^\pi(s,a)$ is the state-action visitation distribution induced by any policy π from initial state distribution ρ.*

Note that if we choose $\nu(s,a) > 0$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, then Assumption 4 is guaranteed to hold true (see Lemma 7 in Appendix B). We are now ready to state the convergence guarantee, whose formal version and proof could be found in Appendix B.

Theorem 5 (Convergence rate of Algorithm 3 (informal))**.** *Let $\boldsymbol{\xi}^{(0)}_1 = \cdots = \boldsymbol{\xi}^{(0)}_N$ in FedNAC. Denoting $\bar{\boldsymbol{\xi}}^{(t)} := \frac{1}{N}\sum_{n=1}^N \boldsymbol{\xi}^{(t)}_n$, and $\bar{f}^{(t)} := f_{\bar{\xi}^{(t)}}$ as the average policy. Then under Assumption 1, 2, 3 and 4, with appropriately chosen learning rates α and β, as long as the number of actor iterations satisfies*

$$T \gtrsim \max\left\{\frac{\sigma}{\varepsilon^{3/2}(1-\gamma)^{17/4}(1-\sigma)^{3/2}}, \frac{1}{\varepsilon(1-\gamma)}, \frac{\sigma^{1/4}}{\varepsilon^{3/4}(1-\sigma)^{3/8}(1-\gamma)^{7/8}N^{3/8}}, \frac{\sigma^4}{(1-\gamma)^2(1-\sigma)^6}\right\}$$

and the number of critic iterations satisfies $K = \mathcal{O}\left(\frac{1}{(1-\gamma)^6\varepsilon^2}\right)$, it holds that

$$V^\star(\rho) - \frac{1}{T}\sum_{t=0}^{T-1} V^{\bar{f}^{(t)}}(\rho) \lesssim \varepsilon + \frac{\bar{\varepsilon}_{approx}}{1-\gamma}. \tag{46}$$

In the server-client setting when $\sigma = 0$, to reach (46), it suffices to choose $T = \mathcal{O}\left(\frac{1}{(1-\gamma)\varepsilon}\right)$ and $K = \mathcal{O}\left(\frac{1}{(1-\gamma)^6\varepsilon^2}\right)$, leading to a total sample complexity of $KT/(1-\gamma) = \mathcal{O}\left(\frac{1}{(1-\gamma)^8\varepsilon^3}\right)$ per agent, and $T = \mathcal{O}\left(\frac{1}{(1-\gamma)\varepsilon}\right)$ rounds of communication. The sample complexity matches that of (centralized) Q-NPG established in Yuan et al. (2022) with a single agent. On the other end, in the fully decentralized setting when σ is not close to 0, FedNAC requires $\mathcal{O}\left(\frac{1}{(1-\gamma)^{45/4}\varepsilon^{7/2}(1-\sigma)^{3/2}}\right)$ samples for each agent and $\mathcal{O}\left(\frac{1}{\varepsilon^{3/2}(1-\gamma)^{17/4}(1-\sigma)^{3/2}}\right)$ rounds of communication to reach (46), for sufficiently small ε. Encouragingly, the dependency on the accuracy level ε — the dominating factor — in the sample complexity matches that of FedNPG given in (33) when assuming access to the generative model, which allows query of arbitrary state-action pairs. In contrast, FedNAC only collects on-policy samples, and therefore is much more challenging to guarantee its convergence.

5 Conclusions

This work proposes the first provably efficient federated NPG (FedNPG) methods for solving vanilla and entropy-regularized multi-task RL problems in the fully decentralized setting. The established finite-time global convergence guarantees are almost independent of the size of the state-action space up to some logarithmic factor, and illuminate the impacts of the size and connectivity of the network. Furthermore, the

proposed FedNPG methods are provably robust vis-a-vis inexactness of local policy evaluations. Last but not least, we also propose FedNAC, which can be viewed as an extension of FedNPG with function approximation and stochastic policy evaluation, and establish its finite-time sample complexity. Future directions include generalizing the framework of federated policy optimization to allow personalized policy learning in a shared environment.

Acknowledgments

The work of T. Yang, S. Cen and Y. Chi are supported in part by the grants ONR N00014-19-1-2404, NSF CCF-1901199, CCF-2106778, AFRL FA8750-20-2-0504, and a CMU Cylab seed grant. The work of Y. Wei is supported in part by the the NSF grants DMS-2147546/2015447, CAREER award DMS-2143215, CCF-2106778, and the Google Research Scholar Award. The work of Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR grant FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009 and CCF-1907661. S. Cen is also gratefully supported by Wei Shen and Xuehong Zhang Presidential Fellowship, Boeing Scholarship, and JP Morgan Chase PhD Fellowship.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506.

Ahmed, Z., Le Roux, N., Norouzi, M., and Schuurmans, D. (2019). Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, pages 151–160.

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.

Anwar, A. and Raychowdhury, A. (2021). Multi-task federated reinforcement learning with adversaries. *arXiv preprint arXiv:2103.06473*.

Assran, M., Romoff, J., Ballas, N., Pineau, J., and Rabbat, M. (2019). Gossip-based actor-learner architectures for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 32.

Bach, F. and Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). *Advances in neural information processing systems*, 26.

Bhandari, J. and Russo, D. (2021). On the linear convergence of policy gradient methods for finite MDPs. In *International Conference on Artificial Intelligence and Statistics*, pages 2386–2394. PMLR.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. (2009). Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2022a). Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578.

Cen, S., Chi, Y., Du, S. S., and Xiao, L. (2022b). Faster last-iterate convergence of policy optimization in zero-sum Markov games. In *The Eleventh International Conference on Learning Representations*.

Cen, S., Wei, Y., and Chi, Y. (2021). Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34:27952–27964.

Chen, J., Feng, J., Gao, W., and Wei, K. (2022). Decentralized natural policy gradient with variance reduction for collaborative multi-agent reinforcement learning. *arXiv preprint arXiv:2209.02179*.

Chen, T., Zhang, K., Giannakis, G. B., and Başar, T. (2021a). Communication-efficient policy gradient methods for distributed reinforcement learning. *IEEE Transactions on Control of Network Systems*, 9(2):917–929.

Chen, Z., Zhou, Y., and Chen, R. (2021b). Multi-agent off-policy TDC with near-optimal sample and communication complexity. In *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pages 504–508. IEEE.

Di Lorenzo, P. and Scutari, G. (2016). Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136.

Duchi, J. C., Agarwal, A., and Wainwright, M. J. (2011). Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. (2018). Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*, pages 1407–1416. PMLR.

Eysenbach, B. and Levine, S. (2021). Maximum entropy RL (provably) solves some robust RL problems. In *International Conference on Learning Representations*.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Kakade, S. M. (2001). A natural policy gradient. *Advances in neural information processing systems*, 14.

Kar, S., Moura, J. M., and Poor, H. V. (2012). Qd-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus. *arXiv preprint arXiv:1205.0047*.

Khodadadian, S., Doan, T. T., Romberg, J., and Maguluri, S. T. (2022a). Finite sample analysis of two-time-scale natural actor-critic algorithm. *IEEE Transactions on Automatic Control*.

Khodadadian, S., Jhunjhunwala, P. R., Varma, S. M., and Maguluri, S. T. (2021). On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3794–3799. IEEE.

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

Lan, G. (2023). Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106.

Lan, G., Li, Y., and Zhao, T. (2023a). Block policy mirror descent. *SIAM Journal on Optimization*, 33(3):2341–2378.

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

Li, B., Cen, S., Chen, Y., and Chi, Y. (2020). Communication-efficient distributed optimization in networks with gradient tracking and variance reduction. *The Journal of Machine Learning Research*, 21(1):7331–7381.

Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2023a). Is q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*.

Li, G., Wei, Y., Chi, Y., and Chen, Y. (2023b). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Operations Research*.

Li, G., Wei, Y., Chi, Y., and Chen, Y. (2023c). Softmax policy gradient methods can take exponential time to converge. *Mathematical Programming*, pages 1–96.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. (2017). Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30.

Lobel, I. and Ozdaglar, A. (2008). Convergence analysis of distributed subgradient methods over random networks. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 353–360. IEEE.

M Alshater, M. (2022). Exploring the role of artificial intelligence in enhancing academic performance: A case study of chatgpt. *Available at SSRN*.

McKelvey, R. D. and Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020). On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR.

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

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. (2017). Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2775–2785.

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

Nedic, A., Olshevsky, A., and Shi, W. (2017). Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633.

Nedic, A. and Ozdaglar, A. (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61.

Omidshafiei, S., Pazis, J., Amato, C., How, J. P., and Vian, J. (2017). Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pages 2681–2690. PMLR.

Petersen, K. B. and Pedersen, M. S. (2008). The matrix cookbook. *Technical University of Denmark*, 7(15):510.

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

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

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

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

Rahman, M. M., Terano, H. J., Rahman, M. N., Salamzadeh, A., and Rahaman, M. S. (2023). Chatgpt and academic research: a review and recommendations based on practical examples. *Rahman, M., Terano, HJR, Rahman, N., Salamzadeh, A., Rahaman, S.(2023). ChatGPT and Academic Research: A Review and Recommendations Based on Practical Examples. Journal of Education, Management and Development Studies*, 3(1):1–12.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shani, L., Efroni, Y., and Mannor, S. (2020). Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5668–5675.

Wang, H., Kaplan, Z., Niu, D., and Li, B. (2020). Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 1698–1707. IEEE.

Wang, J., Hu, J., Mills, J., Min, G., Xia, M., and Georgalas, N. (2023). Federated ensemble model-based reinforcement learning in edge computing. *IEEE Transactions on Parallel and Distributed Systems*.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. (2019). Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*.

Williams, R. J. and Peng, J. (1991). Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268.

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

Woo, J., Shi, L., Joshi, G., and Chi, Y. (2024). Federated offline reinforcement learning: Collaborative single-policy coverage suffices. In *Forty-first International Conference on Machine Learning*.

Xiao, L. (2022). On the convergence rates of policy gradient methods. *The Journal of Machine Learning Research*, 23(1):12887–12922.

Xu, T., Wang, Z., and Liang, Y. (2020). Improving sample complexity bounds for actor-critic algorithms. *arXiv preprint arXiv:2004.12956*.

Yu, T., Li, T., Sun, Y., Nanda, S., Smith, V., Sekar, V., and Seshan, S. (2020). Learning context-aware policies from multiple smart homes via federated multi-task learning. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 104–115. IEEE.

Yuan, R., Du, S. S., Gower, R. M., Lazaric, A., and Xiao, L. (2022). Linear convergence of natural policy gradient methods with log-linear policies. *arXiv preprint arXiv:2210.01400*.

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

Zerka, F., Barakat, S., Walsh, S., Bogowicz, M., Leijenaar, R. T., Jochems, A., Miraglio, B., Townend, D., and Lambin, P. (2020). Systematic review of privacy-preserving distributed machine learning from federated databases in health care. *JCO clinical cancer informatics*, 4:184–200.

Zhan, W., Cen, S., Huang, B., Chen, Y., Lee, J. D., and Chi, Y. (2023). Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091.

Zhao, F., Ren, X., Yang, S., Zhao, P., Zhang, R., and Xu, X. (2023). Federated multi-objective reinforcement learning. *Information Sciences*, 624:811–832.

Zhou, R., Liu, T., Kalathil, D., Kumar, P., and Tian, C. (2022). Anchor-changing regularized natural policy gradient for multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 35:13584–13596.

Zhu, M. and Martínez, S. (2010). Discrete-time dynamic average consensus. *Automatica*, 46(2):322–329.

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

A Convergence analysis of FedNPG

For technical convenience, we present first the analysis for entropy-regularized FedNPG and then for vanilla FedNPG.

A.1 Analysis of entropy-regularized FedNPG with exact policy evaluation

To facilitate analysis, we introduce several notation below. For all $t \geq 0$, we recall $\overline{\pi}^{(t)}$ as the normalized geometric mean of $\{\pi_n^{(t)}\}_{n\in[N]}$:

$$\overline{\pi}^{(t)} \coloneqq \mathrm{softmax}\left(\frac{1}{N}\sum_{n=1}^{N}\log \pi_n^{(t)}\right), \tag{47}$$

from which we can easily see that for each $(s,a) \in \mathcal{S}\times\mathcal{A}$, $\overline{\pi}^{(t)}(a|s) \propto \left(\prod_{n=1}^{N}\pi_n^{(t)}(a|s)\right)^{\frac{1}{N}}$. We denote the soft Q-functions of $\overline{\pi}^{(t)}$ by $\overline{\boldsymbol{Q}}_\tau^{(t)}$:

$$\overline{\boldsymbol{Q}}_\tau^{(t)} \coloneqq \begin{pmatrix} Q_{\tau,1}^{\overline{\pi}^{(t)}} \\ \vdots \\ Q_{\tau,N}^{\overline{\pi}^{(t)}} \end{pmatrix}. \tag{48}$$

In addition, we define $\widehat{Q}_\tau^{(t)}$, $\overline{Q}_\tau^{(t)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\overline{V}_\tau^{(t)} \in \mathbb{R}^{|\mathcal{S}|}$ as follows

$$\widehat{Q}_\tau^{(t)} \coloneqq \frac{1}{N}\sum_{n=1}^{N} Q_{\tau,n}^{\pi_n^{(t)}}, \tag{49a}$$

$$\overline{Q}_\tau^{(t)} \coloneqq Q_\tau^{\overline{\pi}^{(t)}} = \frac{1}{N}\sum_{n=1}^{N} Q_{\tau,n}^{\overline{\pi}^{(t)}}. \tag{49b}$$

$$\overline{V}_\tau^{(t)} \coloneqq V_\tau^{\overline{\pi}^{(t)}} = \frac{1}{N}\sum_{n=1}^{N} V_{\tau,n}^{\overline{\pi}^{(t)}}. \tag{49c}$$

For notational convenience, we also denote

$$\alpha \coloneqq 1 - \frac{\eta\tau}{1-\gamma}. \tag{50}$$

Following Cen et al. (2022a), we introduce the following auxiliary sequence $\{\boldsymbol{\xi}^{(t)} = (\xi_1^{(t)},\cdots,\xi_N^{(t)})^\top \in \mathbb{R}^{N\times|\mathcal{S}||\mathcal{A}|}\}_{t=0,1,\cdots}$, each recursively defined as

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \quad \boldsymbol{\xi}^{(0)}(s,a) \coloneqq \frac{\left\|\exp\left(Q_\tau^\star(s,\cdot)/\tau\right)\right\|_1}{\left\|\exp\left(\frac{1}{N}\sum_{n=1}^{N}\log\pi_n^{(0)}(\cdot|s)\right)\right\|_1}\cdot\boldsymbol{\pi}^{(0)}(a|s), \tag{51a}$$

$$\log\boldsymbol{\xi}^{(t+1)}(s,a) = \boldsymbol{W}\left(\alpha\log\boldsymbol{\xi}^{(t)}(s,a) + (1-\alpha)\boldsymbol{T}^{(t)}(s,a)/\tau\right), \tag{51b}$$

where $\boldsymbol{T}^{(t)}(s,a)$ is updated via (26). Similarly, we introduce an averaged auxiliary sequence $\{\overline{\xi}^{(t)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}\}$ given by

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \quad \overline{\xi}^{(0)}(s,a) \coloneqq \left\|\exp\left(Q_\tau^\star(s,\cdot)/\tau\right)\right\|_1\cdot\overline{\pi}^{(0)}(a|s), \tag{52a}$$

$$\log\overline{\xi}^{(t+1)}(s,a) = \alpha\log\overline{\xi}^{(t)}(s,a) + (1-\alpha)\widehat{Q}_\tau^{(t)}(s,a)/\tau. \tag{52b}$$

We introduces four error metrics defined as

$$\Omega_1^{(t)} \coloneqq \left\|u^{(t)}\right\|_\infty, \tag{53a}$$

$$\Omega_2^{(t)} \coloneqq \left\|v^{(t)}\right\|_\infty, \tag{53b}$$

$$\Omega_3^{(t)} := \left\| Q_\tau^\star - \tau \log \overline{\xi}^{(t)} \right\|_\infty , \tag{53c}$$

$$\Omega_4^{(t)} := \max \left\{ 0, -\min_{s,a} \left(\overline{Q}_\tau^{(t)}(s,a) - \tau \log \overline{\xi}^{(t)}(s,a) \right) \right\} , \tag{53d}$$

where $u^{(t)}, v^{(t)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ are defined as

$$u^{(t)}(s,a) := \left\| \log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\xi}^{(t)}(s,a) \mathbf{1}_N \right\|_2 , \tag{54}$$

$$v^{(t)}(s,a) := \left\| \boldsymbol{T}^{(t)}(s,a) - \widehat{Q}_\tau^{(t)}(s,a) \mathbf{1}_N \right\|_2 . \tag{55}$$

We collect the error metrics above in a vector $\boldsymbol{\Omega}^{(t)} \in \mathbb{R}^4$:

$$\boldsymbol{\Omega}^{(t)} := \left(\Omega_1^{(t)}, \Omega_2^{(t)}, \Omega_3^{(t)}, \Omega_4^{(t)} \right)^\top . \tag{56}$$

With the above preparation, we are ready to state the convergence guarantee of Algorithm 2 in Theorem 6 below, which is the formal version of Theorem 3.

Theorem 6. *For any $N \in \mathbb{N}_+, \tau > 0, \gamma \in (0,1)$, there exists $\eta_0 > 0$ which depends only on $N, \gamma, \tau, \sigma, |\mathcal{A}|$, such that if $0 < \eta \le \eta_0$ and $1 - \sigma > 0$, then the updates of Algorithm 2 satisfy*

$$\left\| \overline{Q}_\tau^{(t)} - Q_\tau^\star \right\|_\infty \le 2\gamma \rho(\eta)^t \left\| \boldsymbol{\Omega}^{(0)} \right\|_2 , \tag{57}$$

$$\left\| \log \pi_\tau^\star - \log \overline{\pi}^{(t)} \right\|_\infty \le \frac{2}{\tau} \rho(\eta)^t \left\| \boldsymbol{\Omega}^{(0)} \right\|_2 , \tag{58}$$

where

$$\rho(\eta) \le \max \left\{ 1 - \frac{\tau\eta}{2}, \frac{3+\sigma}{4} \right\} < 1 .$$

Moreover, the consensus errors satisfy:

$$\forall n \in [N] : \quad \left\| \log \pi_n^{(t)} - \log \overline{\pi}^{(t)} \right\|_\infty \le 2\rho(\eta)^t \left\| \boldsymbol{\Omega}^{(0)} \right\|_2 . \tag{59}$$

The dependency of η_0 on $N, \gamma, \tau, \sigma, |\mathcal{A}|$ is made clear in Lemma 2 that will be presented momentarily in this section. The rest of this section is dedicated to the proof of Theorem 6. We first state a key lemma that tracks the error recursion of Algorithm 2.

Lemma 1. *The following linear system holds for all $t \ge 0$:*

$$\boldsymbol{\Omega}^{(t+1)} \le \underbrace{\begin{pmatrix} \sigma\alpha & \frac{\eta\sigma}{1-\gamma} & 0 & 0 \\ S\sigma & \left(1 + \frac{\eta M\sqrt{N}}{1-\gamma}\sigma\right)\sigma & \frac{(2+\gamma)\eta M N}{1-\gamma}\sigma & \frac{\gamma\eta M N}{1-\gamma}\sigma \\ (1-\alpha)M & 0 & (1-\alpha)\gamma + \alpha & (1-\alpha)\gamma \\ \frac{2\gamma+\eta\tau}{1-\gamma}M & 0 & 0 & \alpha \end{pmatrix}}_{=:\boldsymbol{A}(\eta)} \boldsymbol{\Omega}^{(t)} , \tag{60}$$

where we let

$$S := M\sqrt{N} \left(2\alpha + (1-\alpha) \cdot \sqrt{2N} + \frac{1-\alpha}{\tau} \cdot \sqrt{N} M \right) , \tag{61}$$

and

$$M := \frac{1 + \gamma + 2\tau(1-\gamma)\log|\mathcal{A}|}{(1-\gamma)^2} \cdot \gamma .$$

In addition, it holds for all $t \ge 0$ that

$$\left\| \overline{Q}_\tau^{(t)} - Q_\tau^\star \right\|_\infty \le \gamma \Omega_3^{(t)} + \gamma \Omega_4^{(t)} , \tag{62}$$

$$\left\| \log \overline{\pi}^{(t)} - \log \pi_\tau^\star \right\|_\infty \le \frac{2}{\tau} \Omega_3^{(t)} . \tag{63}$$

Proof. See Appendix C.1. □

Let $\rho(\eta)$ denote the spectral norm of $\boldsymbol{A}(\eta)$. As $\boldsymbol{\Omega}^{(t)} \geq 0$, it is immediate from (60) that

$$\left\|\boldsymbol{\Omega}^{(t)}\right\|_2 \leq \rho(\eta)^t \left\|\boldsymbol{\Omega}^{(0)}\right\|_2 ,$$

and therefore we have

$$\left\|\overline{Q}_\tau^{(t)} - Q_\tau^\star\right\|_\infty \leq 2\gamma \left\|\boldsymbol{\Omega}^{(t)}\right\|_\infty \leq 2\gamma\rho(\eta)^t \left\|\boldsymbol{\Omega}^{(0)}\right\|_2 ,$$

and

$$\left\|\log \overline{\pi}^{(t)} - \log \pi_\tau^\star\right\|_\infty \leq \frac{2}{\tau}\left\|\boldsymbol{\Omega}^{(t)}\right\|_\infty \leq \frac{2}{\tau}\rho(\eta)^t \left\|\boldsymbol{\Omega}^{(0)}\right\|_2 .$$

It remains to bound the spectral radius $\rho(\eta)$, which is achieved by the following lemma.

Lemma 2 (Bounding the spectral norm of $\boldsymbol{A}(\eta)$)**.** *Let*

$$\zeta := \frac{(1-\gamma)(1-\sigma)^2\tau}{8\left(\tau S_0 \sigma^2 + 10Mc\sigma^2/(1-\gamma) + (1-\sigma)^2\tau^2/16\right)} , \tag{64}$$

where $S_0 := M\sqrt{N}\left(2 + \sqrt{2N} + \frac{M\sqrt{N}}{\tau}\right)$, $c := MN/(1-\gamma)$. *For any* $N \in \mathbb{N}_+, \tau > 0, \gamma \in (0,1)$, *if*

$$0 < \eta \leq \eta_0 := \min\left\{\frac{1-\gamma}{\tau}, \zeta\right\} , \tag{65}$$

then we have

$$\rho(\eta) \leq \max\left\{\frac{3+\sigma}{4}, \frac{1+(1-\alpha)\gamma+\alpha}{2}\right\} < 1 . \tag{66}$$

Proof. See Appendix C.2. □

A.2 Analysis of entropy-regularized FedNPG with inexact policy evaluation

We define the collection of *inexact* Q-function estimates as

$$\boldsymbol{q}_\tau^{(t)} := \left(q_{\tau,1}^{\pi_1^{(t)}}, \cdots, q_{\tau,N}^{\pi_N^{(t)}}\right)^\top ,$$

and then the update rule (28) should be understood as

$$\boldsymbol{T}^{(t+1)}(s,a) = \boldsymbol{W}\left(\boldsymbol{T}^{(t)}(s,a) + \boldsymbol{q}_\tau^{(t+1)}(s,a) - \boldsymbol{q}_\tau^{(t)}(s,a)\right) \tag{67}$$

in the inexact setting. For notational simplicity, we define $e_n \in \mathbb{R}$ as

$$e_n := \max_{t\in[T]} \left\|Q_{\tau,n}^{\pi_n^{(t)}} - q_{\tau,n}^{\pi_n^{(t)}}\right\|_\infty , \quad n \in [N] , \tag{68}$$

and let $\boldsymbol{e} = (e_1, \cdots, e_n)^\top$. Define $\widehat{q}_\tau^{(t)}$, the approximation of $\widehat{Q}_\tau^{(t)}$ as

$$\widehat{q}_\tau^{(t)} := \frac{1}{N}\sum_{n=1}^N q_{\tau,n}^{\pi_n^{(t)}} . \tag{69}$$

With slight abuse of notation, we adapt the auxiliary sequence $\{\overline{\xi}^{(t)}\}_{t=0,\cdots}$ to the inexact updates as

$$\overline{\xi}^{(0)}(s,a) := \left\|\exp\left(Q_\tau^\star(s,\cdot)/\tau\right)\right\|_1 \cdot \overline{\pi}^{(0)}(a|s) , \tag{70a}$$

$$\overline{\xi}^{(t+1)}(s,a) := \left[\overline{\xi}^{(t)}(s,a)\right]^\alpha \exp\left((1-\alpha)\frac{\widehat{q}_\tau^{(t)}(s,a)}{\tau}\right) , \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A},\ t \geq 0 . \tag{70b}$$

In addition, we define

$$\Omega_1^{(t)} := \left\| u^{(t)} \right\|_\infty , \tag{71a}$$

$$\Omega_2^{(t)} := \left\| v^{(t)} \right\|_\infty , \tag{71b}$$

$$\Omega_3^{(t)} := \left\| Q_\tau^\star - \tau \log \overline{\xi}^{(t)} \right\|_\infty , \tag{71c}$$

$$\Omega_4^{(t)} := \max \left\{ 0, -\min_{s,a} \left(\overline{q}_\tau^{(t)}(s,a) - \tau \log \overline{\xi}^{(t)}(s,a) \right) \right\} , \tag{71d}$$

where

$$u^{(t)}(s,a) := \left\| \log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\xi}^{(t)}(s,a) \mathbf{1}_N \right\|_2 , \tag{72}$$

$$v^{(t)}(s,a) := \left\| \boldsymbol{T}^{(t)}(s,a) - \widehat{q}_\tau^{(t)}(s,a) \mathbf{1}_N \right\|_2 . \tag{73}$$

We let $\boldsymbol{\Omega}^{(t)}$ be

$$\boldsymbol{\Omega}^{(t)} := \left(\Omega_1^{(t)}, \Omega_2^{(t)}, \Omega_3^{(t)}, \Omega_4^{(t)} \right)^\top . \tag{74}$$

With the above preparation, we are ready to state the inexact convergence guarantee of Algorithm 2 in Theorem 7 below, which is the formal version of Theorem 4.

Theorem 7. *Suppose that $q_{\tau,n}^{\pi_n^{(t)}}$ are used in replace of $Q_{\tau,n}^{\pi_n^{(t)}}$ in Algorithm 2. For any $N \in \mathbb{N}_+, \tau > 0, \gamma \in (0,1)$, there exists $\eta_0 > 0$ which depends only on $N, \gamma, \tau, \sigma, |\mathcal{A}|$, such that if $0 < \eta \le \eta_0$ and $1 - \sigma > 0$, we have*

$$\left\| \overline{Q}_\tau^{(t)} - Q_\tau^\star \right\|_\infty \le 2\gamma \left(\rho(\eta)^t \left\| \boldsymbol{\Omega}^{(0)} \right\|_2 + C_2 \max_{n \in [N], t \in [T]} \left\| Q_{\tau,n}^{\pi_n^{(t)}} - q_{\tau,n}^{\pi_n^{(t)}} \right\|_\infty \right) , \tag{75}$$

$$\left\| \log \pi_\tau^\star - \log \overline{\pi}^{(t)} \right\|_\infty \le \frac{2}{\tau} \left(\rho(\eta)^t \left\| \boldsymbol{\Omega}^{(0)} \right\|_2 + C_2 \max_{n \in [N], t \in [T]} \left\| Q_{\tau,n}^{\pi_n^{(t)}} - q_{\tau,n}^{\pi_n^{(t)}} \right\|_\infty \right) . \tag{76}$$

Moreover, the consensus errors satisfy:

$$\forall n \in [N] : \quad \left\| \log \pi_n^{(t)} - \log \overline{\pi}^{(t)} \right\|_\infty \le 2 \left(\rho(\eta)^t \left\| \boldsymbol{\Omega}^{(0)} \right\|_2 + C_2 \max_{n \in [N], t \in [T]} \left\| Q_{\tau,n}^{\pi_n^{(t)}} - q_{\tau,n}^{\pi_n^{(t)}} \right\|_\infty \right) , \tag{77}$$

where $\rho(\eta) \le \max\{1 - \frac{\tau\eta}{2}, \frac{3+\sigma}{4}\} < 1$ is the same as in Theorem 6, and $C_2 := \frac{\sigma\sqrt{N}(2(1-\gamma)+M\sqrt{N}\eta)+2\gamma^2+\eta\tau}{(1-\gamma)(1-\rho(\eta))}$.

From Theorem 7, we can conclude that if

$$\max_{n \in [N], t \in [T]} \left\| Q_{\tau,n}^{\pi_n^{(t)}} - q_{\tau,n}^{\pi_n^{(t)}} \right\|_\infty \le \frac{(1-\gamma)(1-\rho(\eta))\varepsilon}{2\gamma \left(\sigma\sqrt{N}(2(1-\gamma) + M\sqrt{N}\eta) + 2\gamma^2 + \eta\tau \right)} , \tag{78}$$

then inexact entropy-regularized FedNPG could still achieve 2ε-accuracy (i.e. $\left\| \overline{Q}_\tau^{(t)} - Q_\tau^\star \right\|_\infty \le 2\varepsilon$) within $\max\left\{ \frac{2}{\tau\eta}, \frac{4}{1-\sigma} \right\} \log \frac{2\gamma \left\| \boldsymbol{\Omega}^{(0)} \right\|_2}{\varepsilon}$ iterations.

Remark 2. *When $\eta = \eta_0$ (cf. (65) and (64)) and $\tau \le 1$, the RHS of (78) is of the order*

$$\mathcal{O}\left(\frac{(1-\gamma)\tau\eta_0\varepsilon}{\gamma(\gamma^2 + \sigma\sqrt{N}(1-\gamma))} \right) = \mathcal{O}\left(\frac{(1-\gamma)^8\tau^2(1-\sigma)^2\varepsilon}{\gamma(\gamma^2 + \sigma\sqrt{N}(1-\gamma))(\gamma^2 N\sigma^2 + (1-\sigma)^2\tau^2(1-\gamma)^6)} \right) ,$$

which can be translated into a crude sample complexity bound when using fresh samples to estimate the soft Q-functions in each iteration.

The rest of this section outlines the proof of Theorem 7. We first state a key lemma that tracks the error recursion of Algorithm 2 with inexact policy evaluation, which is a modified version of Lemma 1.

Lemma 3. *The following linear system holds for all $t \geq 0$:*

$$\boldsymbol{\Omega}^{(t+1)} \leq \boldsymbol{A}(\eta)\boldsymbol{\Omega}^{(t)} + \underbrace{\begin{pmatrix} 0 \\ \sigma\sqrt{N}\left(2 + \frac{M\sqrt{N}\eta}{1-\gamma}\right) \\ \frac{\eta\tau}{1-\gamma} \\ \frac{2\gamma^2}{1-\gamma} \end{pmatrix}}_{=:\boldsymbol{b}(\eta)} \|\boldsymbol{e}\|_\infty , \tag{79}$$

where $\boldsymbol{A}(\eta)$ is provided in Lemma 1. In addition, it holds for all $t \geq 0$ that

$$\left\|\overline{Q}_\tau^{(t)} - Q_\tau^\star\right\|_\infty \leq \gamma\Omega_3^{(t)} + \gamma\Omega_4^{(t)} , \tag{80}$$

$$\left\|\log\overline{\pi}^{(t)} - \log\pi_\tau^\star\right\|_\infty \leq \frac{2}{\tau}\Omega_3^{(t)} . \tag{81}$$

Proof. See Appendix C.3. □

By (79), we have

$$\forall t \in N_+ : \quad \boldsymbol{\Omega}^{(t)} \leq \boldsymbol{A}(\eta)^t\boldsymbol{\Omega}^{(0)} + \sum_{s=1}^{t} \boldsymbol{A}(\eta)^{t-s}\boldsymbol{b}(\eta) ,$$

which gives

$$\begin{aligned} \left\|\boldsymbol{\Omega}^{(t)}\right\|_2 &\leq \rho(\eta)^t \left\|\boldsymbol{\Omega}^{(0)}\right\|_2 + \sum_{s=1}^{t} \rho(\eta)^{t-s} \|\boldsymbol{b}(\eta)\|_2 \|\boldsymbol{e}\|_\infty \\ &\leq \rho(\eta)^t \left\|\boldsymbol{\Omega}^{(0)}\right\|_2 + \frac{\sigma\sqrt{N}(2(1-\gamma) + M\sqrt{N}\eta) + 2\gamma^2 + \eta\tau}{(1-\gamma)(1-\rho(\eta))} \|\boldsymbol{e}\|_\infty . \end{aligned} \tag{82}$$

Here, (82) follows from $\|\boldsymbol{b}(\eta)\|_2 \leq \|\boldsymbol{b}(\eta)\|_1 = \frac{\sigma\sqrt{N}(2(1-\gamma)+M\sqrt{N}\eta)+2\gamma^2+\eta\tau}{1-\gamma} \|\boldsymbol{e}\|_\infty$ and $\sum_{s=1}^{t} \rho(\eta)^{t-s} \leq 1/(1-\rho(\eta))$. Recall that the bound on $\rho(\eta)$ has already been established in Lemma 2. Therefore we complete the proof of Theorem 7 by combining the above inequality with (80) and (81) in a similar fashion as before. We omit further details for conciseness.

A.3 Analysis of FedNPG with exact policy evaluation

We state the formal version of Theorem 1 below.

Theorem 8. *Suppose all $\pi_n^{(0)}$ in Algorithm 1 are initialized as uniform distribution. When*

$$0 < \eta \leq \eta_1 := \frac{(1-\sigma)^2(1-\gamma)^3}{8(1+\gamma)\gamma\sqrt{N}\sigma^2} ,$$

we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(V^\star(\rho) - V^{\overline{\pi}^{(t)}}(\rho)\right) \leq \frac{V^\star(d_\rho^{\pi^\star})}{(1-\gamma)T} + \frac{\log|\mathcal{A}|}{\eta T} + \frac{8(1+\gamma)^2\gamma^2 N\sigma^2}{(1-\gamma)^9(1-\sigma)^2}\eta^2 \tag{83}$$

for any fixed state distribution ρ. Furthermore, we have

$$\forall n \in [N] : \quad \left\|\log\pi_n^{(t)} - \log\overline{\pi}^{(t)}\right\|_\infty \leq \frac{32N\sigma}{3(1-\gamma)^4(1-\sigma)}\eta . \tag{84}$$

The rest of this section is dedicated to prove Theorem 8. Similar to (48), we denote the Q-functions of $\overline{\pi}^{(t)}$ by $\overline{\boldsymbol{Q}}^{(t)}$:

$$\overline{\boldsymbol{Q}}^{(t)} := \begin{pmatrix} Q_1^{\overline{\pi}^{(t)}} \\ \vdots \\ Q_N^{\overline{\pi}^{(t)}} \end{pmatrix}. \tag{85}$$

In addition, similar to (49), we define $\widehat{Q}^{(t)}, \overline{Q}^{(t)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\overline{V}^{(t)} \in \mathbb{R}^{|\mathcal{S}|}$ as follows

$$\widehat{Q}^{(t)} := \frac{1}{N}\sum_{n=1}^{N} Q_n^{\pi_n^{(t)}}, \tag{86a}$$

$$\overline{Q}^{(t)} := Q^{\overline{\pi}^{(t)}} = \frac{1}{N}\sum_{n=1}^{N} Q_n^{\overline{\pi}^{(t)}}. \tag{86b}$$

$$\overline{V}^{(t)} := V^{\overline{\pi}^{(t)}} = \frac{1}{N}\sum_{n=1}^{N} V_n^{\overline{\pi}^{(t)}}. \tag{86c}$$

Following the same strategy in the analysis of entropy-regularized FedNPG, we introduce the auxiliary sequence $\{\boldsymbol{\xi}^{(t)} = (\xi_1^{(t)}, \cdots, \xi_N^{(t)})^\top \in \mathbb{R}^{N\times|\mathcal{S}||\mathcal{A}|}\}$ recursively:

$$\boldsymbol{\xi}^{(0)}(s,a) := \frac{1}{\left\|\exp\left(\frac{1}{N}\sum_{n=1}^{N}\log \pi_n^{(0)}(\cdot|s)\right)\right\|_1} \cdot \boldsymbol{\pi}^{(0)}(a|s), \tag{87a}$$

$$\log \boldsymbol{\xi}^{(t+1)}(s,a) = \boldsymbol{W}\left(\log \boldsymbol{\xi}^{(t)}(s,a) + \frac{\eta}{1-\gamma}\boldsymbol{T}^{(t)}(s,a)\right), \tag{87b}$$

as well as the averaged auxiliary sequence $\{\overline{\xi}^{(t)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}\}$:

$$\overline{\xi}^{(0)}(s,a) := \overline{\pi}^{(0)}(a|s), \tag{88a}$$

$$\log \overline{\xi}^{(t+1)}(s,a) := \log \overline{\xi}^{(t)}(s,a) + \frac{\eta}{1-\gamma}\widehat{Q}^{(t)}(s,a), \quad \forall (s,a)\in\mathcal{S}\times\mathcal{A},\ t\geq 0. \tag{88b}$$

As usual, we collect the consensus errors in a vector $\boldsymbol{\Omega}^{(t)} = (\|u^{(t)}\|_\infty, \|v^{(t)}\|_\infty)^\top$, where $u^{(t)}, v^{(t)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ are defined as:

$$u^{(t)}(s,a) := \left\|\log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\xi}^{(t)}(s,a)\mathbf{1}_N\right\|_2, \tag{89}$$

$$v^{(t)}(s,a) := \left\|\boldsymbol{T}^{(t)}(s,a) - \widehat{Q}^{(t)}(s,a)\mathbf{1}_N\right\|_2. \tag{90}$$

Step 1: establishing the error recursion. The next key lemma establishes the error recursion of Algorithm 1.

Lemma 4. *The updates of FedNPG satisfy*

$$\boldsymbol{\Omega}^{(t+1)} \leq \underbrace{\begin{pmatrix} \sigma & \frac{\eta}{1-\gamma}\sigma \\ J\sigma & \sigma\left(1 + \frac{(1+\gamma)\gamma\sqrt{N}\eta}{(1-\gamma)^3}\sigma\right)\end{pmatrix}}_{=:\boldsymbol{B}(\eta)} \boldsymbol{\Omega}^{(t)} + \underbrace{\begin{pmatrix} 0 \\ \frac{(1+\gamma)\gamma N\sigma}{(1-\gamma)^4}\eta \end{pmatrix}}_{=:\boldsymbol{d}(\eta)} \tag{91}$$

for all $t \geq 0$, where

$$J := \frac{2(1+\gamma)\gamma}{(1-\gamma)^2}\sqrt{N}. \tag{92}$$

In addition, we have

$$\phi^{(t+1)}(\eta) \leq \phi^{(t)}(\eta) + \frac{2(1+\gamma)\gamma}{(1-\gamma)^4}\eta\|u^{(t)}\|_\infty - \eta\left(V^\star(\rho) - \overline{V}^{(t)}(\rho)\right), \tag{93}$$

where

$$\phi^{(t)}(\eta) \coloneqq \mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathsf{KL}\big(\pi^\star(\cdot|s)\,\|\,\overline{\pi}^{(t)}(\cdot|s)\big)\right] - \frac{\eta}{1-\gamma}\overline{V}^{(t)}(d_\rho^{\pi^\star})\,, \quad \forall t\geq 0\,. \tag{94}$$

Moreover, when $\eta \leq \eta_1$, we have

$$\forall n\in[N]: \quad \left\|\log\pi_n^{(t)} - \log\overline{\pi}^{(t)}\right\|_\infty \leq 2\left(\frac{3}{8}\sigma+\frac{5}{8}\right)^t \left\|\mathbf{\Omega}^{(0)}\right\|_2 + \frac{32N\sigma}{3(1-\gamma)^4(1-\sigma)}\eta\,. \tag{95}$$

Proof. See Appendix C.4. □

Note that when all $\pi_n^{(0)}$ in Algorithm 1 are initialized as uniform distribution, $\mathbf{\Omega}^{(0)} = \mathbf{0}$ and (95) indicates (84) in Theorem 8.

Step 2: bounding the value functions. Let $\boldsymbol{p}\in\mathbb{R}^2$ be defined as:

$$\boldsymbol{p}(\eta) = \begin{pmatrix} p_1(\eta) \\ p_2(\eta)\end{pmatrix} \coloneqq \frac{2(1+\gamma)\gamma}{(1-\gamma)^4}\begin{pmatrix} \frac{\sigma(1-\gamma)\left(1-\sigma-(1+\gamma)\gamma\sqrt{N}\sigma\eta/(1-\gamma)^3\right)\eta}{(1-\gamma)\left(1-\sigma-(1+\gamma)\gamma\sqrt{N}\sigma^2\eta/(1-\gamma)^3\right)(1-\sigma)-J\sigma^2\eta} \\ \frac{\sigma\eta^2}{(1-\gamma)\left(1-\sigma-(1+\gamma)\gamma\sqrt{N}\sigma^2\eta/(1-\gamma)^3\right)(1-\sigma)-J\sigma^2\eta}\end{pmatrix}; \tag{96}$$

the rationale for this choice will be made clear momentarily. We define the following Lyapunov function

$$\Phi^{(t)}(\eta) = \phi^{(t)}(\eta) + \boldsymbol{p}(\eta)^\top\mathbf{\Omega}^{(t)}\,, \quad \forall t\geq 0\,, \tag{97}$$

which satisfies

$$\begin{aligned}
\Phi^{(t+1)}(\eta) &= \phi^{(t+1)}(\eta) + \boldsymbol{p}(\eta)^\top\mathbf{\Omega}^{(t+1)} \\
&\leq \phi^{(t)}(\eta) + \frac{2(1+\gamma)\gamma}{(1-\gamma)^4}\eta\left\|u^{(t)}\right\|_\infty - \eta\left(V^\star(\rho)-\overline{V}^{(t)}(\rho)\right) + \boldsymbol{p}(\eta)^\top\left(\boldsymbol{B}(\eta)\mathbf{\Omega}^{(t)}+\boldsymbol{d}(\eta)\right) \\
&= \Phi^{(t)}(\eta) + \left[\boldsymbol{p}(\eta)^\top\left(\boldsymbol{B}(\eta)-\boldsymbol{I}\right) + \left(\frac{2(1+\gamma)\gamma}{(1-\gamma)^4}\eta, 0\right)\right]\mathbf{\Omega}^{(t)} - \eta\left(V^\star(\rho)-\overline{V}^{(t)}(\rho)\right) \\
&\quad + p_2(\eta)\frac{(1+\gamma)\gamma N\sigma}{(1-\gamma)^4}\eta\,.
\end{aligned} \tag{98}$$

Here, the second inequality follows from (93). One can verify that the second term vanishes due to the choice of $\boldsymbol{p}(\eta)$:

$$\boldsymbol{p}(\eta)^\top\left(\boldsymbol{B}(\eta)-\boldsymbol{I}\right) + \left(\frac{2(1+\gamma)\gamma}{(1-\gamma)^4}\eta, 0\right) = (0,0)\,. \tag{99}$$

Therefore, we conclude that

$$V^\star(\rho) - \overline{V}^{(t)}(\rho) \leq \frac{\Phi^{(t)}(\eta)-\Phi^{(t+1)}(\eta)}{\eta} + p_2(\eta)\frac{(1+\gamma)\gamma N\sigma}{(1-\gamma)^4}\,.$$

Averaging over $t = 0,\cdots,T-1$,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(V^\star(\rho)-\overline{V}^{(t)}(\rho)\right) \\
&\leq \frac{\Phi^{(0)}(\eta)-\Phi^{(T)}(\eta)}{\eta T} + \frac{2(1+\gamma)^2\gamma^2}{(1-\gamma)^8}\cdot\frac{N\sigma^2\eta^2}{(1-\gamma)(1-\sigma-(1+\gamma)\gamma\sqrt{N}\sigma^2\eta/(1-\gamma)^3)(1-\sigma)-\sigma^2J\eta}\,.
\end{aligned} \tag{100}$$

Step 3: simplifying the expression. We first upper bound the first term in the RHS of (100). Assuming uniform initialization for all $\pi_n^{(0)}$ in Algorithm 1, we have $\left\|u^{(0)}\right\|_\infty = \left\|v^{(0)}\right\|_\infty = 0$, and

$$\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathsf{KL}\big(\pi^\star(\cdot|s)\,\|\,\overline{\pi}^{(0)}(\cdot|s)\big)\right] \le \log|\mathcal{A}|.$$

Therefore, putting together relations (97) and (232) we have

$$\frac{\Phi^{(0)}(\eta)-\Phi^{(T)}(\eta)}{\eta T} \le \frac{\log|\mathcal{A}|}{T\eta} + \frac{1}{T}\left(\boldsymbol{p}(\eta)^\top \boldsymbol{\Omega}^{(0)}/\eta + \frac{V^\star(d_\rho^{\pi^\star})}{1-\gamma}\right) = \frac{\log|\mathcal{A}|}{T\eta} + \frac{V^\star(d_\rho^{\pi^\star})}{T(1-\gamma)}\,, \tag{101}$$

To continue, we upper bound the second term in the RHS of (100). Note that

$$\eta \le \eta_1 \le \frac{(1-\sigma)(1-\gamma)^3}{2(1+\gamma)\gamma\sqrt{N}\sigma^2}\,,$$

which gives

$$\frac{(1+\gamma)\gamma\sqrt{N}\sigma^2}{(1-\gamma)^3}\eta \le \frac{1-\sigma}{2}. \tag{102}$$

Thus we have

$$\begin{aligned}
&(1-\gamma)(1-\sigma-(1+\gamma)\gamma\sqrt{N}\sigma^2\eta/(1-\gamma)^3)(1-\sigma) - J\sigma^2\eta \\
&\ge (1-\gamma)(1-\sigma)^2/2 - J\sigma^2\eta_1 \\
&\ge (1-\gamma)(1-\sigma)^2/4\,,
\end{aligned} \tag{103}$$

where the first inequality follows from (102) and the second inequality follows from the definition of η_1 and J. By (103), we deduce

$$\frac{2(1+\gamma)^2\gamma^2}{(1-\gamma)^8}\cdot\frac{N\sigma^2\eta^2}{(1-\gamma)(1-\sigma-(1+\gamma)\gamma\sqrt{N}\sigma^2\eta/(1-\gamma)^3)(1-\sigma)-J\sigma^2\eta} \le \frac{8(1+\gamma)^2\gamma^2N\sigma^2}{(1-\gamma)^9(1-\sigma)^2}\eta^2\,, \tag{104}$$

and our advertised bound (83) thus follows from plugging (101) and (104) into (100).

A.4 Analysis of FedNPG with inexact policy evaluation

We state the formal version of Theorem 2 below.

Theorem 9. *Suppose that $q_n^{\pi_n^{(t)}}$ are used in replace of $Q_n^{\pi_n^{(t)}}$ in Algorithm 1. Suppose all $\pi_n^{(0)}$ in Algorithm 1 set to uniform distribution. Let*

$$0 < \eta \le \eta_1 := \frac{(1-\sigma)^2(1-\gamma)^3}{8(1+\gamma)\gamma\sqrt{N}\sigma^2}\,,$$

we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(V^\star(\rho) - V^{\overline{\pi}^{(t)}}(\rho)\right) \\
&\le \frac{V^\star(d_\rho^{\pi^\star})}{(1-\gamma)T} + \frac{\log|\mathcal{A}|}{\eta T} + \frac{8(1+\gamma)^2\gamma^2N\sigma^2}{(1-\gamma)^9(1-\sigma)^2}\eta^2 \\
&\quad + \left[\frac{8(1+\gamma)\gamma}{(1-\gamma)^5(1-\sigma)^2}\sqrt{N}\sigma\eta\left(\frac{(1+\gamma)\gamma\eta\sqrt{N}}{(1-\gamma)^3}+2\right) + \frac{2}{(1-\gamma)^2}\right]\max_{n\in[N],t\in[T]}\left\|Q_n^{\pi_n^{(t)}} - q_n^{\pi_n^{(t)}}\right\|_\infty
\end{aligned}$$

for any fixed state distribution ρ.

Furthermore, we have

$$\forall n\in[N]: \quad \left\|\log\pi_n^{(t)} - \log\overline{\pi}^{(t)}\right\|_\infty \le \frac{32}{3(1-\sigma)}\left(\frac{N\sigma}{(1-\gamma)^4}\eta + \sqrt{N}\sigma\left(\frac{\eta\sqrt{N}}{(1-\gamma)^3}+1\right)\max_{n\in[N],t\in[T]}\left\|Q_n^{\pi_n^{(t)}} - q_n^{\pi_n^{(t)}}\right\|_\infty\right). \tag{105}$$

We next outline the proof of Theorem 9. With slight abuse of notation, we again define $e_n \in \mathbb{R}$ as

$$e_n := \max_{t\in[T]} \left\| Q_n^{\pi_n^{(t)}} - q_n^{\pi_n^{(t)}} \right\|_\infty , \quad n \in [N] , \tag{106}$$

and let $\boldsymbol{e} = (e_1, \cdots, e_n)^\top$. We define the collection of *inexact* Q-function estimates as

$$\boldsymbol{q}^{(t)} := \left(q_1^{\pi_1^{(t)}}, \cdots, q_N^{\pi_N^{(t)}} \right)^\top ,$$

and then the update rule (26) should be understood as

$$\boldsymbol{T}^{(t+1)}(s,a) = \boldsymbol{W} \left(\boldsymbol{T}^{(t)}(s,a) + \boldsymbol{q}^{(t+1)}(s,a) - \boldsymbol{q}^{(t)}(s,a) \right) \tag{107}$$

in the inexact setting. Define $\widehat{q}^{(t)}$, the approximation of $\widehat{Q}^{(t)}$ as

$$\widehat{q}^{(t)} := \frac{1}{N} \sum_{n=1}^{N} q_n^{\pi_n^{(t)}} , \tag{108}$$

we adapt the averaged auxiliary sequence $\{\overline{\xi}^{(t)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}\}$ to the inexact updates as follows:

$$\overline{\xi}^{(0)}(s,a) := \overline{\pi}^{(0)}(a|s) , \tag{109a}$$

$$\overline{\xi}^{(t+1)}(s,a) := \overline{\xi}^{(t)}(s,a) \exp\left(\frac{\eta}{1-\gamma} \widehat{q}^{(t)}(s,a) \right) , \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A},\ t \geq 0 . \tag{109b}$$

As usual, we define the consensus error vector as $\boldsymbol{\Omega}^{(t)} = (\|u^{(t)}\|_\infty, \|v^{(t)}\|_\infty)^\top$, where $u^{(t)}, v^{(t)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ are given by

$$u^{(t)}(s,a) := \left\| \log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\xi}^{(t)}(s,a) \mathbf{1}_N \right\|_2 , \tag{110}$$

$$v^{(t)}(s,a) := \left\| \boldsymbol{T}^{(t)}(s,a) - \widehat{q}^{(t)}(s,a) \mathbf{1}_N \right\|_2 . \tag{111}$$

The following lemma characterizes the dynamics of the error vector $\boldsymbol{\Omega}^{(t)}$, perturbed by additional approximation error.

Lemma 5. *The updates of inexact FedNPG satisfy*

$$\boldsymbol{\Omega}^{(t+1)} \leq \boldsymbol{B}(\eta)\boldsymbol{\Omega}^{(t)} + \boldsymbol{d}(\eta) + \underbrace{\begin{pmatrix} 0 \\ \sqrt{N}\sigma \left(\frac{(1+\gamma)\gamma\eta\sqrt{N}}{(1-\gamma)^3} + 2 \right) \end{pmatrix} \|\boldsymbol{e}\|_\infty}_{=:\boldsymbol{c}(\eta)} . \tag{112}$$

In addition, we have

$$\phi^{(t+1)}(\eta) \leq \phi^{(t)}(\eta) + \frac{2(1+\gamma)\gamma}{(1-\gamma)^4}\eta \left\| u^{(t)} \right\|_\infty + \frac{2\eta}{(1-\gamma)^2} \|\boldsymbol{e}\|_\infty - \eta \left(V^\star(\rho) - \overline{V}^{(t)}(\rho) \right) , \tag{113}$$

where $\phi^{(t)}(\eta)$ is defined in (94).

Moreover, when $\eta \leq \eta_1$, we have

$$\forall n \in [N] : \quad \left\| \log \pi_n^{(t)} - \log \overline{\pi}^{(t)} \right\|_\infty \leq 2 \left(\frac{3}{8}\sigma + \frac{5}{8} \right)^t \left\| \boldsymbol{\Omega}^{(0)} \right\|_2 + \frac{32}{3(1-\sigma)} \left(\frac{N\sigma}{(1-\gamma)^4}\eta + \sqrt{N}\sigma \left(\frac{\eta\sqrt{N}}{(1-\gamma)^3} + 1 \right) \|\boldsymbol{e}\|_\infty \right) . \tag{114}$$

Proof. See Appendix C.5. □

Similar to (98), we can recursively bound $\Phi^{(t)}(\eta)$ (defined in (97)) as

$$\begin{aligned}
\Phi^{(t+1)}(\eta) &= \phi^{(t+1)}(\eta) + \boldsymbol{p}(\eta)^\top \boldsymbol{\Omega}^{(t+1)} \\
&\overset{(113)}{\leq} \phi^{(t)}(\eta) + \frac{2(1+\gamma)\gamma}{(1-\gamma)^4}\eta \left\| u^{(t)} \right\|_\infty + \frac{2\eta}{(1-\gamma)^2} \|\boldsymbol{e}\|_\infty - \eta \left(V^\star(\rho) - \overline{V}^{(t)}(\rho) \right) \\
&\qquad + \boldsymbol{p}(\eta)^\top \left(\boldsymbol{B}(\eta)\boldsymbol{\Omega}^{(t)} + \boldsymbol{d}(\eta) + \boldsymbol{c}(\eta) \right) \\
&= \Phi^{(t)}(\eta) + \underbrace{\left[\boldsymbol{p}(\eta)^\top \left(\boldsymbol{B}(\eta) - \boldsymbol{I} \right) + \left(\frac{2(1+\gamma)\gamma}{(1-\gamma)^4}\eta, 0 \right) \right]}_{=(0,0) \text{ via } (99)} \boldsymbol{\Omega}^{(t)} - \eta \left(V^\star(\rho) - \overline{V}^{(t)}(\rho) \right) \\
&\qquad + p_2(\eta) \frac{(1+\gamma)\gamma N \sigma}{(1-\gamma)^4}\eta + \left[p_2(\eta)\sqrt{N}\sigma \left(\frac{(1+\gamma)\gamma\eta\sqrt{N}}{(1-\gamma)^3} + 2 \right) + \frac{2\eta}{(1-\gamma)^2} \right] \|\boldsymbol{e}\|_\infty . \qquad (115)
\end{aligned}$$

From the above expression we know that

$$V^\star(\rho) - \overline{V}^{(t)}(\rho) \leq \frac{\Phi^{(t)}(\eta) - \Phi^{(t+1)}(\eta)}{\eta} + p_2(\eta)\frac{(1+\gamma)\gamma N\sigma}{(1-\gamma)^4} + \left[p_2(\eta)\sqrt{N}\sigma \left(\frac{(1+\gamma)\gamma\sqrt{N}}{(1-\gamma)^3} + \frac{2}{\eta} \right) + \frac{2}{(1-\gamma)^2} \right] \|\boldsymbol{e}\|_\infty ,$$

which gives

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \left(V^\star(\rho) - \overline{V}^{(t)}(\rho) \right) &\leq \frac{\Phi^{(0)}(\eta) - \Phi^{(T)}(\eta)}{\eta T} + p_2(\eta)\frac{(1+\gamma)\gamma N\sigma}{(1-\gamma)^4} \\
&\qquad + \left[p_2(\eta)\sqrt{N}\sigma \left(\frac{(1+\gamma)\gamma\sqrt{N}}{(1-\gamma)^3} + \frac{2}{\eta} \right) + \frac{2}{(1-\gamma)^2} \right] \|\boldsymbol{e}\|_\infty \qquad (116)
\end{aligned}$$

via telescoping. Combining the above expression with (101), (103) and (104), we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \left(V^\star(\rho) - \overline{V}^{(t)}(\rho) \right) &\leq \frac{\log|\mathcal{A}|}{T\eta} + \frac{V^\star(d_\rho^{\pi^\star})}{T(1-\gamma)} + \frac{8(1+\gamma)^2\gamma^2 N\sigma}{(1-\gamma)^9(1-\sigma)^2}\eta^2 \\
&\qquad + \left[\frac{8(1+\gamma)\gamma}{(1-\gamma)^5(1-\sigma)^2}\sqrt{N}\sigma\eta \left(\frac{(1+\gamma)\gamma\eta\sqrt{N}}{(1-\gamma)^3} + 2 \right) + \frac{2}{(1-\gamma)^2} \right] \|\boldsymbol{e}\|_\infty , \qquad (117)
\end{aligned}$$

which establishes (105).

B Convergence analysis of FedNAC

Let $\pi^\star$ be an optimal policy and does not need to belong to the log-linear policy class. Fix a state distribution $\rho \in \Delta(\mathcal{S})$ and a state-action distribution ν. To simplify the notation, we denote $d_\rho^{\pi^\star}$ as $d_\star$, $d^{f_{\boldsymbol{\xi}^{(t)}}}$ as $d^{(t)}$, $\tilde{d}_n^{(t)}$ as $\tilde{d}_\nu^{f_{\xi_n^{(t)}}}$, and define $d_n^{(t)}$ and $\bar{d}_n^{(t)}$ analogously. We also let $Q_n^{(t)}$ denote $Q_n^{\xi_n^{(t)}}$.

Define

$$\vartheta_\rho \coloneqq \frac{1}{1-\gamma}\left\| \frac{d_\star}{\rho} \right\|_\infty \geq \frac{1}{1-\gamma} \qquad (118)$$

and assume $\vartheta_\rho < \infty$.

We also introduce a weighted KL divergence given by

$$D_\star^{(t)} \coloneqq \mathbb{E}_{s\sim d_\star}\left[\mathsf{KL}\big(\pi^\star(\cdot|s) \,\|\, \pi^{(t)}(\cdot|s)\big) \right] , \qquad (119)$$

where $\mathsf{KL}(\cdot\,\|\,\cdot) : \mathbb{R}^{|\mathcal{A}|} \times \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}$ is the Kullback-Leibler (KL) divergence:

$$\forall f, g \in \mathbb{R}^{|\mathcal{A}|} : \quad \mathsf{KL}(f\,\|\,g) \coloneqq \sum_{a\in\mathcal{A}} f(a)\log\left(\frac{f(a)}{g(a)} \right) . \qquad (120)$$

Given a state distribution ρ and an optimal policy $\pi^\star$, we define a state-action measure $\tilde{d}^\star$ as

$$\tilde{d}^\star(s,a) \coloneqq d_\star(s) \cdot \mathrm{Unif}_{\mathcal{A}}(a) = \frac{d_\star(s)}{|\mathcal{A}|}. \tag{121}$$

The following theorem guarantees that for any fixed policy π and state-action distribution $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$, the Q-Sampler algorithm (cf. Algorithm 5) samples (s,a) from $\tilde{d}_\nu^\pi$ and gives an unbiased estimate $\widehat{Q}^\pi(s,a)$ of $Q^\pi(s,a)$, whose proof can be found in Yuan et al. (2022, Lemma 4).

Lemma 6 (Lemma 4 in Yuan et al. (2022))**.** *Consider the output (s_h, a_h) and $\widehat{Q}^\pi(s_h, a_h)$ of Algorithm 5. It follows that*

$$\begin{aligned}
\mathbb{E}[h+1] &= \frac{1}{1-\gamma}\,, \\
P(s_h = s, a_h = a) &= \tilde{d}_\nu^\pi(s,a)\,, \\
\mathbb{E}\left[\widehat{Q}^\pi(s_h,a_h) | s_h, a_h\right] &= Q^\pi(s_h,a_h)\,.
\end{aligned}$$

To present the convergence results of FedNAC, we further introduce the following notation, where $t \in \mathbb{N}$ represents the iteration step in FedNAC:

$$\hat{\boldsymbol{w}}^{(t)} \coloneqq \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{w}_n^{(t)}, \tag{122a}$$

$$\bar{\boldsymbol{\xi}}^{(t)} \coloneqq \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\xi}_n^{(t)}, \tag{122b}$$

$$\bar{f}^{(t)} \coloneqq f_{\bar{\xi}^{(t)}}, \tag{122c}$$

$$f_n^{(t)} \coloneqq f_{\xi_n^{(t)}}, \tag{122d}$$

$$\boldsymbol{w}_{\star,n}^{(t)} \in \arg\min_{\boldsymbol{w}} \ell\left(\boldsymbol{w}, Q_n^{(t)}, \tilde{d}_n^{(t)}\right), \tag{122e}$$

$$\hat{\boldsymbol{w}}_\star^{(t)} \coloneqq \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{w}_{\star,n}^{(t)}. \tag{122f}$$

For convenience of narration, we introduce the following bounded statistical error assumption.

Assumption 5 (Bounded statistical error)**.** *For all $n \in [N]$, there exists $\varepsilon_{stat}^n > 0$ such that for all $t \in \mathbb{N}$ in Algorithm 3, we have*

$$\mathbb{E}\left[\ell\left(\boldsymbol{w}_n^{(t)}, Q_n^{(t)}, \tilde{d}_n^{(t)}\right) - \ell\left(\boldsymbol{w}_{\star,n}^{(t)}, Q_n^{(t)}, \tilde{d}_n^{(t)}\right)\right] \le \varepsilon_{stat}^n. \tag{123}$$

When solving the regression problem with sampling based approaches, we can expect $\varepsilon_{\mathrm{stat}}^n = \mathcal{O}(1/K)$, where K is the iteration number of Algorithm 4.

Theorem 10 (Convergence rate of Critic (Algorithm 4))**.** *For Algorithm 4, let $\boldsymbol{w}_0 = \mathbf{0}$ and $\beta = \frac{1}{2C_\phi}$. Then under Assumption 2, we have*

$$\mathbb{E}\left[\ell\left(\boldsymbol{w}_{out}, Q_\xi, \tilde{d}_\xi\right)\right] - \ell\left(\boldsymbol{w}^\star, Q_\xi, \tilde{d}_\xi\right) \le \frac{4}{K}\left(\frac{\sqrt{2p}}{1-\gamma}\left(\frac{C_\phi^2}{\mu(1-\gamma)} + 1\right) + \frac{C_\phi^2}{\mu(1-\gamma)^2}\right)^2, \tag{124}$$

where $\boldsymbol{w}^\star \in \arg\min_{\boldsymbol{w}} \ell\left(\boldsymbol{w}, Q^\xi, \tilde{d}_\xi\right)$.

The proof of Theorem 10 is postponed to Appendix B.2.

The following lemma provide a (very pessimistic) upper bound of C_ν in Assumption 4.

Lemma 7 (Upper bound of C_ν)**.** *If $\nu(s,a) > 0$ for all state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$, then we have*

$$C_\nu \leq \frac{1}{(1-\gamma)^2 \nu_{\min}^2}.$$

Proof. We only need to note that

$$\sqrt{\mathbb{E}_{(s.a)\sim\tilde{d}^{(t)}}\left[\left(\frac{h^{(t)}(s,a)}{\tilde{d}_n^{(t)}(s,a)}\right)^2\right]} \leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{h^{(t)}(s,a)}{\tilde{d}_n^{(t)}(s,a)} \leq \frac{1}{(1-\gamma)\nu_{\min}},$$

where the last inequality follows from (10). □

We give some key lemmas which will be used in our proof of Theorem 5.

Lemma 8 (consensus properties)**.** *For all $t \in \mathbb{N}$, we have*

$$\bar{\boldsymbol{\xi}}^{(t+1)} = \bar{\boldsymbol{\xi}}^{(t)} + \alpha \hat{\boldsymbol{w}}^{(t)}, \tag{125}$$

$$\frac{1}{N}\mathbf{1}^\top \boldsymbol{h}^{(t)} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{h}_n^{(t)} = \hat{\boldsymbol{w}}^{(t)}. \tag{126}$$

Proof. (126) could be obtained directly by using mathematical induction and update rule (40) (note that $\frac{1}{N}\mathbf{1}^\top \boldsymbol{h}^{(-1)} = \hat{\boldsymbol{w}}^{(-1)} = \mathbf{0}$, see line 2 of Algorithm 3), and (125) could be obtained by averaging both sides of (41) and using (126). □

Lemma 9 (Young's inequalities)**.** *Let $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$ be a set of m vectors in $\mathbb{R}^l$. Then for any $\zeta > 0$, we have*

$$\|\boldsymbol{x}_i + \boldsymbol{x}_j\|_2^2 \leq (1+\zeta)\|\boldsymbol{x}_i\|_2^2 + (1+1/\zeta)\|\boldsymbol{x}_j\|_2^2, \tag{127}$$

$$\left\|\sum_{i=1}^{m} \boldsymbol{x}_i\right\|_2^2 \leq m \sum_{i=1}^{m} \|\boldsymbol{x}_i\|_2^2. \tag{128}$$

Lemma 10 (Lipschitzness of Q-function with function approximation)**.** *Assume that $r(s,a) \in [0,1], \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. For any $\boldsymbol{\xi}$, $\boldsymbol{\xi}' \in \mathbb{R}^p$, we have*

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad |Q^{f_{\xi'}}(s,a) - Q^{f_\xi}(s,a)| \leq \underbrace{\frac{2C_\phi \gamma(1+\gamma)}{(1-\gamma)^2}}_{:=L_Q} \|\boldsymbol{\xi}' - \boldsymbol{\xi}\|_2. \tag{129}$$

Proof. See Appendix D.1. □

For each iteration step t in Algorithm 3, we let $\bar{\boldsymbol{\xi}}^{(t)} := \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\xi}_n^{(t)} = \frac{1}{N}\boldsymbol{\xi}^{(t)\top}\mathbf{1}_N$. We define

$$\Omega_1^{(t)} := \mathbb{E}\left\|\boldsymbol{\xi}^{(t)} - \mathbf{1}_N \bar{\boldsymbol{\xi}}^{(t)\top}\right\|_{\mathrm{F}}^2, \tag{130}$$

$$\Omega_2^{(t)} := \mathbb{E}\left\|\boldsymbol{h}^{(t)} - \mathbf{1}_N \hat{\boldsymbol{w}}^{(t)\top}\right\|_{\mathrm{F}}^2, \tag{131}$$

We let

$$\bar{\varepsilon}_{\text{stat}} := \frac{1}{N}\sum_{n=1}^{N} \varepsilon_{\text{stat}}^n, \tag{132}$$

$$\bar{\varepsilon}_{\text{approx}} := \frac{1}{N}\sum_{n=1}^{N} \varepsilon_{\text{approx}}^n, \tag{133}$$

and define $\delta^{(t)} := V^\star - \bar{V}^{(t)}(\rho)$, where $\bar{V}^{(t)}$ is shorthand for $V^{\bar{f}^{(t)}}$. We give the following performance improvement lemma.

Lemma 11 (Performance improvement of FedNAC)**.** *Fix a state distribution ρ, then we have*

$$\vartheta_\rho \delta^{(t+1)} + \frac{D_\star^{(t+1)}}{(1-\gamma)\alpha} \le \vartheta_\rho \delta^{(t)} + \frac{D_\star^{(t)}}{(1-\gamma)\alpha} - \delta^{(t)} + \frac{2\sqrt{C_\nu}(\vartheta_\rho+1)}{1-\gamma}\left(\sqrt{\bar{\varepsilon}_{stat}} + \sqrt{2\left(\bar{\varepsilon}_{approx} + \frac{L_Q^2}{N}\left\|\boldsymbol{\xi}^{(t)} - \mathbf{1}_N \bar{\boldsymbol{\xi}}^{(t)\top}\right\|_{\mathrm{F}}^2\right)}\right). \tag{134}$$

Proof. See Appendix D.2. □

Lemma 12 (linear system)**.** *For any $t \in \mathbb{N}$, we let $\boldsymbol{\Omega}^{(t)} = (\Omega_1^{(t)}, \Omega_2^{(t)})^\top$. Then for any $\zeta > 0$, we have*

$$\boldsymbol{\Omega}^{(t+1)} \le \boldsymbol{C}\boldsymbol{\Omega}^{(t)} + \boldsymbol{s}, \tag{135}$$

where

$$\boldsymbol{C} = (c_{ij}) = \begin{pmatrix} (1+\zeta)\sigma^2 & \alpha^2(1+1/\zeta)\sigma^2 \\ (1+1/\zeta)\frac{96\sigma^2 L_Q^2}{(1-\gamma)\mu} & \sigma^2\left(1+\zeta+(1+1/\zeta)\frac{24L_Q^2\alpha^2}{(1-\gamma)\mu}\right) \end{pmatrix}, \tag{136}$$

and

$$\boldsymbol{s} = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} 0 \\ (1+1/\zeta)\frac{6\sigma^2}{(1-\gamma)\mu}\left(N(\bar{\varepsilon}_{stat} + C_\nu \bar{\varepsilon}_{approx}) + 4L_Q^2\left(\frac{\alpha^2 N \bar{\varepsilon}_{stat}}{(1-\gamma)\mu} + \frac{\alpha^2 N C_\phi^2}{\mu^2(1-\gamma)^2}\right)\right) \end{pmatrix}. \tag{137}$$

Proof. See Appendix D.3. □

Now we are ready to give the formal version of Theorem 5 and its proof.

Theorem 11 (Convergence rate of FedNAC (formal))**.** *Let $\boldsymbol{\xi}_1^{(0)} = \cdots = \boldsymbol{\xi}_N^{(0)}$ in FedNAC (Algorithm 3), let the $\boldsymbol{w}^{(0)} = \mathbf{0}$ and the critic stepsize $\beta = \frac{1}{2C_\phi}$ in Algorithm 4. Then under Assumptions 1, 2, 3 and 4, when the actor stepsize satisfies*

$$\alpha \le \alpha_1 := \frac{(1-\sigma^2)^3\sqrt{(1-\gamma)\mu}}{768\sqrt{6}\sigma L_Q}, \tag{138}$$

where L_Q is defined in Lemma 10, we have

$$\begin{aligned} & V^\star(\rho) - \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\bar{V}^{(t)}(\rho)\right] \\ & \le \frac{D_\star^{(0)} + \alpha\vartheta_\rho}{T(1-\gamma)\alpha} + \frac{1}{T}\cdot\frac{512\sqrt{6}C_\phi\sqrt{C_\nu}(\vartheta_\rho+1)\sigma\alpha}{(1-\sigma^2)^{3/2}(1-\gamma)^3\sqrt{N}}\sqrt{\Omega_2^{(0)}} \\ & \quad + \left[\frac{2\sqrt{C_\nu}(\vartheta_\rho+1)}{1-\gamma} + \sqrt{1+\frac{64C_\phi^2\alpha^2}{(1-\gamma)^5\mu}}\cdot\frac{3072\sqrt{3}C_\phi\sqrt{C_\nu}(\vartheta_\rho+1)\sigma^2\alpha}{(1-\sigma^2)^3(1-\gamma)^{7/2}\sqrt{\mu}}\right] \\ & \quad \cdot\frac{2}{(1-\gamma)^2\sqrt{K}}\left((\sqrt{2p}+1)C_\phi^2 + \sqrt{2p}\mu(1-\gamma)\right) \\ & \quad + \left[\frac{2\sqrt{2C_\nu}(\vartheta_\rho+1)}{1-\gamma} + \frac{3072\sqrt{3}C_\phi C_\nu(\vartheta_\rho+1)\sigma^2\alpha}{(1-\sigma^2)^3(1-\gamma)^{7/2}\sqrt{\mu}}\right]\sqrt{\bar{\varepsilon}_{approx}} + \frac{6144\sqrt{2}\sigma^2 C_\nu(\vartheta_\rho+1)C_\phi^3\alpha^2}{(1-\gamma)^{13/2}\mu^{3/2}(1-\sigma^2)^3}. \end{aligned} \tag{139}$$

Moreover, the consensus errors could be upper bounded by

$$\mathbb{E}\left\|\boldsymbol{\xi}^{(t)} - \mathbf{1}_N\bar{\boldsymbol{\xi}}^{(t)\top}\right\|_{\mathrm{F}}^2 \le \left(\frac{49}{64}\sigma^2 + \frac{15}{64}\right)^t \mathbb{E}\left\|\boldsymbol{h}^{(0)} - \mathbf{1}_N\hat{\boldsymbol{w}}^{(0)\top}\right\|_{\mathrm{F}}^2 + \frac{64\delta(\alpha,K)}{15(1-\sigma^2)}, \tag{140}$$

where

$$\delta(\alpha,K) := \frac{18\sigma^2 N}{(1-\sigma^2)(1-\gamma)\mu}\left(\bar{\varepsilon}_{stat} + C_\nu\bar{\varepsilon}_{approx}\right) + \frac{72\sigma^2 L_Q^2 N}{(1-\gamma)^3\mu^3(1-\sigma^2)}\left((1-\gamma)\mu\bar{\varepsilon}_{stat} + C_\phi^2\right)\alpha^2, \tag{141}$$

and

$$\bar{\varepsilon}_{stat} \leq \frac{4}{(1-\gamma)^4 K}\left((\sqrt{2p}+1)C_\phi^2+\sqrt{2p}\mu(1-\gamma)\right)^2.$$

Remark 3 (Sample and communication complexity)**.** *When $\sigma > 0$ and*

$$\alpha = \frac{\sqrt{\mu}(D_\star^{(0)})^{1/3}}{6144^{1/3}2^{1/6}C_\nu^{1/3}(1+\vartheta_\rho)^{1/3}C_\phi}\cdot\frac{(1-\gamma)^{11/6}(1-\sigma^2)}{T^{1/3}\sigma^{2/3}},$$

it follows from Theorem 11 that

$$\begin{aligned}
&V^\star(\rho)-\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\bar{V}^{(t)}(\rho)\right]\\
&\leq \frac{3^{1/3}\cdot 2^{29/6}(D_\star^{(0)})^{2/3}C_\nu^{1/3}(1+\vartheta_\rho)^{1/3}C_\phi\sigma^{2/3}}{T^{2/3}(1-\gamma)^{17/6}(1-\sigma^2)\sqrt{\mu}}+\frac{\vartheta_\rho}{(1-\gamma)T}+\frac{2^{17/3}3^{1/6}C_\nu^{1/6}(1+\vartheta_\rho)^{2/3}\sigma^{1/3}\sqrt{\mu}(D_\star^{(0)})^{1/3}}{T^{4/3}(1-\sigma^2)^{1/2}(1-\gamma)^{7/6}\sqrt{N}}\\
&\quad+\left[\frac{2\sqrt{C_\nu}(\vartheta_\rho+1)}{1-\gamma}+\sqrt{1+\frac{(D_\star^{(0)})^{2/3}(1-\sigma^2)^2}{3^{3/2}\cdot 4C_\nu^{2/3}(1-\gamma)^{4/3}(1+\vartheta_\rho)^{1/3}T^{2/3}\sigma^{4/3}}}\cdot\frac{2^{37/6}\cdot 3^{7/6}C_\nu^{1/6}(\vartheta_\rho+1)^{2/3}\sigma^{4/3}(D_\star^{(0)})^{1/3}}{(1-\sigma^2)^2(1-\gamma)^{5/3}T^{1/3}}\right]\\
&\quad\cdot\frac{2}{(1-\gamma)^2\sqrt{K}}\left((\sqrt{2p}+1)C_\phi^2+\sqrt{2p}\mu(1-\gamma)\right)\\
&\quad+\left[\frac{2\sqrt{2C_\nu}(\vartheta_\rho+1)}{1-\gamma}+\frac{2^{37/6}\cdot 3^{7/6}C_\nu^{1/6}(\vartheta_\rho+1)^{2/3}\sigma^{4/3}(D_\star^{(0)})^{1/3}}{(1-\sigma^2)^2(1-\gamma)^{5/3}T^{1/3}}\right]\sqrt{\bar{\varepsilon}_{approx}}.
\end{aligned}\tag{142}$$

Consequently, we need

$$T \gtrsim \left\{\frac{\sigma}{\varepsilon^{3/2}(1-\gamma)^{17/4}(1-\sigma^2)^{3/2}},\frac{1}{\varepsilon(1-\gamma)},\frac{\sigma^{1/4}}{\varepsilon^{3/4}(1-\sigma^2)^{3/8}(1-\gamma)^{7/8}N^{3/8}},\frac{\sigma^4}{(1-\gamma)^2(1-\gamma^2)^6}\right\}$$

and

$$K=\mathcal{O}\left(\frac{1}{(1-\gamma)^6\varepsilon^2}\right)$$

such that $V^\star(\rho)-\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\bar{V}^{(t)}(\rho)\right]\lesssim\varepsilon+\frac{\bar{\varepsilon}_{approx}}{1-\gamma}$. In Algorithm 5, each trajectory has the expected length $1/(1-\gamma)$. Consider only the term where ε dominates, FedNAC requires $\mathcal{O}\left(\frac{1}{(1-\gamma)^{45/4}\varepsilon^{7/2}(1-\sigma^2)^{3/2}}\right)$ samples for each agent and $\mathcal{O}\left(\frac{1}{\varepsilon^{3/2}(1-\gamma)^{17/4}(1-\sigma^2)^{3/2}}\right)$ rounds of communication.

On the other end, when $\sigma=0$, (139) becomes:

$$\begin{aligned}
V^\star(\rho)-\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\bar{V}^{(t)}(\rho)\right]\leq\frac{D_\star^{(0)}+\alpha\vartheta_\rho}{T(1-\gamma)\alpha}+\frac{4\sqrt{C_\nu}(\vartheta_\rho+1)}{(1-\gamma)^3\sqrt{K}}\left((\sqrt{2p}+1)C_\phi^2+\sqrt{2p}\mu(1-\gamma)\right)\\
+\frac{2\sqrt{2C_\nu}(\vartheta_\rho+1)}{1-\gamma}\sqrt{\bar{\varepsilon}_{approx}},
\end{aligned}\tag{143}$$

Consequently, for any fixed $\alpha>0$, when $\sigma=0$ or close to 0, with $T=\mathcal{O}\left(\frac{1}{(1-\gamma)\varepsilon}\right)$ and $K=\mathcal{O}\left(\frac{1}{(1-\gamma)^6\varepsilon^2}\right)$, FedNAC requires $KT/(1-\gamma)=\mathcal{O}\left(\frac{1}{(1-\gamma)^8\varepsilon^3}\right)$ samples for each agent and $T=\mathcal{O}\left(\frac{1}{(1-\gamma)\varepsilon}\right)$ rounds of communication such that $V^\star(\rho)-\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\bar{V}^{(t)}(\rho)\right]\lesssim\varepsilon+\frac{\bar{\varepsilon}_{approx}}{1-\gamma}$.

B.1 Proof of Theorem 11

We suppose Assumptions 1, 5, 2, 3 and 4 holds. By Lemma 12 and nonnegativity of each entry of $\boldsymbol{C}$, $\boldsymbol{s}$ and $\boldsymbol{\Omega}^{(t)}$ where $t\in\mathbb{N}$, it's easy to see that

$$\sqrt{\boldsymbol{\Omega}^{(t+1)}}\leq\sqrt{\boldsymbol{C}}\sqrt{\boldsymbol{\Omega}^{(t)}}+\sqrt{\boldsymbol{s}},\tag{144}$$

where $\sqrt{\cdot}$ is exerted element-wise.

In addition, taking expectation on both sides of (134) and using the act that

$$\mathbb{E}\left[\sqrt{2\left(\bar{\varepsilon}_{\text{approx}}+\frac{L_Q^2}{N}\left\|\boldsymbol{\xi}^{(t)}-\mathbf{1}_N\bar{\boldsymbol{\xi}}^{(t)\top}\right\|_{\text{F}}^2\right)}\right]\leq\sqrt{2\bar{\varepsilon}_{\text{approx}}}+\sqrt{\frac{2L_Q^2}{N}\Omega_1^{(t)}},$$

we have

$$\begin{aligned}\vartheta_\rho\mathbb{E}[\delta^{(t+1)}]+\frac{\mathbb{E}[D_\star^{(t+1)}]}{(1-\gamma)\alpha}\leq\vartheta_\rho\mathbb{E}[\delta^{(t)}]+\frac{\mathbb{E}[D_\star^{(t)}]}{(1-\gamma)\alpha}-\mathbb{E}[\delta^{(t)}]\\ +\frac{2\sqrt{C_\nu}(\vartheta_\rho+1)}{1-\gamma}\left(\sqrt{\bar{\varepsilon}_{\text{stat}}}+\sqrt{2\bar{\varepsilon}_{\text{approx}}}+\sqrt{\frac{2L_Q^2}{N}\Omega_1^{(t)}}\right).\end{aligned}\tag{145}$$

We define the Lyapunov function $\Phi^{(t)}$ as follows:

$$\Phi^{(t)}\coloneqq\vartheta_\rho\mathbb{E}[\delta^{(t)}]+\frac{\mathbb{E}[D_\star^{(t)}]}{(1-\gamma)\alpha}+\boldsymbol{q}^\top\sqrt{\boldsymbol{\Omega}^{(t)}},\tag{146}$$

where

$$\boldsymbol{q}=\begin{pmatrix}q_1\\q_2\end{pmatrix}=\begin{pmatrix}\frac{2L_Q\sqrt{2C_\nu}(\vartheta_\rho+1)}{(1-\gamma)\sqrt{N}}\cdot\frac{1}{1-\sqrt{1+\zeta}\sigma-\sqrt{(1+1/\zeta)c_{21}}\sigma\alpha/(1-\sqrt{c_{22}})}\\ \frac{2L_Q\sqrt{2C_\nu}(\vartheta_\rho+1)}{(1-\gamma)\sqrt{N}}\cdot\frac{\sqrt{1+1/\zeta}\sigma\alpha}{(1-\sqrt{1+\zeta}\sigma)(1-\sqrt{c_{22}})-\sqrt{(1+1/\zeta)c_{21}}\sigma\alpha}\end{pmatrix}.\tag{147}$$

It's straightforward to verify that when $\zeta=\frac{1-\sigma^2}{2}$, we have the entries in $\boldsymbol{C}$ (cf. (136)) satisfies

$$c_{11}<\frac{1+\sigma^2}{2},\tag{148}$$

$$c_{12}\leq\frac{3\sigma^2\alpha^2}{1-\sigma^2}.\tag{149}$$

Moreover, from $\alpha\leq\frac{\sqrt{(1-\gamma)\mu}(1-\sigma^2)}{12\sqrt{2}\sigma L_Q}$ we deduce

$$c_{22}\leq\frac{3+\sigma^2}{4},\tag{150}$$

which gives

$$1-\sqrt{c_{22}}\geq1-\sqrt{\frac{3+\sigma^2}{4}}\geq\frac{1-\sigma^2}{8},\tag{151}$$

Also note that $\alpha\leq\frac{(1-\sigma^2)^3\sqrt{(1-\gamma)\mu}}{768\sqrt{6}\sigma^2L_Q}$ yields

$$\sqrt{(1+1/\zeta)c_{21}}\sigma\alpha\leq\frac{(1-\sqrt{1+\zeta}\sigma)(1-\sqrt{c_{22}})}{2}.$$

which together with (151) and the fact $1-\sqrt{1+\zeta}\sigma\geq\frac{1-\sigma^2}{4}$ indicates $q_1,q_2>0$ and that

$$q_1\leq\frac{16\sqrt{2}L_Q\sqrt{C_\nu}(\vartheta_\rho+1)}{(1-\sigma^2)(1-\gamma)\sqrt{N}},\tag{152}$$

$$q_2\leq\frac{128\sqrt{6}L_Q\sqrt{C_\nu}(\vartheta_\rho+1)\sigma\alpha}{(1-\sigma^2)^{5/2}(1-\gamma)\sqrt{N}}.\tag{153}$$

Thus by (144) and (145) we have

$$\Phi^{(t+1)}=\vartheta_\rho\mathbb{E}[\delta^{(t+1)}]+\frac{\mathbb{E}[D_\star^{(t+1)}]}{(1-\gamma)\alpha}+\boldsymbol{q}^\top\sqrt{\boldsymbol{\Omega}^{(t+1)}}$$

$$
\begin{aligned}
&\leq \vartheta_\rho \mathbb{E}[\delta^{(t)}] + \frac{\mathbb{E}[D_\star^{(t)}]}{(1-\gamma)\alpha} - \mathbb{E}[\delta^{(t)}] + \boldsymbol{q}^\top \left(\sqrt{\boldsymbol{C}}\sqrt{\boldsymbol{\Omega}^{(t)}} + \sqrt{\boldsymbol{s}}\right) \\
&\quad + \frac{2\sqrt{C_\nu}(\vartheta_\rho+1)}{1-\gamma}\left(\sqrt{\bar{\varepsilon}_{\text{stat}}} + \sqrt{2\bar{\varepsilon}_{\text{approx}}} + \sqrt{\frac{2L_Q^2}{N}\Omega_1^{(t)}}\right) \\
&= \Phi^{(t)} + \left(\underbrace{\boldsymbol{q}^\top(\sqrt{\boldsymbol{C}} - \boldsymbol{I}) + \left(\frac{2L_Q\sqrt{2C_\nu}(\vartheta_\rho+1)}{(1-\gamma)\sqrt{N}}, 0\right)}_{=(0,0)}\right)\sqrt{\boldsymbol{\Omega}^{(t)}} \\
&\quad + \frac{2\sqrt{C_\nu}(\vartheta_\rho+1)}{1-\gamma}\left(\sqrt{\bar{\varepsilon}_{\text{stat}}} + \sqrt{2\bar{\varepsilon}_{\text{approx}}}\right) + q_2\sqrt{s_2} - \mathbb{E}[\delta^{(t)}],
\end{aligned} \tag{154}
$$

which gives

$$
\mathbb{E}[\delta^{(t)}] \leq \Phi^{(t)} - \Phi^{(t+1)} + \frac{2\sqrt{C_\nu}(\vartheta_\rho+1)}{1-\gamma}\left(\sqrt{\bar{\varepsilon}_{\text{stat}}} + \sqrt{2\bar{\varepsilon}_{\text{approx}}}\right) + q_2\sqrt{s_2}. \tag{155}
$$

Summing the above inequality over $t = 0, 1, \cdots, T-1$ and divide both sides by T, we have

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\delta^{(t)}] \leq \frac{\Phi^{(0)} - \Phi^{(t)}}{T} + \frac{2\sqrt{C_\nu}(\vartheta_\rho+1)}{1-\gamma}\left(\sqrt{\bar{\varepsilon}_{\text{stat}}} + \sqrt{2\bar{\varepsilon}_{\text{approx}}}\right) + q_2\sqrt{s_2}. \tag{156}
$$

Since

$$
s_2 \leq \frac{18\sigma^2 N}{(1-\sigma^2)(1-\gamma)\mu}\left(\bar{\varepsilon}_{\text{stat}} + C_\nu\bar{\varepsilon}_{\text{approx}}\right) + \frac{72\sigma^2 L_Q^2 N}{(1-\gamma)^3\mu^3(1-\sigma^2)}\left((1-\gamma)\mu\bar{\varepsilon}_{\text{stat}} + C_\phi^2\right)\alpha^2, \tag{157}
$$

and

$$
\Phi^{(0)} - \Phi^{(t)} \leq \Phi^{(0)} \leq \frac{\vartheta_\rho}{1-\gamma} + \frac{\mathbb{E}[D_\star^{(0)}]}{(1-\gamma)\alpha} + \frac{16\sqrt{2}L_Q\sqrt{C_\nu}(\vartheta_\rho+1)}{(1-\sigma^2)(1-\gamma)\sqrt{N}}\left(\sqrt{\Omega_1^{(0)}} + \frac{8\sqrt{3}\sigma\alpha}{\sqrt{1-\sigma^2}}\sqrt{\Omega_2^{(0)}}\right), \tag{158}
$$

we have (recall that $L_Q = \frac{2C_\phi\gamma(1+\gamma)}{(1-\gamma)^2} \leq \frac{4C_\phi}{(1-\gamma)^2}$)

$$
\begin{aligned}
&V^\star(\rho) - \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\bar{V}^{(t)}(\rho)\right] \\
&\leq \frac{D_\star^{(0)} + \alpha\vartheta_\rho}{T(1-\gamma)\alpha} + \frac{1}{T}\cdot\frac{64\sqrt{2}C_\phi\sqrt{C_\nu}(\vartheta_\rho+1)}{(1-\sigma^2)(1-\gamma)^3\sqrt{N}}\left(\sqrt{\Omega_1^{(0)}} + \frac{8\sqrt{3}\sigma\alpha}{\sqrt{1-\sigma^2}}\sqrt{\Omega_2^{(0)}}\right) \\
&\quad + \left[\frac{2\sqrt{C_\nu}(\vartheta_\rho+1)}{1-\gamma} + \sqrt{\frac{18\sigma^2 N}{(1-\sigma^2)(1-\gamma)\mu} + \frac{1152\sigma^2 C_\phi^2 N\alpha^2}{(1-\gamma)^6\mu^2(1-\sigma^2)}}\cdot\frac{512\sqrt{6}C_\phi\sqrt{C_\nu}(\vartheta_\rho+1)\sigma\alpha}{(1-\sigma^2)^{5/2}(1-\gamma)^3\sqrt{N}}\right]\sqrt{\bar{\varepsilon}_{\text{stat}}} \\
&\quad + \left[\frac{2\sqrt{2C_\nu}(\vartheta_\rho+1)}{1-\gamma} + \sqrt{\frac{18\sigma^2 N C_\nu}{(1-\sigma^2)(1-\gamma)\mu}}\cdot\frac{512\sqrt{6}C_\phi\sqrt{C_\nu}(\vartheta_\rho+1)\sigma\alpha}{(1-\sigma^2)^{5/2}(1-\gamma)^3\sqrt{N}}\right]\sqrt{\bar{\varepsilon}_{\text{approx}}} \\
&\quad + \frac{6144\sqrt{2}\sigma^2\sqrt{C_\nu}(\vartheta_\rho+1)C_\phi^3\alpha^2}{(1-\gamma)^{13/2}\mu^{3/2}(1-\sigma^2)^3}.
\end{aligned} \tag{159}
$$

By Theorem 10 we know that $\sqrt{\bar{\varepsilon}_{\text{stat}}}$ could be upper bounded as follows:

$$
\sqrt{\bar{\varepsilon}_{stat}} \leq \frac{2}{(1-\gamma)^2\sqrt{K}}\left((\sqrt{2p}+1)C_\phi^2 + \sqrt{2p}\mu(1-\gamma)\right). \tag{160}
$$

(139) follows from plugging (160) into (159) and noting that when $\boldsymbol{\xi}_1^{(0)} = \cdots = \boldsymbol{\xi}_N^{(0)}$, $\Omega_1^{(0)} = 0$.

Bounding the consensus errors. Similar to Step 4 in Appendix C.4, to bound the consensus error $\left\|\log f_n^{(t)} - \log \bar{f}^{(t)}\right\|_\infty$ for all $n \in [N]$, we first upper bound the eigenvalue of $\rho(\boldsymbol{C})$—the spectral norm of $\boldsymbol{C}$.

The characteristic polynomial of $\boldsymbol{C}$ is

$$\begin{aligned} f(\lambda) &= (\lambda - c_{11})(\lambda - c_{22}) - c_{12}c_{21} \\ &= \lambda^2 - (c_{11} + c_{22})\lambda + c_{11}c_{22} - c_{12}c_{21}\,, \end{aligned}$$

which gives

$$\begin{aligned} \rho(\boldsymbol{C}) &\le \frac{c_{11} + c_{22} + \sqrt{(c_{11} + c_{22})^2 - 4(c_{11}c_{12} - c_{12}c_{21})}}{2} \\ &= \frac{c_{11} + c_{22} + \sqrt{(c_{22} - c_{11})^2 + 4c_{12}c_{21}}}{2} \\ &\le \frac{c_{11} + c_{22} + c_{22} - c_{11} + 2\sqrt{c_{12}c_{21}}}{2} \\ &= c_{22} + \sqrt{c_{12}c_{21}} \\ &\le \frac{3 + \sigma^2}{4} + \frac{\sqrt{3}\sigma\alpha}{\sqrt{1 - \sigma^2}} \cdot \frac{12\sqrt{2}L_Q\sigma}{\sqrt{1 - \sigma^2}(1 - \gamma)\mu} \\ &\le \frac{3 + \sigma^2}{4} + \frac{\sigma(1 - \sigma^2)^2}{64} \\ &\le \frac{49 + 15\sigma^2}{64} < 1\,, \end{aligned} \tag{161}$$

where the third inequality uses (149), (150), and the fourth inequality uses (138).

Therefore, similar to (241), when $\alpha \le \alpha_1$, we have

$$\left\|\boldsymbol{\Omega}^{(t)}\right\|_2 \le \left(\frac{49}{64}\sigma + \frac{15}{64}\right)^t \left\|\boldsymbol{\Omega}^{(0)}\right\|_2 + \frac{64s_2}{15(1 - \sigma^2)}\,. \tag{162}$$

Combining the above inequality with (157), and (160), we obtain (140).

B.2 Proof of Theorem 10

The proof of Theorem 10 could be found in Appendix C.5 in Yuan et al. (2022). We present it for completeness. To prove Theorem 10, we need the following Theorem 12.

Theorem 12 (Theorem 1 in Bach and Moulines (2013))**.** *Consider the following assumptions:*

(i) The observations $(\boldsymbol{a}_k, \boldsymbol{b}_k) \in \mathbb{R}^p \times \mathbb{R}^p$ are independent and identically distributed.

(ii) $\mathbb{E}\left[\|\boldsymbol{a}_k\|^2\right]$[2] and $\mathbb{E}\left[\|\boldsymbol{b}_k\|^2\right]$ are finite. The covariance $\mathbb{E}\left[\boldsymbol{a}_k\boldsymbol{a}_k^\top\right]$ is invertible.

(iii) The global minimum of $g(w) = \frac{1}{2}\mathbb{E}\left[\langle \boldsymbol{w}, \boldsymbol{a}_k\rangle^2 - 2\langle \boldsymbol{w}, \boldsymbol{b}_k\rangle\right]$ is attained at a certain $\boldsymbol{w}^\star \in \mathbb{R}^p$. Let $\Delta_k = \boldsymbol{b}_k - \langle \boldsymbol{w}^\star, \boldsymbol{a}_k\rangle \boldsymbol{a}_k$ denote the residual. We have $\mathbb{E}[\Delta_k] = 0$.

(iv) $\exists R > 0$ and $\sigma > 0$ such that $\mathbb{E}\left[\Delta_k\Delta_k^\top\right] \le \sigma^2\mathbb{E}\left[\boldsymbol{a}_k\boldsymbol{a}_k^\top\right]$ and $\mathbb{E}\left[\|\boldsymbol{a}_k\|^2\,\boldsymbol{a}_k\boldsymbol{a}_k^\top\right] \le R^2\mathbb{E}\left[\boldsymbol{a}_k\boldsymbol{a}_k^\top\right]$.

Consider the stochastic gradient recursion

$$w_{k+1} = w_k - \eta\left(\langle w_k, a_k\rangle a_k - b_k\right)$$

started from $w_0 \in \mathbb{R}^p$. Let $w_{out} = \frac{1}{K}\sum_{k=1}^K w_k$. When $\eta = \frac{1}{4R^2}$, we have

$$\mathbb{E}\left[g(w_{out}) - g(w^\star)\right] \le \frac{2}{K}(\sigma\sqrt{p} + R\,\|w_0 - w^\star\|)^2. \tag{163}$$

[2] Here $\|\cdot\|$ could be any norm in $\mathbb{R}^p$.

In the proof of Theorem 10 we'll show that for Algorithm 4, the assumptions in Theorem 12 are all satisfied and thus we can use the result (163).

Proof of Theorem 10. We let a_k and b_k in Theorem 12 be $\phi(s,a)$ and $\widehat{Q}_\xi\phi(s,a)$ in Algorithm 4, respectively. And we let $\|\cdot\| = \|\cdot\|_2$ in Theorem 12. Since the observations $\left(\phi(s,a), \widehat{Q}_\xi(s,a)\phi(s,a)\right) \in \mathbb{R}^p \times \mathbb{R}^p$ are i.i.d., (i) is satisfied.

As we assume $\|\phi(s,a)\|_2 \le C_\phi$, $\mathbb{E}\left[\|\phi(s,a)\|_2^2\right]$ is finite. From Assumption 2 we know that $\mathbb{E}\left[\phi(s,a)\phi(s,a)^\top\right]$ is invertible.

Let H be the length of trajectory for estimating $\widehat{Q}_\xi(s,a)$. Then $\left(\widehat{Q}_\xi(s,a)\right)^2$ is bounded by

$$\begin{aligned}
\mathbb{E}\left[\left(\widehat{Q}_\xi(s,a)\right)^2\right] &= \mathbb{E}_{(s,a)\sim\tilde{d}_\nu^{\pi_\xi}}\left[\sum_{\tau=0}^{\infty} Pr(H=\tau)\mathbb{E}\left[\left(\sum_{t=0}^{\tau} r(s_t,a_t)\right)^2 \middle| H=\tau, s_0=s, a_0=a\right]\right] \\
&= \mathbb{E}_{(s,a)\sim\tilde{d}_\nu^{\pi_\xi}}\left[(1-\gamma)\sum_{\tau=0}^{\infty}\gamma^\tau\mathbb{E}\left[\left(\sum_{t=0}^{\tau} r(s_t,a_t)\right)^2 \middle| H=\tau, s_0=s, a_0=a\right]\right] \\
&\le \mathbb{E}_{(s,a)\sim\tilde{d}_\nu^{\pi_\xi}}\left[(1-\gamma)\sum_{\tau=0}^{\infty}\gamma^\tau(\tau+1)^2\right] \le \frac{2}{(1-\gamma)^2}\,, \qquad (164)
\end{aligned}$$

from which we deduce $\mathbb{E}\left[\left\|\widehat{Q}_\xi(s,a)\phi(s,a)\right\|_2^2\right] \le C_\phi^2\mathbb{E}\left[\widehat{Q}_\xi(s,a)^2\right]$ is bounded. Thus (ii) holds.

Furthermore, we introduce the residual

$$\Delta := \left(\widehat{Q}_\xi(s,a) - \phi(s,a)^\top w^\star\right)\phi(s,a)\,, \qquad (165)$$

then from Lemma 7 in Yuan et al. (2022) we know that $\mathbb{E}[\Delta] = \frac{1}{2}\nabla_w \ell(w, \widehat{Q}_\xi, d_\nu^{\pi_\xi}) = 0$, which gives (iii).

To verify (iv), we let $R = C_\phi$ in Theorem 12, then $\mathbb{E}\left[\|\phi(s,a)\|_2^2\,\phi(s,a)\phi(s,a)^\top\right] \le C_\phi^2\mathbb{E}\left[\phi(s,a)\phi(s,a)^\top\right]$. Also note that

$$\begin{aligned}
w^\star &= \left(\mathbb{E}_{(s,a)\sim\tilde{d}_\nu^{\pi_\xi}}\left[\phi(s,a)\phi(s,a)^\top\right]\right)^\dagger \mathbb{E}_{(s,a)\sim\tilde{d}_\nu^{\pi_\xi}}\left[\widehat{Q}_\xi(s,a)\phi(s,a)\right] \\
&\le \frac{1}{1-\gamma}\left(\mathbb{E}_{(s,a)\sim\nu}\left[\phi(s,a)\phi(s,a)^\top\right]\right)^\dagger \mathbb{E}_{(s,a)\sim\tilde{d}_\nu^{\pi_\xi}}\left[\widehat{Q}_\xi(s,a)\phi(s,a)\right]\,, \qquad (166)
\end{aligned}$$

from which we deduce

$$\|w^\star\|_2 \le \frac{B}{\mu(1-\gamma)^2}\,. \qquad (167)$$

$$\begin{aligned}
\mathbb{E}\left[\left(\widehat{Q}_\xi(s,a) - \phi(s,a)^\top w^\star\right)^2 |s,a\right] &= \mathbb{E}\left[\left(\widehat{Q}_\xi(s,a)\right)^2 |s,a\right] - 2Q_\xi(s,a)\phi(s,a)^\top w^\star + (\phi(s,a)^\top w^\star)^2 \qquad (168) \\
&\le \frac{2}{(1-\gamma)^2} + \frac{2C_\phi^2}{\mu(1-\gamma)^3} + \frac{C_\phi^4}{\mu^2(1-\gamma)^4} \\
&\le \frac{2}{(1-\gamma)^2}\left(\frac{C_\phi^2}{\mu(1-\gamma)} + 1\right)^2. \qquad (169)
\end{aligned}$$

The above expression implies

$$\begin{aligned}
\mathbb{E}\left[\Delta\Delta^\top\right] &= \mathbb{E}_{(s,a)\sim\tilde{d}_\nu^{\pi_\xi}}\left[\left(\widehat{Q}_\xi(s,a) - \phi(s,a)^\top w^\star\right)^2 \phi(s,a)\phi(s,a)^\top |s,a\right] \\
&= \mathbb{E}_{(s,a)\sim\tilde{d}_\nu^{\pi_\xi}}\left[\mathbb{E}\left[\left(\widehat{Q}_\xi(s,a) - \phi(s,a)^\top w^\star\right)^2 |s,a\right]\phi(s,a)\phi(s,a)^\top\right]
\end{aligned}$$

$$\leq \left(\underbrace{\frac{\sqrt{2}}{1-\gamma}\left(\frac{C_\phi^2}{\mu(1-\gamma)}+1\right)}_{\sigma} \right) \mathbb{E}[\phi(s,a)\phi(s,a)^\top]\,. \tag{170}$$

Therefore, (iv) is verified.

Thus by (163), with stepsize $\beta = \frac{1}{2C_\phi^2}$, initialization $w_0 = 0$ and K steps of critic updates, we have

$$\begin{aligned}\mathbb{E}\left[\ell\left(w_{\text{out}}, \widehat{Q}_\xi, \tilde{d}_\xi\right)\right] - \ell\left(w^\star, \widehat{Q}_\xi, \tilde{d}_\xi\right) &\leq \frac{4}{K}\left(\sigma\sqrt{p} + C_\phi \left\|w^\star\right\|_2\right)^2 \\ &\leq \frac{4}{K}\left(\frac{\sqrt{2p}}{1-\gamma}\left(\frac{C_\phi^2}{\mu(1-\gamma)}+1\right) + \frac{C_\phi^2}{\mu(1-\gamma)^2}\right)^2,\end{aligned}$$

which gives (124). □

C Proof of key lemmas for FedNPG

C.1 Proof of Lemma 1

Before proceeding, we summarize several useful properties of the auxiliary sequences (cf. (51) and (52)), whose proof is postponed to Appendix C.6.

Lemma 13 (Properties of auxiliary sequences $\{\overline{\xi}^{(t)}\}$ and $\{\boldsymbol{\xi}^{(t)}\}$)**.** *$\{\overline{\xi}^{(t)}\}$ and $\{\boldsymbol{\xi}^{(t)}\}$ have the following properties:*

1. *$\boldsymbol{\xi}^{(t)}$ can be viewed as an unnormalized version of $\boldsymbol{\pi}^{(t)}$, i.e.,*

$$\pi_n^{(t)}(\cdot|s) = \frac{\xi_n^{(t)}(s,\cdot)}{\left\|\xi_n^{(t)}(s,\cdot)\right\|_1}\,,\ \forall n \in [N],\ s \in \mathcal{S}\,. \tag{171}$$

2. *For any $t \geq 0$, $\log \overline{\xi}^{(t)}$ keeps track of the average of $\log \boldsymbol{\xi}^{(t)}$, i.e.,*

$$\frac{1}{N}\mathbf{1}_N^\top \log \boldsymbol{\xi}^{(t)} = \log \overline{\xi}^{(t)}\,. \tag{172}$$

It follows that

$$\forall s \in \mathcal{S},\ t \geq 0: \quad \overline{\pi}^{(t)}(\cdot|s) = \frac{\overline{\xi}^{(t)}(s,\cdot)}{\left\|\overline{\xi}^{(t)}(s,\cdot)\right\|_1}. \tag{173}$$

Lemma 14 ((Cen et al., 2022a, Appendix. A.2))**.** *For any vector $\theta = [\theta_a]_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$, we denote by $\pi_\theta \in \mathbb{R}^{|\mathcal{A}|}$ the softmax transform of θ such that*

$$\pi_\theta(a) = \frac{\exp(\theta_a)}{\sum_{a'\in\mathcal{A}}\exp(\theta_{a'})}\,, \quad a \in \mathcal{A}\,. \tag{174}$$

For any $\theta_1, \theta_2 \in \mathbb{R}^{|\mathcal{A}|}$, we have

$$\left|\log(\|\exp(\theta_1)\|_1) - \log(\|\exp(\theta_2)\|_1)\right| \leq \|\theta_1 - \theta_2\|_\infty\,, \tag{175}$$

$$\|\log \pi_{\theta_1} - \log \pi_{\theta_2}\|_\infty \leq 2\,\|\theta_1 - \theta_2\|_\infty\,. \tag{176}$$

Step 1: bound $u^{(t+1)}(s,a) = \big\| \log \boldsymbol{\xi}^{(t+1)}(s,a) - \log \overline{\xi}^{(t+1)}(s,a)\mathbf{1}_N \big\|_2$**.** By (51b) and (52b) we have

$$\begin{aligned}
u^{(t+1)}(s,a) &= \big\| \log \boldsymbol{\xi}^{(t+1)}(s,a) - \log \overline{\xi}^{(t+1)}(s,a)\mathbf{1}_N \big\|_2 \\
&= \Big\| \alpha\Big(\boldsymbol{W} \log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\xi}^{(t)}(s,a)\mathbf{1}_N \Big) + (1-\alpha)\Big(\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a) - \widehat{Q}_\tau^{(t)}(s,a)\mathbf{1}_N \Big)/\tau \Big\|_2 \\
&\le \sigma\alpha \big\| \log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\xi}^{(t)}(s,a)\mathbf{1}_N \big\|_2 + \frac{1-\alpha}{\tau}\sigma \big\| \boldsymbol{T}^{(t)}(s,a) - \widehat{Q}_\tau^{(t)}(s,a)\mathbf{1}_N \big\|_2 \\
&\le \sigma\alpha \big\| u^{(t)} \big\|_\infty + \frac{1-\alpha}{\tau}\sigma \big\| v^{(t)} \big\|_\infty,
\end{aligned} \tag{177}$$

where the penultimate step results from the averaging property of $\boldsymbol{W}$ (property (21)). Taking maximum over $(s,a) \in \mathcal{S} \times \mathcal{A}$ establishes the bound on $\Omega_1^{(t+1)}$ in (60).

Step 2: bound $v^{(t+1)}(s,a) = \big\| \boldsymbol{T}^{(t+1)}(s,a) - \widehat{Q}_\tau^{(t+1)}(s,a)\mathbf{1}_N \big\|_2$**.** By (28) we have

$$\begin{aligned}
&\big\| \boldsymbol{T}^{(t+1)}(s,a) - \widehat{Q}_\tau^{(t+1)}(s,a)\mathbf{1}_N \big\|_2 \\
&= \Big\| \boldsymbol{W}\left(\boldsymbol{T}^{(t)}(s,a) + \boldsymbol{Q}_\tau^{(t+1)}(s,a) - \boldsymbol{Q}_\tau^{(t)}(s,a) \right) - \widehat{Q}_\tau^{(t+1)}(s,a)\mathbf{1}_N \Big\|_2 \\
&= \Big\| \Big(\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a) - \widehat{Q}_\tau^{(t)}(s,a)\mathbf{1}_N \Big) + \boldsymbol{W}\left(\boldsymbol{Q}_\tau^{(t+1)}(s,a) - \boldsymbol{Q}_\tau^{(t)}(s,a) \right) + \left(\widehat{Q}_\tau^{(t)}(s,a) - \widehat{Q}_\tau^{(t+1)}(s,a) \right)\mathbf{1}_N \Big\|_2 \\
&\le \sigma \big\| \boldsymbol{T}^{(t)}(s,a) - \widehat{Q}_\tau^{(t)}(s,a)\mathbf{1}_N \big\|_2 + \sigma \Big\| \Big(\boldsymbol{Q}_\tau^{(t+1)}(s,a) - \boldsymbol{Q}_\tau^{(t)}(s,a) \Big) + \left(\widehat{Q}_\tau^{(t)}(s,a) - \widehat{Q}_\tau^{(t+1)}(s,a) \right)\mathbf{1}_N \Big\|_2 \\
&\le \sigma \big\| \boldsymbol{T}^{(t)}(s,a) - \widehat{Q}_\tau^{(t)}(s,a)\mathbf{1}_N \big\|_2 + \sigma \big\| \boldsymbol{Q}_\tau^{(t+1)}(s,a) - \boldsymbol{Q}_\tau^{(t)}(s,a) \big\|_2,
\end{aligned} \tag{178}$$

where the penultimate step uses property (21), and the last step is due to

$$\begin{aligned}
&\Big\| \Big(\boldsymbol{Q}_\tau^{(t+1)}(s,a) - \boldsymbol{Q}_\tau^{(t)}(s,a) \Big) + \left(\widehat{Q}_\tau^{(t)}(s,a) - \widehat{Q}_\tau^{(t+1)}(s,a) \right)\mathbf{1}_N \Big\|_2^2 \\
&= \big\| \boldsymbol{Q}_\tau^{(t+1)}(s,a) - \boldsymbol{Q}_\tau^{(t)}(s,a) \big\|_2^2 + N\big(\widehat{Q}_\tau^{(t)}(s,a) - \widehat{Q}_\tau^{(t+1)}(s,a) \big)^2 \\
&\qquad - 2\sum_{n=1}^N \left(Q_{\tau,n}^{\pi_n^{(t+1)}}(s,a) - Q_{\tau,n}^{\pi_n^{(t)}}(s,a) \right)\left(\widehat{Q}_\tau^{(t+1)}(s,a) - \widehat{Q}_\tau^{(t)}(s,a) \right) \\
&= \big\| \boldsymbol{Q}_\tau^{(t+1)}(s,a) - \boldsymbol{Q}_\tau^{(t)}(s,a) \big\|_2^2 - N\big(\widehat{Q}_\tau^{(t)}(s,a) - \widehat{Q}_\tau^{(t+1)}(s,a) \big)^2 \\
&\le \big\| \boldsymbol{Q}_\tau^{(t+1)}(s,a) - \boldsymbol{Q}_\tau^{(t)}(s,a) \big\|_2^2.
\end{aligned}$$

Step 3: bound $\big\| Q_\tau^\star - \tau \log \overline{\xi}^{(t+1)} \big\|_\infty$**.** We decompose the term of interest as

$$\begin{aligned}
Q_\tau^\star - \tau \log \overline{\xi}^{(t+1)} &= Q_\tau^\star - \tau\alpha \log \overline{\xi}^{(t)} - (1-\alpha)\widehat{Q}_\tau^{(t)} \\
&= \alpha(Q_\tau^\star - \tau \log \overline{\xi}^{(t)}) + (1-\alpha)(Q_\tau^\star - \overline{Q}_\tau^{(t)}) + (1-\alpha)(\overline{Q}_\tau^{(t)} - \widehat{Q}_\tau^{(t)}),
\end{aligned}$$

which gives

$$\big\| Q_\tau^\star - \tau \log \overline{\xi}^{(t+1)} \big\|_\infty \le \alpha \big\| Q_\tau^\star - \tau \log \overline{\xi}^{(t)} \big\|_\infty + (1-\alpha)\big\| Q_\tau^\star - \overline{Q}_\tau^{(t)} \big\|_\infty + (1-\alpha)\big\| \overline{Q}_\tau^{(t)} - \widehat{Q}_\tau^{(t)} \big\|_\infty. \tag{179}$$

Note that we can upper bound $\big\| \overline{Q}_\tau^{(t)} - \widehat{Q}_\tau^{(t)} \big\|_\infty$ by

$$\begin{aligned}
\big\| \overline{Q}_\tau^{(t)} - \widehat{Q}_\tau^{(t)} \big\|_\infty &= \left\| \frac{1}{N}\sum_{n=1}^N Q_{\tau,n}^{\pi_n^{(t)}} - \frac{1}{N}\sum_{n=1}^N Q_{\tau,n}^{\overline{\pi}^{(t)}} \right\|_\infty \\
&\le \frac{1}{N}\sum_{n=1}^N \big\| Q_{\tau,n}^{\pi_n^{(t)}} - Q_{\tau,n}^{\overline{\pi}^{(t)}} \big\|_\infty
\end{aligned}$$

$$\leq \frac{M}{N}\sum_{n=1}^{N}\left\|\log\xi_n^{(t)}-\log\overline{\xi}^{(t)}\right\|_\infty \leqslant M\left\|u^{(t)}\right\|_\infty. \tag{180}$$

The last step is due to $\left|\log\xi_n^{(t)}(s,a)-\log\overline{\xi}^{(t)}(s,a)\right| \leq u^{(t)}(s,a)$, while the penultimate step results from writing

$$\overline{\pi}^{(t)}(\cdot|s)=\text{softmax}\left(\log\overline{\xi}^{(t)}(s,\cdot)\right),$$
$$\pi_n^{(t)}(\cdot|s)=\text{softmax}\left(\log\xi_n^{(t)}(s,\cdot)\right),$$

and applying the following lemma.

Lemma 15 (Lipschitz constant of soft Q-function)**.** *Assume that $r(s,a)\in[0,1], \forall(s,a)\in\mathcal{S}\times\mathcal{A}$ and $\tau\geq 0$. For any θ, $\theta'\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we have*

$$\|Q_\tau^{\pi_{\theta'}}-Q_\tau^{\pi_\theta}\|_\infty \leq \underbrace{\frac{1+\gamma+2\tau(1-\gamma)\log|\mathcal{A}|}{(1-\gamma)^2}\cdot\gamma}_{=:M}\|\theta'-\theta\|_\infty. \tag{181}$$

Plugging (180) into (179) gives

$$\left\|Q_\tau^\star-\tau\log\overline{\xi}^{(t+1)}\right\|_\infty \leq \alpha\left\|Q_\tau^\star-\tau\log\overline{\xi}^{(t)}\right\|_\infty+(1-\alpha)\left\|Q_\tau^\star-\overline{Q}_\tau^{(t)}\right\|_\infty+(1-\alpha)M\left\|u^{(t)}\right\|_\infty. \tag{182}$$

Step 4: bound $\left\|\boldsymbol{Q}_\tau^{(t+1)}(s,a)-\boldsymbol{Q}_\tau^{(t)}(s,a)\right\|_2$. Let $w^{(t)}:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ be defined as

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}: \quad w^{(t)}(s,a):=\left\|\log\boldsymbol{\xi}^{(t+1)}(s,a)-\log\boldsymbol{\xi}^{(t)}(s,a)-(1-\alpha)V_\tau^\star(s)\mathbf{1}_N/\tau\right\|_2. \tag{183}$$

Again, we treat $w^{(t)}$ as vectors in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ whenever it is clear from context. For any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $n\in[N]$, by Lemma 15 it follows that

$$\begin{aligned}\left|Q_{\tau,n}^{\pi_n^{(t+1)}}(s,a)-Q_{\tau,n}^{\pi_n^{(t)}}(s,a)\right| &\leq M\max_{s\in\mathcal{S}}\left\|\log\xi_n^{(t+1)}(s,\cdot)-\log\xi_n^{(t)}(s,\cdot)-(1-\alpha)V_\tau^\star(s)\mathbf{1}_{|\mathcal{A}|}/\tau\right\|_\infty\\ &\leq M\max_{s\in\mathcal{S}}\max_{a\in\mathcal{A}}w^{(t)}(s,a)\leq M\left\|w^{(t)}\right\|_\infty,\end{aligned} \tag{184}$$

and consequently

$$\left\|\boldsymbol{Q}_\tau^{(t+1)}(s,a)-\boldsymbol{Q}_\tau^{(t)}(s,a)\right\|_2 \leq M\sqrt{N}\left\|w^{(t)}\right\|_\infty. \tag{185}$$

It boils down to control $\left\|w^{(t)}\right\|_\infty$. To do so, we first note that for each $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned}&w^{(t)}(s,a)\\ &=\left\|\boldsymbol{W}\left(\alpha\log\boldsymbol{\xi}^{(t)}(s,a)+(1-\alpha)\boldsymbol{T}^{(t)}(s,a)/\tau\right)-\log\boldsymbol{\xi}^{(t)}(s,a)-(1-\alpha)V_\tau^\star(s)\mathbf{1}_N/\tau\right\|_2\\ &\overset{(a)}{=}\left\|\alpha(\boldsymbol{W}-\boldsymbol{I}_N)\left(\log\boldsymbol{\xi}^{(t)}(s,a)-\log\overline{\xi}^{(t)}(s,a)\mathbf{1}_N\right)+(1-\alpha)\left(\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a)/\tau-\log\boldsymbol{\xi}^{(t)}(s,a)-V_\tau^\star(s)\mathbf{1}_N/\tau\right)\right\|_2\\ &\overset{(b)}{\leq}2\alpha\left\|\log\boldsymbol{\xi}^{(t)}(s,a)-\log\overline{\xi}^{(t)}(s,a)\mathbf{1}_N\right\|_2+\frac{1-\alpha}{\tau}\left\|\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a)-\tau\log\boldsymbol{\xi}^{(t)}(s,a)-V_\tau^\star(s)\mathbf{1}_N\right\|_2\end{aligned} \tag{186}$$

where (a) is due to the doubly stochasticity property of $\boldsymbol{W}$ and (b) is from the fact $\|\boldsymbol{W}-\boldsymbol{I}_N\|_2\leq 2$. We further bound the second term as follows:

$$\begin{aligned}&\left\|\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a)-\tau\log\boldsymbol{\xi}^{(t)}(s,a)-V_\tau^\star(s)\mathbf{1}_N\right\|_2\\ &=\left\|\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a)-\tau\log\boldsymbol{\xi}^{(t)}(s,a)-\left(Q_\tau^\star(s,a)-\tau\log\pi_\tau^\star(a|s)\right)\mathbf{1}_N\right\|_2\\ &\leq\left\|\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a)-Q_\tau^\star(s,a)\mathbf{1}_N\right\|_2+\tau\left\|\log\boldsymbol{\xi}^{(t)}(s,a)-\log\pi_\tau^\star(a|s)\mathbf{1}_N\right\|_2\end{aligned}$$

$$
\begin{aligned}
&\leq \left\|\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a) - \widehat{Q}_\tau(s,a)\mathbf{1}_N\right\|_2 + \left\|\widehat{Q}_\tau(s,a)\mathbf{1}_N - Q^\star_\tau(s,a)\mathbf{1}_N\right\|_2 \\
&\qquad + \tau\left\|\log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\pi}^{(t)}(a|s)\mathbf{1}_N\right\|_2 + \tau\left\|\log \overline{\pi}^{(t)}(a|s)\mathbf{1}_N - \log \pi^\star_\tau(a|s)\mathbf{1}_N\right\|_2 \\
&= \sigma\left\|\boldsymbol{T}^{(t)}(s,a) - \widehat{Q}^{(t)}_\tau(s,a)\mathbf{1}_N\right\|_2 + \sqrt{N}\left|\widehat{Q}^{(t)}_\tau(s,a) - Q^\star_\tau(s,a)\right| \\
&\qquad + \tau\left\|\log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\pi}^{(t)}(a|s)\mathbf{1}_N\right\|_2 + \tau\sqrt{N}\left|\log \overline{\pi}^{(t)}(a|s) - \log \pi^\star_\tau(a|s)\right|. \qquad (187)
\end{aligned}
$$

Here, the first step results from the following relation established in Nachum et al. (2017):

$$
\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad V^\star_\tau(s) = -\tau \log \pi^\star_\tau(a|s) + Q^\star_\tau(s,a)\,, \qquad (188)
$$

which also leads to

$$
\left\|\log \overline{\pi}^{(t)} - \log \pi^\star_\tau\right\|_\infty \leq \frac{2}{\tau}\left\|Q^\star_\tau - \tau \log \overline{\xi}^{(t)}\right\|_\infty \qquad (189)
$$

by Lemma 14. For the remaining terms in (187), we have

$$
\left|\widehat{Q}^{(t)}_\tau(s,a) - Q^\star_\tau(s,a)\right| \leq \left\|\widehat{Q}^{(t)}_\tau - \overline{Q}^{(t)}_\tau\right\|_\infty + \left\|\overline{Q}^{(t)}_\tau - Q^\star_\tau\right\|_\infty, \qquad (190)
$$

and

$$
\begin{aligned}
\left\|\log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\pi}^{(t)}(a|s)\mathbf{1}_N\right\|_2 &= \sqrt{\sum_{n=1}^N \left(\log \xi^{(t)}_n(s,a) - \log \overline{\pi}^{(t)}(a|s)\right)^2} \\
&\leq \sqrt{\sum_{n=1}^N 2\left\|\log \xi^{(t)}_n - \log \overline{\xi}^{(t)}\right\|_\infty^2} \\
&\leq \sqrt{\sum_{n=1}^N 2\left\|u^{(t)}\right\|_\infty^2} = \sqrt{2N}\left\|u^{(t)}\right\|_\infty, \qquad (191)
\end{aligned}
$$

where the first inequality again results from Lemma 14. Plugging (189), (190), (191) into (187) and using the definition of $u^{(t)}, v^{(t)}$, we arrive at

$$
\begin{aligned}
w^{(t)}(s,a) \leq &\left(2\alpha + (1-\alpha)\cdot\sqrt{2N}\right)\left\|u^{(t)}\right\|_\infty + \frac{1-\alpha}{\tau}\left\|v^{(t)}\right\|_\infty + \frac{1-\alpha}{\tau}\cdot\sqrt{N}\left(\left\|\widehat{Q}^{(t)}_\tau - \overline{Q}^{(t)}_\tau\right\|_\infty + \left\|\overline{Q}^{(t)}_\tau - Q^\star_\tau\right\|_\infty\right) \\
&+ \frac{1-\alpha}{\tau}\cdot 2\sqrt{N}\left\|Q^\star_\tau - \tau\log\overline{\xi}^{(t)}\right\|_\infty.
\end{aligned}
$$

Using previous display, we can write (185) as

$$
\begin{aligned}
&\left\|\boldsymbol{Q}^{(t+1)}_\tau(s,a) - \boldsymbol{Q}^{(t)}_\tau(s,a)\right\|_2 \\
&\leq M\sqrt{N}\Bigg\{\left(2\alpha + (1-\alpha)\cdot\sqrt{2N}\right)\left\|u^{(t)}\right\|_\infty + \frac{1-\alpha}{\tau}\sigma\left\|v^{(t)}\right\|_\infty \\
&\qquad + \frac{1-\alpha}{\tau}\cdot\sqrt{N}\left(M\left\|u^{(t)}\right\|_\infty + \left\|\overline{Q}^{(t)}_\tau - Q^\star_\tau\right\|_\infty\right) + \frac{1-\alpha}{\tau}\cdot 2\sqrt{N}\left\|Q^\star_\tau - \tau\log\overline{\xi}^{(t)}\right\|_\infty\Bigg\}. \qquad (192)
\end{aligned}
$$

Combining (178) with the above expression (192), we get

$$
\begin{aligned}
\left\|v^{(t+1)}\right\|_\infty \leq \sigma\left(1 + \frac{\eta M\sqrt{N}}{1-\gamma}\sigma\right)\left\|v^{(t)}\right\|_\infty + \sigma M\sqrt{N}\Bigg\{&\left(2\alpha + (1-\alpha)\cdot\sqrt{2N} + \frac{1-\alpha}{\tau}\cdot\sqrt{N}M\right)\left\|u^{(t)}\right\|_\infty \\
&+ \frac{1-\alpha}{\tau}\cdot\sqrt{N}\left\|\overline{Q}^{(t)}_\tau - Q^\star_\tau\right\|_\infty + \frac{1-\alpha}{\tau}\cdot 2\sqrt{N}\left\|Q^\star_\tau - \tau\log\overline{\xi}^{(t)}\right\|_\infty\Bigg\}. \qquad (193)
\end{aligned}
$$

Step 5: bound $\|\overline{Q}_\tau^{(t+1)} - Q_\tau^\star\|_\infty$. For any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, we observe that

$$\begin{aligned}
&Q_\tau^\star(s,a) - \overline{Q}_\tau^{(t+1)}(s,a) \\
&= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}[V_\tau^\star(s')] - \left(r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[V_\tau^{\overline{\pi}^{(t+1)}}(s')\right]\right) \\
&= \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[\tau\log\left(\left\|\exp\left(\frac{Q_\tau^\star(s',\cdot)}{\tau}\right)\right\|_1\right)\right] - \gamma \mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a),\\ a'\sim\overline{\pi}^{(t+1)}(\cdot|s')}}\left[\overline{Q}_\tau^{(t+1)}(s',a') - \tau\log\overline{\pi}^{(t+1)}(a'|s')\right], \qquad (194)
\end{aligned}$$

where the first step invokes the definition of Q_τ (cf. (7a)), and the second step is due to the following expression of $V_\tau^\star$ established in Nachum et al. (2017):

$$V_\tau^\star(s) = \tau\log\left(\left\|\exp\left(\frac{Q_\tau^\star(s,\cdot)}{\tau}\right)\right\|_1\right). \qquad (195)$$

To continue, note that by (173) and (52b) we have

$$\begin{aligned}
\log\overline{\pi}^{(t+1)}(a|s) &= \log\overline{\xi}^{(t+1)}(s,a) - \log\left(\|\overline{\xi}^{(t+1)}(s,\cdot)\|_1\right) \\
&= \alpha\log\overline{\xi}^{(t)}(s,a) + (1-\alpha)\frac{\widehat{Q}_\tau^{(t)}(s,a)}{\tau} - \log\left(\|\overline{\xi}^{(t+1)}(s,\cdot)\|_1\right). \qquad (196)
\end{aligned}$$

Plugging (196) into (194) and (192) establishes the bounds on

$$\begin{aligned}
Q_\tau^\star(s,a) - \overline{Q}_\tau^{(t+1)}(s,a) = \gamma &\mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[\tau\log\left(\left\|\exp\left(\frac{Q_\tau^\star(s',\cdot)}{\tau}\right)\right\|_1\right) - \tau\log\left(\left\|\overline{\xi}^{(t+1)}(s',\cdot)\right\|_1\right)\right] \\
&- \gamma \mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a),\\ a'\sim\overline{\pi}^{(t+1)}(\cdot|s')}}\left[\overline{Q}_\tau^{(t+1)}(s',a') - \tau\underbrace{\left(\alpha\log\overline{\xi}^{(t)}(s',a') + (1-\alpha)\frac{\widehat{Q}_\tau^{(t)}(s',a')}{\tau}\right)}_{=\log\overline{\xi}^{(t+1)}(s',a')}\right] \qquad (197)
\end{aligned}$$

for any $(s,a) \in \mathcal{S}\times\mathcal{A}$. In view of property (175), the first term on the right-hand side of (197) can be bounded by

$$\tau\log\left(\left\|\exp\left(\frac{Q_\tau^\star(s',\cdot)}{\tau}\right)\right\|_1\right) - \tau\log\left(\|\overline{\xi}^{(t+1)}(s',\cdot)\|_1\right) \le \|Q_\tau^\star - \tau\log\overline{\xi}^{(t+1)}\|_\infty.$$

Plugging the above expression into (197), we have

$$0 \le Q_\tau^\star(s,a) - \overline{Q}_\tau^{(t+1)}(s,a) \le \gamma\|Q_\tau^\star - \tau\log\overline{\xi}^{(t+1)}\|_\infty - \gamma\min_{s,a}\left(\overline{Q}_\tau^{(t+1)}(s,a) - \tau\log\overline{\xi}^{(t+1)}(s,a)\right),$$

which gives

$$\|Q_\tau^\star - \overline{Q}_\tau^{(t+1)}\|_\infty \le \gamma\|Q_\tau^\star - \tau\log\overline{\xi}^{(t+1)}\|_\infty + \gamma\max\left\{0, -\min_{s,a}\left(\overline{Q}_\tau^{(t+1)}(s,a) - \tau\log\overline{\xi}^{(t+1)}(s,a)\right)\right\}. \qquad (198)$$

Plugging the above inequality into (182) and (193) establishes the bounds on $\Omega_3^{(t+1)}$ and $\Omega_2^{(t+1)}$ in (60), respectively.

Step 6: bound $-\min_{s,a}\left(\overline{Q}_\tau^{(t+1)}(s,a) - \tau\log\overline{\xi}^{(t+1)}(s,a)\right)$. We need the following lemma which is adapted from Lemma 1 in Cen et al. (2022a):

Lemma 16 (Performance improvement of FedNPG with entropy regularization)**.** *Suppose $0 < \eta \le (1-\gamma)/\tau$. For any state-action pair $(s_0,a_0) \in \mathcal{S}\times\mathcal{A}$, one has*

$$\overline{V}_\tau^{(t+1)}(s_0) - \overline{V}_\tau^{(t)}(s_0) \ge \frac{1}{\eta}\mathop{\mathbb{E}}_{s\sim d_{s_0}^{\overline{\pi}^{(t+1)}}}\left[\alpha\mathsf{KL}\big(\overline{\pi}^{(t+1)}(\cdot|s_0)\,\|\,\overline{\pi}^{(t)}(\cdot|s_0)\big) + \mathsf{KL}\big(\overline{\pi}^{(t)}(\cdot|s_0)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s_0)\big)\right]$$

$$- \frac{2}{1-\gamma} \left\| \widehat{Q}_\tau^{(t)} - \overline{Q}_\tau^{(t)} \right\|_\infty , \tag{199}$$

$$\overline{Q}_\tau^{(t+1)}(s_0, a_0) - \overline{Q}_\tau^{(t)}(s_0, a_0) \geq - \frac{2\gamma}{1-\gamma} \left\| \widehat{Q}_\tau^{(t)} - \overline{Q}_\tau^{(t)} \right\|_\infty . \tag{200}$$

Proof. See Appendix C.8. □

Using (200), we have

$$\begin{aligned}
&\overline{Q}_\tau^{(t+1)}(s,a) - \tau \left(\alpha \log \overline{\xi}^{(t)}(s,a) + (1-\alpha) \frac{\widehat{Q}_\tau^{(t)}(s,a)}{\tau} \right) \\
&\geq \overline{Q}_\tau^{(t)}(s,a) - \tau \left(\alpha \log \overline{\xi}^{(t)}(s,a) + (1-\alpha) \frac{\widehat{Q}_\tau^{(t)}(s,a)}{\tau} \right) - \frac{2\gamma}{1-\gamma} \left\| \widehat{Q}_\tau^{(t)} - \overline{Q}_\tau^{(t)} \right\|_\infty \\
&\geq \alpha \left(\overline{Q}_\tau^{(t)}(s,a) - \tau \log \overline{\xi}^{(t)}(s,a) \right) - \frac{2\gamma + \eta\tau}{1-\gamma} \left\| \widehat{Q}_\tau^{(t)} - \overline{Q}_\tau^{(t)} \right\|_\infty ,
\end{aligned} \tag{201}$$

which gives

$$\begin{aligned}
&- \min_{s,a} \left(\overline{Q}_\tau^{(t+1)}(s,a) - \tau \log \overline{\xi}^{(t+1)}(s,a) \right) \\
&\leq -\alpha \min_{s,a} \left(\overline{Q}_\tau^{(t)}(s,a) - \tau \log \overline{\xi}^{(t)}(s,a) \right) + \frac{2\gamma + \eta\tau}{1-\gamma} M \left\| u^{(t)} \right\|_\infty \\
&\leq \alpha \max \left\{ 0, \min_{s,a} \left(\overline{Q}_\tau^{(t)}(s,a) - \tau \log \overline{\xi}^{(t)}(s,a) \right) \right\} + \frac{2\gamma + \eta\tau}{1-\gamma} M \left\| u^{(t)} \right\|_\infty .
\end{aligned} \tag{202}$$

This establishes the bounds on $\Omega_4^{(t+1)}$ in (60).

C.2 Proof of Lemma 2

Let $f(\lambda)$ denote the characteristic function. In view of some direct calculations, we obtain

$$\begin{aligned}
f(\lambda) = (\lambda - \alpha) \Bigg\{ & \underbrace{(\lambda - \sigma\alpha)(\lambda - \sigma(1+\sigma b\eta))(\lambda - (1-\alpha)\gamma - \alpha)}_{=: f_0(\lambda)} \\
& - \frac{\eta\sigma^2}{1-\gamma} \underbrace{[S(\lambda - (1-\alpha)\gamma - \alpha) + \gamma c d M \eta + (1-\alpha)(2+\gamma) M c \eta]}_{=: f_1(\lambda)} \Bigg\} \\
& - \frac{\tau\eta^3\gamma}{(1-\gamma)^2} \cdot 2cdM\sigma^2 ,
\end{aligned} \tag{203}$$

where, for the notation simplicity, we let

$$b := \frac{M\sqrt{N}}{1-\gamma} , \tag{204a}$$

$$c := \frac{MN}{1-\gamma} = \sqrt{N} b , \tag{204b}$$

$$d := \frac{2\gamma + \eta\tau}{1-\gamma} . \tag{204c}$$

Note that among all these new notation we introduce, S, d are dependent of η. To decouple the dependence, we give their upper bounds as follows

$$d_0 := \frac{1+\gamma}{1-\gamma} \geq d , \tag{205}$$

$$S_0 \coloneqq M\sqrt{N}\left(2+\sqrt{2N}+\frac{M\sqrt{N}}{\tau}\right) \geq S\,, \tag{206}$$

where (205) follows from $\eta \leq (1-\gamma)/\tau$, and (206) uses the fact that $\alpha \leq 1$ and $1-\alpha \leq 1$.

Let

$$\lambda^\star \coloneqq \max\left\{\frac{3+\sigma}{4}, \frac{1+(1-\alpha)\gamma+\alpha}{2}\right\}. \tag{207}$$

Since $\boldsymbol{A}(\rho)$ is a nonnegative matrix, by Perron-Frobenius Theorem (see Horn and Johnson (2012), Theorem 8.3.1), $\rho(\eta)$ is an eigenvalue of $\boldsymbol{A}(\rho)$. So to verify (66), it suffices to show that $f(\lambda) > 0$ for any $\lambda \in [\lambda^\star, \infty)$. To do so, in the following we first show that $f(\lambda^\star) > 0$, and then we prove that f is nondecreasing on $[\lambda^\star, \infty)$.

- *Showing* $f(\lambda^\star) > 0$. We first lower bound $f_0(\lambda^\star)$. Since $\lambda^\star \geq \frac{3+\sigma}{4}$, we have

$$\lambda^\star - \sigma(1+\sigma b\eta) \geq \frac{1-\sigma}{4}\,, \tag{208}$$

and from $\lambda^\star \geq \frac{1+(1-\alpha)\gamma+\alpha}{2}$ we deduce

$$\lambda^\star - (1-\alpha)\gamma - \alpha \geq \frac{(1-\gamma)(1-\alpha)}{2} \tag{209}$$

and

$$\lambda^\star > \frac{1+\alpha}{2}\,, \tag{210}$$

which gives

$$\lambda^\star - \sigma\alpha \geq \frac{1+\alpha}{2} - \sigma\alpha\,. \tag{211}$$

Combining (211), (208), (209), we have that

$$f_0(\lambda^\star) \geq \frac{1-\sigma}{8}\left(\frac{1+\alpha}{2} - \sigma\alpha\right)\eta\tau\,. \tag{212}$$

To continue, we upper bound $f_1(\lambda^\star)$ as follows.

$$\begin{aligned} f_1(\lambda^\star) &\leq S\tau\eta + \gamma cdM\eta + \frac{2+\gamma}{1-\gamma}cM\tau\eta^2 \\ &= \eta\left(\tau\left(S + \frac{2+\gamma}{1-\gamma}Mc\eta\right) + \gamma cdM\right)\,. \end{aligned} \tag{213}$$

Plugging (212),(213) into (203) and using (210), we have

$$\begin{aligned} f(\lambda^\star) &> \frac{1-\alpha}{2}\left(f_0(\lambda^\star) - \frac{\eta\sigma^2}{1-\gamma}f_1(\lambda^\star)\right) - \frac{\tau\eta^3\gamma}{(1-\gamma)^2}\cdot 2cdM\sigma^2 \\ &\geq \frac{\tau\eta^2}{2(1-\gamma)}\left[\frac{1-\sigma}{8}\tau\left(1-\sigma+(1-\alpha)(\sigma-\frac{1}{2})\right) - \frac{\eta\sigma^2}{1-\gamma}\left(\tau\left(S+\frac{2+\gamma}{1-\gamma}Mc\eta\right)+5\gamma cdM\right)\right] \\ &= \frac{\tau\eta^2}{2(1-\gamma)}\left[\frac{(1-\sigma)^2}{8}\tau - \frac{\eta}{1-\gamma}\left(S\tau\sigma^2 + \frac{2+\gamma}{1-\gamma}Mc\sigma^2\tau\eta + \tau^2\left(\frac{1}{2}-\sigma^2\right)\cdot\frac{1-\sigma}{8} + 5\gamma cdM\sigma^2\right)\right] \\ &\geq \frac{\tau\eta^2}{2(1-\gamma)}\left[\frac{(1-\sigma)^2}{8}\tau - \frac{\eta}{1-\gamma}\left(S_0\tau\sigma^2 + \frac{(1-\sigma)^2}{16}\tau^2 + (2+\gamma+5\gamma d_0)\,cM\sigma^2\right)\right] \geq 0\,, \end{aligned}$$

where the penultimate inequality uses $\frac{1}{2}-\sigma \leq \frac{1-\sigma}{2}$, and the last inequality follows from the definition of ζ (cf. (64)).

- *Proving f is non-decreasing on $[\lambda^\star, \infty)$.* Note that

$$\eta \le \zeta \le \frac{(1-\gamma)(1-\sigma)^2}{8S_0\sigma^2}\,,$$

thus we have

$$\forall \lambda \ge \lambda^\star: \quad f_0'(\lambda) - \frac{\eta\sigma^2}{1-\gamma} f_1'(\lambda) \ge (\lambda - \sigma\alpha)(\lambda - \sigma(1+\sigma b\eta)) - \frac{\eta}{1-\gamma} S\sigma^2 \ge 0\,,$$

which indicates that $f_0 - f_1$ is non-decreasing on $[\lambda^\star, \infty)$. Therefore, f is non-decreasing on $[\lambda^\star, \infty)$.

C.3 Proof of Lemma 3

Note that bounding $u^{(t+1)}(s,a)$ is identical to the proof in Appendix C.1 and shall be omitted. The rest of the proof also follows closely that of Lemma 1, and we only highlight the differences due to approximation error for simplicity.

Step 2: bound $v^{(t+1)}(s,a) = \left\|\boldsymbol{T}^{(t+1)}(s,a) - \widehat{q}_\tau^{(t+1)}(s,a)\mathbf{1}_N\right\|_2$. Let $\boldsymbol{q}_\tau^{(t)} := \left(q_{\tau,1}^{\pi_1^{(t)}}, \cdots, q_{\tau,N}^{\pi_N^{(t)}}\right)^\top$. Similar to (178) we have

$$\begin{aligned}
&\left\|\boldsymbol{T}^{(t+1)}(s,a) - \widehat{q}_\tau^{(t+1)}(s,a)\mathbf{1}_N\right\|_2 \\
&\le \sigma\left\|\boldsymbol{T}^{(t)}(s,a) - \widehat{q}_\tau^{(t)}(s,a)\mathbf{1}_N\right\|_2 + \sigma\left\|\boldsymbol{q}_\tau^{(t+1)}(s,a) - \boldsymbol{q}_\tau^{(t)}(s,a)\right\|_2 \\
&\le \sigma\left\|\boldsymbol{T}^{(t)}(s,a) - \widehat{q}_\tau^{(t)}(s,a)\mathbf{1}_N\right\|_2 + \sigma\left\|\boldsymbol{Q}_\tau^{(t+1)}(s,a) - \boldsymbol{Q}_\tau^{(t)}(s,a)\right\|_2 + 2\sigma\left\|\boldsymbol{e}\right\|_2 .
\end{aligned} \tag{214}$$

Step 3: bound $\left\|Q_\tau^\star - \tau\log\overline{\xi}^{(t+1)}\right\|_\infty$. In the context of inexact updates, (179) writes

$$\left\|Q_\tau^\star - \tau\log\overline{\xi}^{(t+1)}\right\|_\infty \le \alpha\left\|Q_\tau^\star - \tau\log\overline{\xi}^{(t)}\right\|_\infty + (1-\alpha)\left\|Q_\tau^\star - \overline{Q}_\tau^{(t)}\right\|_\infty + (1-\alpha)\left\|\overline{Q}_\tau^{(t)} - \widehat{q}_\tau^{(t)}\right\|_\infty .$$

For the last term, following a similar argument in (180) leads to

$$\begin{aligned}
\left\|\overline{Q}_\tau^{(t)} - \widehat{q}_\tau^{(t)}\right\|_\infty &= \left\|\frac{1}{N}\sum_{n=1}^N Q_{\tau,n}^{\pi_n^{(t)}} - \frac{1}{N}\sum_{n=1}^N Q_{\tau,n}^{\overline{\pi}^{(t)}}\right\|_\infty + \left\|\frac{1}{N}\sum_{n=1}^N \left(Q_{\tau,n}^{\pi_n^{(t)}} - q_{\tau,n}^{\pi_n^{(t)}}\right)\right\|_\infty \\
&\le M\cdot\frac{1}{N}\sum_{n=1}^N \left\|\log\xi_n^{(t)} - \log\overline{\xi}^{(t)}\right\|_\infty + \frac{1}{N}\sum_{n=1}^N e_n \\
&\le M\left\|u^{(t)}\right\|_\infty + \left\|\boldsymbol{e}\right\|_\infty .
\end{aligned}$$

Combining the above two inequalities, we obtain

$$\left\|Q_\tau^\star - \tau\log\overline{\xi}^{(t+1)}\right\|_\infty \le \alpha\left\|Q_\tau^\star - \tau\log\overline{\xi}^{(t)}\right\|_\infty + (1-\alpha)\left\|Q_\tau^\star - \overline{Q}_\tau^{(t)}\right\|_\infty + (1-\alpha)\left(M\left\|u^{(t)}\right\|_\infty + \left\|\boldsymbol{e}\right\|_\infty\right) . \tag{215}$$

Step 4: bound $\left\|\boldsymbol{Q}_\tau^{(t+1)}(s,a) - \boldsymbol{Q}_\tau^{(t)}(s,a)\right\|_2$. We remark that the bound established in (185) still holds in the inexact setting, with the same definition for $w^{(t)}$:

$$\left\|\boldsymbol{Q}_\tau^{(t+1)}(s,a) - \boldsymbol{Q}_\tau^{(t)}(s,a)\right\|_2 \le M\sqrt{N}\left\|w^{(t)}\right\|_\infty . \tag{216}$$

To deal with the approximation error, we rewrite (187) as

$$\begin{aligned}
&\left\|\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a) - \tau\log\boldsymbol{\xi}^{(t)}(s,a) - V_\tau^\star(s)\mathbf{1}_N\right\|_2 \\
&= \left\|\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a) - \tau\log\boldsymbol{\xi}^{(t)}(s,a) - \left(Q_\tau^\star(s,a) - \tau\log\pi_\tau^\star(a|s)\right)\mathbf{1}_N\right\|_2
\end{aligned}$$

$$
\begin{aligned}
&\le \left\|\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a) - Q_\tau^\star(s,a)\mathbf{1}_N\right\|_2 + \tau\left\|\log \boldsymbol{\xi}^{(t)}(s,a) - \log \pi_\tau^\star(a|s)\mathbf{1}_N\right\|_2 \\
&\le \left\|\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a) - \widehat{q}_\tau(s,a)\mathbf{1}_N\right\|_2 + \left\|\widehat{q}_\tau(s,a)\mathbf{1}_N - Q_\tau^\star(s,a)\mathbf{1}_N\right\|_2 \\
&\qquad + \tau\left\|\log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\pi}^{(t)}(a|s)\mathbf{1}_N\right\|_2 + \tau\left\|\log \overline{\pi}^{(t)}(a|s)\mathbf{1}_N - \log \pi_\tau^\star(a|s)\mathbf{1}_N\right\|_2 \\
&\le \sigma\left\|\boldsymbol{T}^{(t)}(s,a) - \widehat{q}_\tau^{(t)}(s,a)\mathbf{1}_N\right\|_2 + \sqrt{N}\left|\widehat{q}_\tau^{(t)}(s,a) - Q_\tau^\star(s,a)\right| \\
&\qquad + \tau\left\|\log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\pi}^{(t)}(a|s)\mathbf{1}\right\|_2 + \tau\sqrt{N}\left|\log \overline{\pi}^{(t)}(a|s) - \log \pi_\tau^\star(a|s)\right|,
\end{aligned} \tag{217}
$$

where the second term can be upper-bounded by

$$
\begin{aligned}
\left|\widehat{q}_\tau^{(t)}(s,a) - Q_\tau^\star(s,a)\right| &\le \left\|\widehat{Q}_\tau^{(t)} - \overline{Q}_\tau^{(t)}\right\|_\infty + \left\|\overline{Q}_\tau^{(t)} - Q_\tau^\star\right\|_\infty + \left\|\widehat{q}_\tau^{(t)}(s,a) - \widehat{Q}_\tau^{(t)}(s,a)\right\|_\infty \\
&\le \left\|\widehat{Q}_\tau^{(t)} - \overline{Q}_\tau^{(t)}\right\|_\infty + \left\|\overline{Q}_\tau^{(t)} - Q_\tau^\star\right\|_\infty + \|\boldsymbol{e}\|_\infty .
\end{aligned} \tag{218}
$$

Combining (218), (217) and the established bounds in (186), (189), (191) leads to

$$
\begin{aligned}
w^{(t)}(s,a) \le{}& \left(2\alpha + (1-\alpha)\cdot\sqrt{2N}\right)\left\|u^{(t)}\right\|_\infty + \frac{1-\alpha}{\tau}\left\|v^{(t)}\right\|_\infty \\
&+ \frac{1-\alpha}{\tau}\cdot\sqrt{N}\left(\left\|\widehat{Q}_\tau^{(t)} - \overline{Q}_\tau^{(t)}\right\|_\infty + \left\|\overline{Q}_\tau^{(t)} - Q_\tau^\star\right\|_\infty + \|\boldsymbol{e}\|_\infty\right) + \frac{1-\alpha}{\tau}\cdot 2\sqrt{N}\left\|Q_\tau^\star - \tau\log\overline{\xi}^{(t)}\right\|_\infty .
\end{aligned}
$$

Combining the above inequality with (216) and (214) gives

$$
\begin{aligned}
\left\|v^{(t+1)}\right\|_\infty \le{}& \sigma\left(1 + \frac{\eta M\sqrt{N}}{1-\gamma}\sigma\right)\left\|v^{(t)}\right\|_\infty + \sigma M\sqrt{N}\Bigg\{\left(2\alpha + (1-\alpha)\cdot\sqrt{2N} + \frac{1-\alpha}{\tau}\cdot\sqrt{N}M\right)\left\|u^{(t)}\right\|_\infty \\
&+ \frac{1-\alpha}{\tau}\cdot\sqrt{N}\left(\left\|\overline{Q}_\tau^{(t)} - Q_\tau^\star\right\|_\infty + \|\boldsymbol{e}\|_\infty\right) + \frac{1-\alpha}{\tau}\cdot 2\sqrt{N}\left\|Q_\tau^\star - \tau\log\overline{\xi}^{(t)}\right\|_\infty\Bigg\} + 2\sigma\sqrt{N}\,\|\boldsymbol{e}\|_\infty .
\end{aligned} \tag{219}
$$

Step 5: bound $\left\|\overline{Q}_\tau^{(t+1)} - Q_\tau^\star\right\|_\infty$. It is straightforward to verify that (198) applies to the inexact updates as well:

$$
\left\|Q_\tau^\star - \overline{Q}_\tau^{(t+1)}\right\|_\infty \le \gamma\left\|Q_\tau^\star - \tau\log\overline{\xi}^{(t+1)}\right\|_\infty + \gamma\left(-\min_{s,a}\left(\overline{Q}_\tau^{(t+1)}(s,a) - \tau\log\overline{\xi}^{(t+1)}(s,a)\right)\right) .
$$

Plugging the above inequality into (215) and (219) establishes the bounds on $\Omega_3^{(t+1)}$ and $\Omega_2^{(t+1)}$ in (79), respectively.

Step 6: bound $-\min_{s,a}\left(\overline{Q}_\tau^{(t+1)}(s,a) - \tau\log\overline{\xi}^{(t+1)}(s,a)\right)$. We obtain the following lemma by interpreting the approximation error $\boldsymbol{e}$ as part of the consensus error $\left\|\widehat{Q}_\tau^{(t)} - \overline{Q}_\tau^{(t)}\right\|_\infty$ in Lemma 16.

Lemma 17 (inexact version of Lemma 16)**.** *Suppose $0 < \eta \le (1-\gamma)/\tau$. For any state-action pair $(s_0, a_0) \in \mathcal{S}\times\mathcal{A}$, one has*

$$
\begin{aligned}
\overline{V}_\tau^{(t+1)}(s_0) - \overline{V}_\tau^{(t)}(s_0) \ge{}& \frac{1}{\eta}\mathop{\mathbb{E}}_{s\sim d_{s_0}^{\overline{\pi}^{(t+1)}}}\left[\alpha\mathsf{KL}\big(\overline{\pi}^{(t+1)}(\cdot|s_0)\,\|\,\overline{\pi}^{(t)}(\cdot|s_0)\big) + \mathsf{KL}\big(\overline{\pi}^{(t)}(\cdot|s_0)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s_0)\big)\right] \\
&- \frac{2}{1-\gamma}\left(\left\|\widehat{Q}_\tau^{(t)} - \overline{Q}_\tau^{(t)}\right\|_\infty + \|\boldsymbol{e}\|_\infty\right) ,
\end{aligned} \tag{220}
$$

$$
\overline{Q}_\tau^{(t+1)}(s_0,a_0) - \overline{Q}_\tau^{(t)}(s_0,a_0) \ge -\frac{2\gamma}{1-\gamma}\left(\left\|\widehat{Q}_\tau^{(t)} - \overline{Q}_\tau^{(t)}\right\|_\infty + \|\boldsymbol{e}\|_\infty\right) . \tag{221}
$$

Using (221), we have

$$\begin{aligned}
&\overline{Q}_\tau^{(t+1)}(s,a) - \tau\left(\alpha \log \overline{\xi}^{(t)}(s,a) + (1-\alpha)\frac{\widehat{Q}_\tau^{(t)}(s,a)}{\tau}\right) \\
&\geq \overline{Q}_\tau^{(t)}(s,a) - \tau\left(\alpha \log \overline{\xi}^{(t)}(s,a) + (1-\alpha)\frac{\widehat{Q}_\tau^{(t)}(s,a)}{\tau}\right) - \frac{2\gamma}{1-\gamma}\left(\left\|\widehat{Q}_\tau^{(t)} - \overline{Q}_\tau^{(t)}\right\|_\infty + \|\boldsymbol{e}\|_\infty\right) \\
&\geq \alpha\left(\overline{Q}_\tau^{(t)}(s,a) - \tau \log \overline{\xi}^{(t)}(s,a)\right) - \frac{2\gamma+\eta\tau}{1-\gamma}\left\|\widehat{Q}_\tau^{(t)} - \overline{Q}_\tau^{(t)}\right\|_\infty - \frac{2\gamma}{1-\gamma}\|\boldsymbol{e}\|_\infty ,
\end{aligned} \tag{222}$$

which gives

$$\begin{aligned}
&-\min_{s,a}\left(\overline{Q}_\tau^{(t+1)}(s,a) - \tau \log \overline{\xi}^{(t+1)}(s,a)\right) \\
&\leq -\alpha \min_{s,a}\left(\overline{Q}_\tau^{(t)}(s,a) - \tau \log \overline{\xi}^{(t)}(s,a)\right) + \frac{2\gamma+\eta\tau}{1-\gamma} M \left\|u^{(t)}\right\|_\infty + \frac{2\gamma}{1-\gamma}\|\boldsymbol{e}\|_\infty .
\end{aligned} \tag{223}$$

C.4 Proof of Lemma 4

Step 1: bound $u^{(t+1)}(s,a) = \left\|\log \boldsymbol{\xi}^{(t+1)}(s,a) - \log \overline{\xi}^{(t+1)}(s,a)\mathbf{1}_N\right\|_2$. Following the same strategy in establishing (177), we have

$$\begin{aligned}
&\left\|\log \boldsymbol{\xi}^{(t+1)}(s,a) - \log \overline{\xi}^{(t+1)}(s,a)\mathbf{1}_N\right\|_2 \\
&= \left\|\left(\boldsymbol{W}\log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\xi}^{(t)}(s,a)\mathbf{1}_N\right) + \frac{\eta}{1-\gamma}\left(\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a) - \widehat{Q}^{(t)}(s,a)\mathbf{1}_N\right)\right\|_2 \\
&\leq \sigma\left\|\log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\xi}^{(t)}(s,a)\mathbf{1}_N\right\|_2 + \frac{\eta}{1-\gamma}\sigma\left\|\boldsymbol{T}^{(t)}(s,a) - \widehat{Q}^{(t)}(s,a)\mathbf{1}_N\right\|_2 ,
\end{aligned} \tag{224}$$

or equivalently

$$\left\|u^{(t+1)}\right\|_\infty \leq \sigma\left\|u^{(t)}\right\|_\infty + \frac{\eta}{1-\gamma}\sigma\left\|v^{(t)}\right\|_\infty . \tag{225}$$

Step 2: bound $v^{(t+1)}(s,a) = \left\|\boldsymbol{T}^{(t+1)}(s,a) - \widehat{Q}^{(t+1)}(s,a)\mathbf{1}_N\right\|_2$. In the same vein of establishing (178), we have

$$\begin{aligned}
&\left\|\boldsymbol{T}^{(t+1)}(s,a) - \widehat{Q}^{(t+1)}(s,a)\mathbf{1}_N\right\|_2 \\
&\leq \sigma\left\|\boldsymbol{T}^{(t)}(s,a) - \widehat{Q}^{(t)}(s,a)\mathbf{1}_N\right\|_2 + \sigma\left\|\boldsymbol{Q}^{(t+1)}(s,a) - \boldsymbol{Q}^{(t)}(s,a)\right\|_2 ,
\end{aligned} \tag{226}$$

The term $\left\|\boldsymbol{Q}^{(t+1)}(s,a) - \boldsymbol{Q}^{(t)}(s,a)\right\|_2$ can be bounded in a similar way in (185):

$$\left\|\boldsymbol{Q}^{(t+1)}(s,a) - \boldsymbol{Q}^{(t)}(s,a)\right\|_2 \leq \frac{(1+\gamma)\gamma}{(1-\gamma)^2}\sqrt{N}\left\|w_0^{(t)}\right\|_\infty , \tag{227}$$

where the coefficient $\frac{(1+\gamma)\gamma}{(1-\gamma)^2}$ comes from M in Lemma 15 when $\tau = 0$, and $w_0^{(t)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is defined as

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad w_0^{(t)}(s,a) := \left\|\log \boldsymbol{\xi}^{(t+1)}(s,a) - \log \boldsymbol{\xi}^{(t)}(s,a) - \frac{\eta}{1-\gamma}V^\star(s)\mathbf{1}_N\right\|_2 . \tag{228}$$

It remains to bound $\left\|w_0^{(t)}\right\|_\infty$. Towards this end, we rewrite (186) as

$$\begin{aligned}
&w_0^{(t)}(s,a) \\
&= \left\|\boldsymbol{W}\left(\log \boldsymbol{\xi}^{(t)}(s,a) + \frac{\eta}{1-\gamma}\boldsymbol{T}^{(t)}(s,a)\right) - \log \boldsymbol{\xi}^{(t)}(s,a) - \frac{\eta}{1-\gamma}V^\star(s)\mathbf{1}_N\right\|_2
\end{aligned}$$

$$
\begin{aligned}
&= \left\| (\boldsymbol{W} - \boldsymbol{I}) \left(\log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\xi}^{(t)}(s,a) \mathbf{1}_N \right) + \frac{\eta}{1-\gamma} \left(\boldsymbol{W} \boldsymbol{T}^{(t)}(s,a) - V^\star(s) \mathbf{1}_N \right) \right\|_2 \\
&\le 2 \left\| \log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\xi}^{(t)}(s,a) \mathbf{1}_N \right\|_2 + \frac{\eta}{1-\gamma} \left\| \boldsymbol{W} \boldsymbol{T}^{(t)}(s,a) - V^\star(s) \mathbf{1}_N \right\|_2 \\
&\le 2 \left\| \log \boldsymbol{\xi}^{(t)}(s,a) - \log \overline{\xi}^{(t)}(s,a) \mathbf{1}_N \right\|_2 + \frac{\eta}{1-\gamma} \left\| \boldsymbol{W} \boldsymbol{T}^{(t)}(s,a) - \widehat{Q}^{(t)}(s,a) \mathbf{1}_N \right\|_2 + \frac{\eta}{1-\gamma} \cdot \sqrt{N} \left| \widehat{Q}^{(t)}(s,a) - V^\star(s) \right| .
\end{aligned} \tag{229}
$$

Note that it holds for all $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$
\left| \widehat{Q}^{(t)}(s,a) - V^\star(s) \right| \le \frac{1}{1-\gamma}
$$

since $\widehat{Q}^{(t)}(s,a)$ and $V^\star(s)$ are both in $[0, 1/(1-\gamma)]$. This along with (229) gives

$$
w_0^{(t)}(s,a) \le 2 \left\| u^{(t)} \right\|_\infty + \frac{\eta}{1-\gamma} \left\| v^{(t)} \right\|_\infty + \frac{\eta \sqrt{N}}{(1-\gamma)^2} .
$$

Combining the above inequality with (227) and (226), we arrive at

$$
\left\| v^{(t+1)} \right\|_\infty \le \sigma \left(1 + \frac{(1+\gamma)\gamma \sqrt{N} \eta}{(1-\gamma)^3} \sigma \right) \left\| v^{(t)} \right\|_\infty + \frac{(1+\gamma)\gamma}{(1-\gamma)^2} \sqrt{N} \sigma \left\{ 2 \left\| u^{(t)} \right\|_\infty + \frac{\eta}{(1-\gamma)^2} \cdot \sqrt{N} \right\} . \tag{230}
$$

Step 3: establish the descent equation. The following lemma characterizes the improvement in $\phi^{(t)}(\eta)$ for every iteration of Algorithm 1, with the proof postponed to Appendix C.9.

Lemma 18 (Performance improvement of exact FedNPG)**.** *For all starting state distribution $\rho \in \Delta(\mathcal{S})$, we have the iterates of FedNPG satisfy*

$$
\phi^{(t+1)}(\eta) \le \phi^{(t)}(\eta) + \frac{2\eta}{(1-\gamma)^2} \left\| \widehat{Q}^{(t)} - \overline{Q}^{(t)} \right\|_\infty - \eta \left(V^\star(\rho) - \overline{V}^{(t)}(\rho) \right) , \tag{231}
$$

where

$$
\phi^{(t)}(\eta) := \mathbb{E}_{s \sim d_\rho^{\pi^\star}} \left[\mathsf{KL}\big(\pi^\star(\cdot|s) \,\|\, \overline{\pi}^{(t)}(\cdot|s) \big) \right] - \frac{\eta}{1-\gamma} \overline{V}^{(t)}(d_\rho^{\pi^\star}) , \quad \forall t \ge 0 . \tag{232}
$$

It remains to control the term $\left\| \overline{Q}^{(t)} - \widehat{Q}^{(t)} \right\|_\infty$. Similar to (180), for all $t \ge 0$, we have

$$
\begin{aligned}
\left\| \overline{Q}^{(t)} - \widehat{Q}^{(t)} \right\|_\infty &= \left\| \frac{1}{N} \sum_{n=1}^N Q_n^{\pi_n^{(t)}} - \frac{1}{N} \sum_{n=1}^N Q_n^{\overline{\pi}^{(t)}} \right\|_\infty \\
&\overset{(a)}{\le} \frac{(1+\gamma)\gamma}{(1-\gamma)^2} \cdot \frac{1}{N} \sum_{n=1}^N \left\| \log \xi_n^{(t)} - \log \overline{\xi}^{(t)} \right\|_\infty \\
&\overset{(b)}{\le} \frac{(1+\gamma)\gamma}{(1-\gamma)^2} \left\| u^{(t)} \right\|_\infty ,
\end{aligned} \tag{233}
$$

where (a) invokes Lemma 15 with $\tau = 0$ and (b) stems from the definition of $u^{(t)}$. This along with (231) gives

$$
\phi^{(t+1)}(\eta) \le \phi^{(t)}(\eta) + \frac{2(1+\gamma)\gamma}{(1-\gamma)^4} \eta \left\| u^{(t)} \right\|_\infty - \eta \left(V^\star(\rho) - \overline{V}^{(t)}(\rho) \right) .
$$

Step 4: bound the consensus error. To bound the consensus error $\left\| \log \pi_n^{(t)} - \log \overline{\pi}^{(t)} \right\|_\infty$ for all $n \in [N]$, we first upper bound the spectral norm of $\boldsymbol{B}(\eta)$ which we denote as $\rho(\boldsymbol{B}(\eta))$. Since $\boldsymbol{B}(\eta)$ is a nonnegative matrix, by Perron-Frobenius Theorem, $\rho(\boldsymbol{B}(\eta))$ is an eigenvalue of $\boldsymbol{B}(\eta)$. Hence, we only need to upper bound the eigenvalue of $\rho(\boldsymbol{B}(\eta))$.

The characteristic polynomial of $\boldsymbol{B}(\eta)$ is

$$
\begin{aligned}
f(\lambda) &= (\lambda-\sigma)\left(\lambda-\sigma\left(1+\frac{(1+\gamma)\gamma\sqrt{N}\eta}{(1-\gamma)^3}\sigma\right)\right)-\frac{\eta J}{1-\gamma}\sigma^2 \\
&= \lambda^2-\left(2+\frac{(1+\gamma)\gamma\sqrt{N}\eta}{(1-\gamma)^3}\sigma\right)\sigma\lambda+\left(1+\frac{(1+\gamma)\gamma\sqrt{N}\eta}{(1-\gamma)^3}\sigma-\frac{\eta J}{1-\gamma}\right)\sigma^2 .
\end{aligned}
$$

This gives

$$
\begin{aligned}
\rho(\boldsymbol{B}(\eta)) &\le \frac{\sigma}{2}\left[\left(2+\frac{(1+\gamma)\gamma\sqrt{N}\eta}{(1-\gamma)^3}\sigma\right)+\sqrt{\left(2+\frac{(1+\gamma)\gamma\sqrt{N}\eta}{(1-\gamma)^3}\sigma\right)^2-4\left(1+\frac{(1+\gamma)\gamma\sqrt{N}\eta}{(1-\gamma)^3}\sigma\right)+4\frac{\eta J}{1-\gamma}}\right] \\
&\le \frac{\sigma}{2}\left[\left(2+\frac{(1+\gamma)\gamma\sqrt{N}\eta}{(1-\gamma)^3}\sigma\right)+\sqrt{\left(\frac{(1+\gamma)\gamma\sqrt{N}\eta}{(1-\gamma)^3}\sigma\right)^2+4\frac{\eta J}{1-\gamma}}\right] \\
&\le \sigma\left[1+\frac{(1+\gamma)\gamma\sqrt{N}\eta}{(1-\gamma)^3}\sigma+\sqrt{\frac{\eta J}{1-\gamma}}\right] . \qquad (234)
\end{aligned}
$$

Note that when $\eta \le \eta_1$, we have (recall that $J = \frac{2(1+\gamma)\gamma}{(1-\gamma)^2}\sqrt{N}$):

$$
\frac{(1+\gamma)\gamma\sqrt{N}\eta}{(1-\gamma)^3}\sigma \le \frac{(1-\sigma)^2}{8}, \quad \text{and} \quad \frac{\eta J}{1-\gamma} \le \frac{(1-\sigma)^2}{4\sigma} .
$$

Plugging the above two expressions into (234) yields

$$
\begin{aligned}
\rho(\boldsymbol{B}(\eta)) &\le \sigma\left(1+(1-\sigma)^2/8+(1-\sigma)/(2\sqrt{\sigma})\right) \\
&\le \sigma\left(1+(1-\sigma)/(8\sigma)+(1-\sigma)/(2\sigma)\right) = \frac{3}{8}\sigma+\frac{5}{8} < 1 .
\end{aligned}
$$

Therefore, when $\eta \le \eta_1$, we have

$$
\begin{aligned}
\left\|\boldsymbol{\Omega}^{(t)}\right\|_2 &\le \rho(\boldsymbol{B}(\eta))\left\|\boldsymbol{\Omega}^{(t-1)}\right\|_2 + d_2(\eta) \\
&\le \cdots \le \rho^t(\boldsymbol{B}(\eta))\left\|\boldsymbol{\Omega}^{(0)}\right\|_2 + \sum_{i=0}^{t-1}\rho^i(\boldsymbol{B}(\eta))\frac{(1+\gamma)\gamma N\sigma}{(1-\gamma)^4}\eta \\
&\le \rho^t(\boldsymbol{B}(\eta))\left\|\boldsymbol{\Omega}^{(0)}\right\|_2 + \frac{2N\sigma}{(1-\gamma)^4(1-\rho(\boldsymbol{B}(\eta)))}\eta \\
&\le \left(\frac{3}{8}\sigma+\frac{5}{8}\right)^t\left\|\boldsymbol{\Omega}^{(0)}\right\|_2 + \frac{16N\sigma}{3(1-\gamma)^4(1-\sigma)}\eta . \qquad (235)
\end{aligned}
$$

Combining the above inequality with the following fact:

$$
\forall n \in [N]: \quad \left\|\log \pi_n^{(t)} - \log \overline{\pi}^{(t)}\right\|_\infty \le 2\left\|\log \xi_n^{(t)} - \log \overline{\xi}^{(t)}\right\|_\infty \le \Omega_1^{(t)} \le \left\|\boldsymbol{\Omega}^{(t)}\right\|_2 ,
$$

where the first inequality uses (176), we obtain (95).

C.5 Proof of Lemma 5

The bound on $u^{(t+1)}(s,a)$ is already established in Step 1 in Appendix C.1 and shall be omitted. As usual we only highlight the key differences with the proof of Lemma 4 due to approximation error.

Step 1: bound $v^{(t+1)}(s,a) = \left\|\boldsymbol{T}^{(t+1)}(s,a) - \widehat{q}^{(t+1)}(s,a)\mathbf{1}_N\right\|_2$. Let $\boldsymbol{q}^{(t)} := \left(q_1^{\pi_1^{(t)}}, \cdots, q_N^{\pi_N^{(t)}}\right)^\top$. From (107), we have

$$\begin{aligned}
&\left\|\boldsymbol{T}^{(t+1)}(s,a) - \widehat{q}^{(t+1)}(s,a)\mathbf{1}_N\right\|_2 \\
&= \left\|\boldsymbol{W}\left(\boldsymbol{T}^{(t)}(s,a) + \boldsymbol{q}^{(t+1)}(s,a) - \boldsymbol{q}^{(t)}(s,a)\right) - \widehat{q}^{(t+1)}(s,a)\mathbf{1}_N\right\|_2 \\
&= \left\|\left(\boldsymbol{W}\boldsymbol{T}^{(t)}(s,a) - \widehat{q}^{(t)}(s,a)\mathbf{1}_N\right) + \boldsymbol{W}\left(\boldsymbol{q}^{(t+1)}(s,a) - \boldsymbol{q}^{(t)}(s,a)\right) + \left(\widehat{q}^{(t)}(s,a) - \widehat{q}^{(t+1)}(s,a)\right)\mathbf{1}_N\right\|_2 \\
&\le \sigma\left\|\boldsymbol{T}^{(t)}(s,a) - \widehat{q}^{(t)}(s,a)\mathbf{1}_N\right\|_2 + \sigma\left\|\left(\boldsymbol{q}^{(t+1)}(s,a) - \boldsymbol{q}^{(t)}(s,a)\right) + \left(\widehat{q}^{(t)}(s,a) - \widehat{q}^{(t+1)}(s,a)\right)\mathbf{1}_N\right\|_2 \\
&\le \sigma\left\|\boldsymbol{T}^{(t)}(s,a) - \widehat{q}^{(t)}(s,a)\mathbf{1}_N\right\|_2 + \sigma\left\|\boldsymbol{q}^{(t+1)}(s,a) - \boldsymbol{q}^{(t)}(s,a)\right\|_2 \\
&\le \sigma\left\|\boldsymbol{T}^{(t)}(s,a) - \widehat{q}^{(t)}(s,a)\mathbf{1}_N\right\|_2 + \sigma\left\|\boldsymbol{Q}^{(t+1)}(s,a) - \boldsymbol{Q}^{(t)}(s,a)\right\|_2 + 2\sigma\sqrt{N}\left\|\boldsymbol{e}\right\|_\infty .
\end{aligned} \tag{236}$$

Note that (227) still holds for inexact FedNPG:

$$\left\|\boldsymbol{Q}^{(t+1)}(s,a) - \boldsymbol{Q}^{(t)}(s,a)\right\|_2 \le \frac{(1+\gamma)\gamma}{(1-\gamma)^2}\sqrt{N}\left\|w_0^{(t)}\right\|_\infty , \tag{237}$$

where $w_0^{(t)}$ is defined in (228). We rewrite (229), the bound on $w_0^{(t)}(s,a)$, as

$$\begin{aligned}
w_0^{(t)}(s,a) \le\ & 2\left\|\log\boldsymbol{\xi}^{(t)}(s,a) - \log\overline{\xi}^{(t)}(s,a)\mathbf{1}_N\right\|_2 \\
& + \frac{\eta}{1-\gamma}\left\|\boldsymbol{T}^{(t)}(s,a) - \widehat{q}^{(t)}(s,a)\mathbf{1}_N\right\|_2 + \frac{\eta\sigma}{1-\gamma}\cdot\sqrt{N}\left|\widehat{q}^{(t)}(s,a) - V^\star(s)\right| .
\end{aligned} \tag{238}$$

With the following bound

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \quad \left|\widehat{q}^{(t)}(s,a) - V^\star(s)\right| \le \left\|\widehat{q}^{(t)} - \overline{Q}^{(t)}\right\|_\infty + \frac{1}{1-\gamma}$$

in mind, we write (229) as

$$w_0^{(t)}(s,a) \le 2\left\|u^{(t)}\right\|_\infty + \frac{\eta\sigma}{1-\gamma}\left\|v^{(t)}\right\|_\infty + \frac{\eta}{1-\gamma}\cdot\sqrt{N}\left(\left\|\widehat{q}^{(t)} - \overline{q}^{(t)}\right\|_\infty + \frac{1}{1-\gamma}\right) .$$

Putting all pieces together, we obtain

$$\begin{aligned}
\left\|v^{(t+1)}\right\|_\infty \le\ & \sigma\left(1 + \frac{(1+\gamma)\gamma\sqrt{N}\eta}{(1-\gamma)^3}\sigma\right)\left\|v^{(t)}\right\|_\infty \\
& + \frac{(1+\gamma)\gamma}{(1-\gamma)^2}\sqrt{N}\sigma\left\{2\left\|u^{(t)}\right\|_\infty + \frac{\eta\sqrt{N}}{(1-\gamma)^2} + \frac{\eta\sqrt{N}}{1-\gamma}\left\|\boldsymbol{e}\right\|_\infty\right\} \\
& + 2\sigma\sqrt{N}\left\|\boldsymbol{e}\right\|_\infty .
\end{aligned} \tag{239}$$

Step 2: establish the descent equation. Note that Lemma 18 directly applies by replacing $\widehat{Q}^{(t)}$ with $\widehat{q}^{(t)}$:

$$\phi^{(t+1)}(\eta) \le \phi^{(t)}(\eta) + \frac{2\eta}{(1-\gamma)^2}\left\|\widehat{q}^{(t)} - \overline{Q}^{(t)}\right\|_\infty - \eta\left(V^\star(\rho) - \overline{V}^{(t)}(\rho)\right) .$$

To bound the middle term, for all $t \ge 0$, we have

$$\begin{aligned}
\left\|\overline{Q}^{(t)} - \widehat{q}^{(t)}\right\|_\infty &= \left\|\frac{1}{N}\sum_{n=1}^N Q_n^{\pi_n^{(t)}} - \frac{1}{N}\sum_{n=1}^N Q_n^{\overline{\pi}^{(t)}}\right\|_\infty + \frac{1}{N}\left\|\sum_{n=0}^N\left(q_n^{\pi_n^{(t)}} - Q_n^{\pi_n^{(t)}}\right)\right\|_\infty \\
&\le \frac{(1+\gamma)\gamma}{(1-\gamma)^2}\cdot\frac{1}{N}\sum_{n=1}^N\left\|\log\xi_n^{(t)} - \log\overline{\xi}^{(t)}\right\|_\infty + \frac{1}{N}\sum_{n=1}^N e_n \\
&\le \frac{(1+\gamma)\gamma}{(1-\gamma)^2}\left\|u^{(t)}\right\|_\infty + \left\|\boldsymbol{e}\right\|_\infty .
\end{aligned} \tag{240}$$

Hence, (113) is established by combining the above two inequalities.

Step 3: bound the consensus error. Similar to (235), here we have

$$\begin{aligned}
\left\|\boldsymbol{\Omega}^{(t)}\right\|_2 &\le \rho(\boldsymbol{B}(\eta))\left\|\boldsymbol{\Omega}^{(t-1)}\right\|_2 + (d_2(\eta)+c_2(\eta)) \\
&\le \cdots \le \rho^t(\boldsymbol{B}(\eta))\left\|\boldsymbol{\Omega}^{(0)}\right\|_2 + \sum_{i=0}^{t-1}\rho^i(\boldsymbol{B}(\eta))\left(\frac{(1+\gamma)\gamma N\sigma}{(1-\gamma)^4}\eta + \sqrt{N}\sigma\left(\frac{(1+\gamma)\gamma\eta\sqrt{N}}{(1-\gamma)^3}+2\right)\|\boldsymbol{e}\|_\infty\right) \\
&\le \rho^t(\boldsymbol{B}(\eta))\left\|\boldsymbol{\Omega}^{(0)}\right\|_2 + \frac{2}{1-\rho(\boldsymbol{B}(\eta))}\left(\frac{N\sigma}{(1-\gamma)^4}\eta + \sqrt{N}\sigma\left(\frac{\eta\sqrt{N}}{(1-\gamma)^3}+1\right)\|\boldsymbol{e}\|_\infty\right) \\
&\le \left(\frac{3}{8}\sigma+\frac{5}{8}\right)^t\left\|\boldsymbol{\Omega}^{(0)}\right\|_2 + \frac{16}{3(1-\sigma)}\left(\frac{N\sigma}{(1-\gamma)^4}\eta + \sqrt{N}\sigma\left(\frac{\eta\sqrt{N}}{(1-\gamma)^3}+1\right)\|\boldsymbol{e}\|_\infty\right), \qquad (241)
\end{aligned}$$

which indicates (114).

C.6 Proof of Lemma 13

The first claim is easily verified as $\log \xi_n^{(t)}(s,\cdot)$ always deviate from $\log \pi_n^{(t)}(\cdot|s)$ by a global constant shift, as long as it holds for $t=0$:

$$\begin{aligned}
\log \xi_n^{(t+1)}(s,\cdot) &= \sum_{n'=1}^N [W]_{n,n'}\left(\alpha \log \xi_{n'}^{(t)}(s,\cdot) + (1-\alpha)T_n^{(t)}(s,\cdot)/\tau\right) \\
&= \alpha\sum_{n'=1}^N [W]_{n,n'}\left(\alpha\Big(\log \pi_{n'}^{(t)}(s,\cdot) + c_{n'}^{(t)}(s)\mathbf{1}_{|\mathcal{A}|}\Big) + (1-\alpha)T_n^{(t)}(s,\cdot)/\tau\right) \\
&= \alpha\sum_{n'=1}^N [W]_{n,n'}\left(\alpha \log \pi_{n'}^{(t)}(s,\cdot) + (1-\alpha)T_n^{(t)}(s,\cdot)/\tau\right) - \log z_n^{(t)}(s)\mathbf{1}_{|\mathcal{A}|} + c_n^{(t+1)}(s)\mathbf{1}_{|\mathcal{A}|} \\
&= \log \pi_n^{(t+1)}(\cdot|s) + c_n^{(t+1)}(s)\mathbf{1}_{|\mathcal{A}|},
\end{aligned}$$

where $z_n^{(t)}$ is the normalization term (cf. line 5, Algorithm 2) and $\{c_n^{(t)}(s)\}$ are some constants. To prove the second claim, $\forall t\ge 0, \forall (s,a)\in\mathcal{S}\times\mathcal{A}$, let

$$\overline{T}^{(t)}(s,a) \coloneqq \frac{1}{N}\mathbf{1}^\top \boldsymbol{T}^{(t)}(s,a)\,. \qquad (242)$$

Taking inner product with $\frac{1}{N}\mathbf{1}$ for both sides of (28) and using the double stochasticity property of $\boldsymbol{W}$, we get

$$\overline{T}^{(t+1)}(s,a) = \overline{T}^{(t)}(s,a) + \widehat{Q}_\tau^{(t+1)}(s,a) - \widehat{Q}_\tau^{(t)}(s,a)\,. \qquad (243)$$

By the choice of $\boldsymbol{T}^{(0)}$ (line 2 of Algorithm 2), we have $\overline{T}^{(0)} = \widehat{Q}_\tau^{(0)}$ and hence by induction

$$\forall t\ge 0: \quad \overline{T}^{(t)} = \widehat{Q}_\tau^{(t)}\,. \qquad (244)$$

This implies

$$\begin{aligned}
\log \overline{\xi}^{(t+1)}(s,a) - \alpha\log\overline{\xi}^{(t)}(s,a) &= (1-\alpha)\widehat{Q}_\tau^{(t)}(s,a)/\tau \\
&= (1-\alpha)\overline{T}^{(t)}(s,a)/\tau \\
&= \frac{1}{N}\mathbf{1}^\top \log \boldsymbol{\xi}^{(t+1)}(s,a) - \alpha\frac{1}{N}\mathbf{1}^\top\log\boldsymbol{\xi}^{(t)}(s,a).
\end{aligned}$$

Therefore, to prove (172), it suffices to verify the claim for $t=0$:

$$\frac{1}{N}\mathbf{1}^\top\log\boldsymbol{\xi}^{(0)}(s,a) = \log\left\|\exp\left(Q_\tau^\star(s,\cdot)/\tau\right)\right\|_1 + \frac{1}{N}\mathbf{1}^\top\log\boldsymbol{\pi}^{(0)}(a|s) - \log\left\|\exp\left(\frac{1}{N}\sum_{n=1}^N \log\pi_n^{(0)}(\cdot|s)\right)\right\|_1$$

$$= \log \left\| \exp\left(Q_\tau^\star(s,\cdot)/\tau\right) \right\|_1 + \log \overline{\pi}^{(0)}(a|s) = \log \overline{\xi}^{(0)}(s,a)\,.$$

By taking logarithm over both sides of the definition of $\overline{\pi}^{(t+1)}$ (cf. (27)), we get

$$\log \overline{\pi}^{(t+1)}(a|s) = \alpha \log \overline{\pi}^{(t)}(a|s) + (1-\alpha)\widehat{Q}^{(t)}(s,a)/\tau - z^{(t)}(s) \tag{245}$$

for some constant $z^{(t)}(s)$, which deviate from the update rule of $\log \overline{\xi}^{(t+1)}$ by a global constant shift and hence verifies (173).

C.7 Proof of Lemma 15

For notational simplicity, we let $Q_\tau^{\xi'}$ and Q_τ^{ξ} denote $Q_\tau^{\pi_{\xi'}}$ and $Q_\tau^{\pi_\xi}$, respectively. From (7a) we immediately know that to bound $\left\| Q_\tau^{\xi'} - Q_\tau^{\xi} \right\|_\infty$, it suffices to control $\left| V_\tau^{\xi}(s) - V_\tau^{\xi'}(s) \right|$ for each $s \in \mathcal{S}$. By (4) we have

$$\left| V_\tau^{\xi}(s) - V_\tau^{\xi'}(s) \right| \le \left| V^{\xi}(s) - V^{\xi'}(s) \right| + \tau \left| \mathcal{H}(s,\pi_\xi) - \mathcal{H}(s,\pi_{\xi'}) \right|, \tag{246}$$

so in the following we bound both terms in the RHS of (246).

Step 1: bounding $\left|\mathcal{H}(s,\pi_\xi) - \mathcal{H}(s,\pi_{\xi'})\right|$. We first bound $\left|\mathcal{H}(s,\pi_\xi) - \mathcal{H}(s,\pi_{\xi'})\right|$ using the idea in the proof of Lemma 14 in Mei et al. (2020). We let

$$\xi^{(t)} = \xi + t(\xi' - \xi)\,, \quad \forall t \in \mathbb{R}\,, \tag{247}$$

and let $h^{(t)} \in \mathbb{R}^{|\mathcal{S}|}$ be

$$\forall s \in \mathcal{S}: \quad h^{(t)}(s) := -\sum_{a\in\mathcal{A}} \pi_{\xi^{(t)}}(a|s) \log \pi_{\xi^{(t)}}(a|s)\,. \tag{248}$$

Note that $\left\| h^{(t)} \right\|_\infty \le \log|\mathcal{A}|$. We also denote $H^{(t)}: \mathcal{S} \to \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ by:

$$\forall s \in \mathcal{S}: \quad H^{(t)}(s) := \left.\frac{\partial \pi_\xi(\cdot|s)}{\partial \xi}\right|_{\xi=\xi^{(t)}} = \mathrm{diag}\{\pi_{\xi^{(t)}}(\cdot|s)\} - \pi_{\xi^{(t)}}(\cdot|s)\pi_{\xi^{(t)}}(\cdot|s)^\top\,, \tag{249}$$

then we have

$$\begin{aligned}
\forall s \in \mathcal{S}: \quad \left|\frac{dh^{(t)}(s)}{dt}\right| &= \left|\left\langle \frac{\partial h^{(t)}(s)}{\partial \xi^{(t)}(\cdot|s)}, \xi'(s,\cdot) - \xi(s,\cdot) \right\rangle\right| \\
&= \left|\left\langle H^{(t)}(s) \log \pi_{\xi^{(t)}}(\cdot|s), \xi'(s,\cdot) - \xi(s,\cdot) \right\rangle\right| \\
&\le \left\| H^{(t)}(s) \log \pi_{\xi^{(t)}}(\cdot|s) \right\|_1 \left\| \xi'(s,\cdot) - \xi(s,\cdot) \right\|_\infty\,,
\end{aligned} \tag{250}$$

where $\frac{\partial h^{(t)}(s)}{\partial \xi^{(t)}(\cdot|s)}$ stands for $\frac{\partial h^{(t)}(s)}{\partial \xi(\cdot|s)}\big|_{\xi=\xi^{(t)}}$. The first term in (250) is further upper bounded as

$$\begin{aligned}
\left\| H^{(t)}(s) \log \pi_{\xi^{(t)}}(\cdot|s) \right\|_1 &= \sum_{a\in\mathcal{A}} \pi_{\xi^{(t)}}(a|s) \left| \log \pi_{\xi^{(t)}}(a|s) - \pi_{\xi^{(t)}}(\cdot|s)^\top \log \pi_{\xi^{(t)}}(\cdot|s) \right| \\
&\le \sum_{a\in\mathcal{A}} \pi_{\xi^{(t)}}(a|s) \left(\left|\log \pi_{\xi^{(t)}}(a|s)\right| + \left|\pi_{\xi^{(t)}}(\cdot|s)^\top \log \pi_{\xi^{(t)}}(\cdot|s)\right| \right) \\
&= -2 \sum_{a\in\mathcal{A}} \pi_{\xi^{(t)}}(a,s) \log \pi_{\xi^{(t)}}(a|s) \le 2\log|\mathcal{A}|\,.
\end{aligned}$$

By Lagrange mean value theorem, there exists $t \in (0,1)$ such that

$$|h_1(s) - h_0(s)| = \left|\frac{dh^{(t)}(s)}{dt}\right| \le 2\log|\mathcal{A}| \left\| \xi'(s,\cdot) - \xi(s,\cdot) \right\|_\infty\,,$$

where the inequality follows from (250) and the above inequality. Combining (5) with the above inequality, we arrive at

$$\left|\mathcal{H}(s,\pi_\xi) - \mathcal{H}(s,\pi_{\xi'})\right| \le \frac{2\log|\mathcal{A}|}{1-\gamma} \left\| \xi' - \xi \right\|_\infty\,. \tag{251}$$

Step 2: bounding $\left|V^{\xi}(s)-V^{\xi'}(s)\right|$. Similar to the previous proof, we bound $\left|V^{\xi}(s)-V^{\xi'}(s)\right|$ by bounding $\left|\frac{dV^{\xi^{(t)}}}{dt}(s)\right|$. By Bellman's consistency equation, the value function of $\pi_{\xi^{(t)}}$ is given by

$$V^{\xi^{(t)}}(s) = \sum_{a\in\mathcal{A}} \pi_{\xi^{(t)}}(a|s) r(s,a) + \gamma \sum_{a} \pi_{\xi_\alpha}(a|s) \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a) V^{\xi^{(t)}}(s')\,,$$

which can be represented in a matrix-vector form as

$$V^{\xi_\star^{(t)}}(s) = e_s^\top \boldsymbol{M}_t r_t\,, \tag{252}$$

where $e_s \in \mathbb{R}^{|\mathcal{S}|}$ is a one-hot vector whose s-th entry is 1,

$$\boldsymbol{M}_t \coloneqq (\boldsymbol{I} - \gamma \boldsymbol{P}_t)^{-1}\,, \tag{253}$$

with $\boldsymbol{P}_t \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denoting the induced state transition matrix by $\pi_{\xi^{(t)}}$

$$\boldsymbol{P}_t(s,s') = \sum_{a\in\mathcal{A}} \pi_{\xi^{(t)}}(a|s)\mathcal{P}(s'|s,a)\,, \tag{254}$$

and $r_t \in \mathbb{R}^{|\mathcal{S}|}$ is given by

$$\forall s\in\mathcal{S}: \quad r_t(s) \coloneqq \sum_{a\in\mathcal{A}} \pi_{\xi^{(t)}}(a|s) r(s,a)\,. \tag{255}$$

Taking derivative w.r.t. t in (252), we obtain (Petersen and Pedersen, 2008)

$$\frac{dV^{\xi^{(t)}}(s)}{dt} = \gamma \cdot e_s^\top \boldsymbol{M}_t \frac{d\boldsymbol{P}_t}{dt} \boldsymbol{M}_t r_t + e_s^\top \boldsymbol{M}_t \frac{dr_t}{dt}\,. \tag{256}$$

We now calculate each term respectively.

- For the first term, it follows that

$$\begin{aligned}
\left|\gamma \cdot e_s^\top \boldsymbol{M}_t \frac{d\boldsymbol{P}_t}{dt} \boldsymbol{M}_t r_t\right| &\le \gamma \left\|\boldsymbol{M}_t \frac{d\boldsymbol{P}_t}{dt} \boldsymbol{M}_t r_t\right\|_\infty \\
&\le \frac{\gamma}{1-\gamma} \left\|\frac{d\boldsymbol{P}_t}{dt} \boldsymbol{M}_t r_t\right\|_\infty \\
&\le \frac{2\gamma}{1-\gamma} \left\|\boldsymbol{M}_t r_t\right\|_\infty \left\|\xi'-\xi\right\|_\infty \qquad (257)\\
&\le \frac{2\gamma}{(1-\gamma)^2} \left\|r_t\right\|_\infty \left\|\xi'-\xi\right\|_\infty \\
&\le \frac{2\gamma}{(1-\gamma)^2} \left\|\xi'-\xi\right\|_\infty\,. \qquad (258)
\end{aligned}$$

  where the second and fourth lines use the fact $\|\boldsymbol{M}_t\|_1 \le 1/(1-\gamma)$ (Li et al., 2023b, Lemma 10), and the last line follow from

$$\|r_t\|_\infty = \max_{s\in\mathcal{S}} \left|\sum_{a\in\mathcal{A}} \pi_{\xi^{(t)}}(a|s) r(s,a)\right| \le 1.$$

  We defer the proof of (257) to the end of proof.

- For the second term, it follows that

$$\left|e_s^\top \boldsymbol{M}_t \frac{dr_t}{dt}\right| \le \frac{1}{1-\gamma} \left\|\frac{dr_t}{dt}\right\|_\infty \le \frac{1}{1-\gamma} \left\|\xi'-\xi\right\|_\infty\,. \tag{259}$$

where the first inequality follows again from $\|\boldsymbol{M}_t\|_1 \le 1/(1-\gamma)$, and the second inequality follows from

$$\begin{aligned}
\left\|\frac{dr_t}{dt}\right\|_\infty = \max_{s\in\mathcal{S}}\left|\frac{dr_t(s)}{dt}\right| &= \max_{s\in\mathcal{S}}\left|\left\langle \frac{\partial \pi_{\xi^{(t)}}(\cdot|s)^\top r(s,\cdot)}{\partial \xi^{(t)}(s,\cdot)}, \xi'(s,\cdot)-\xi(s,\cdot)\right\rangle\right| \\
&\le \max_{s\in\mathcal{S}}\left\|\frac{\partial \pi_{\xi^{(t)}}(\cdot|s)^\top}{\partial \xi^{(t)}(s,\cdot)} r(s,\cdot)\right\|_1 \|\xi'(s,\cdot)-\xi(s,\cdot)\|_\infty \\
&= \max_{s\in\mathcal{S}}\left(\sum_{a\in\mathcal{A}} \pi_{\xi^{(t)}}(a|s)\left|r(s,a)-\pi_{\xi^{(t)}}(\cdot|s)^\top r(s,\cdot)\right|\right)\|\xi'(s,\cdot)-\xi(s,\cdot)\|_\infty \\
&\le \max_{s\in\mathcal{S}} \underbrace{\max_{a\in\mathcal{A}}\left|r(s,a)-\pi_{\xi^{(t)}}(\cdot|s)^\top r(s,\cdot)\right|}_{\le 1 \text{ since } r(s,a)\in[0,1]} \|\xi'(s,\cdot)-\xi(s,\cdot)\|_\infty \\
&\le \max_{s\in\mathcal{S}} \|\xi'(s,\cdot)-\xi(s,\cdot)\|_\infty = \|\xi'-\xi\|_\infty . \qquad (260)
\end{aligned}$$

Plugging the above two inequalities into (256) and using Lagrange mean value theorem, we have

$$\left|V^\xi(s)-V^{\xi'}(s)\right| \le \frac{1+\gamma}{(1-\gamma)^2}\|\xi'-\xi\|_\infty . \qquad (261)$$

Step 3: sum up. Combining (261), (251) and (246), we have

$$\forall s\in\mathcal{S}: \quad \left|V_\tau^\xi(s)-V_\tau^{\xi'}(s)\right| \le \frac{1+\gamma+2\tau(1-\gamma)\log|\mathcal{A}|}{(1-\gamma)^2}\|\log\pi-\log\pi'\|_\infty . \qquad (262)$$

Combining (262) and (7a), (181) immediately follows.

Proof of (257). For any vector $x\in\mathbb{R}^{|\mathcal{S}|}$, we have

$$\left[\frac{d\boldsymbol{P}_t}{dt}x\right]_s = \sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{d\pi_{\xi^{(t)}}(a|s)}{dt}\mathcal{P}(s'|s,a)x(s'),$$

from which we can bound the l_∞ norm as

$$\begin{aligned}
\left\|\frac{d\boldsymbol{P}_t}{dt}x\right\|_\infty &\le \max_s \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\mathcal{P}(s'|s,a)\left|\frac{d\pi_{\xi^{(t)}}(a|s)}{dt}\right|\|x\|_\infty \\
&= \max_s \sum_{a\in\mathcal{A}}\left|\frac{d\pi_{\xi^{(t)}}(a|s)}{dt}\right|\|x\|_\infty \\
&\le 2\|\xi'-\xi\|_\infty\|x\|_\infty \qquad (263)
\end{aligned}$$

as desired, where the last line follows from the following fact:

$$\begin{aligned}
\sum_{a\in\mathcal{A}}\left|\frac{d\pi_{\xi^{(t)}}(a|s)}{dt}\right| &= \sum_{a\in\mathcal{A}}\left|\left\langle\frac{\partial\pi_{\xi^{(t)}}(a|s)}{\partial\xi^{(t)}}, \xi'-\xi\right\rangle\right| \\
&= \sum_{a\in\mathcal{A}}\left|\left\langle\frac{\partial\pi_{\xi^{(t)}}(a|s)}{\partial\xi^{(t)}(s,\cdot)}, \xi'(s,\cdot)-\xi(s,\cdot)\right\rangle\right| \\
&= \sum_{a\in\mathcal{A}}\pi_{\xi^{(t)}}(a|s)\left|(\xi'(s,a)-\xi(s,a))-\pi_{\xi^{(t)}}(\cdot|s)^\top(\xi'(s,\cdot)-\xi(s,\cdot))\right| \\
&\le \max_a|\xi'(s,a)-\xi(s,a)|+\left|\pi_{\xi^{(t)}}(\cdot|s)^\top(\xi'(s,\cdot)-\xi(s,\cdot))\right| \\
&\le 2\|\xi'-\xi\|_\infty .
\end{aligned}$$

C.8 Proof of Lemma 16

To simplify the notation, we denote

$$\delta^{(t)} := \widehat{Q}_\tau^{(t)} - \overline{Q}_\tau^{(t)} . \tag{264}$$

We first rearrange the terms of (245) and obtain

$$-\tau \log \overline{\pi}^{(t)}(a|s) + \left(\overline{Q}_\tau^{(t)}(s,a) + \delta^{(t)}(s,a)\right) = \frac{1-\gamma}{\eta}\left(\log \overline{\pi}^{(t+1)}(a|s) - \log \overline{\pi}^{(t)}(a|s)\right) + \frac{1-\gamma}{\eta} z^{(t)}(s) . \tag{265}$$

This in turn allows us to express $\overline{V}_\tau^{(t)}(s_0)$ for any $s_0 \in \mathcal{S}$ as follows

$$\begin{aligned}
\overline{V}_\tau^{(t)}(s_0) &= \mathop{\mathbb{E}}_{a_0 \sim \overline{\pi}^{(t)}(\cdot|s_0)} \left[-\tau \log \overline{\pi}^{(t)}(a_0|s_0) + \overline{Q}_\tau^{(t)}(s_0,a_0)\right] \\
&= \mathop{\mathbb{E}}_{a_0 \sim \overline{\pi}^{(t)}(\cdot|s_0)} \left[\frac{1-\gamma}{\eta} z^{(t)}(s_0)\right] + \mathop{\mathbb{E}}_{a_0 \sim \overline{\pi}^{(t)}(\cdot|s_0)} \left[\frac{1-\gamma}{\eta}\left(\log \overline{\pi}^{(t+1)}(a_0|s_0) - \log \overline{\pi}^{(t)}(a_0|s_0)\right) - \delta^{(t)}(s_0,a_0)\right] \\
&= \frac{1-\gamma}{\eta} z^{(t)}(s_0) - \frac{1-\gamma}{\eta} \mathsf{KL}\big(\overline{\pi}^{(t)}(\cdot|s_0) \,\|\, \overline{\pi}^{(t+1)}(\cdot|s_0)\big) - \mathop{\mathbb{E}}_{a_0 \sim \overline{\pi}^{(t)}(\cdot|s_0)} \left[\delta^{(t)}(s_0,a_0)\right] \\
&= \mathop{\mathbb{E}}_{a_0 \sim \overline{\pi}^{(t+1)}(\cdot|s_0)} \left[\frac{1-\gamma}{\eta} z^{(t)}(s_0)\right] - \frac{1-\gamma}{\eta} \mathsf{KL}\big(\overline{\pi}^{(t)}(\cdot|s_0) \,\|\, \overline{\pi}^{(t+1)}(\cdot|s_0)\big) - \mathop{\mathbb{E}}_{a_0 \sim \overline{\pi}^{(t)}(\cdot|s_0)} \left[\delta^{(t)}(s_0,a_0)\right] ,
\end{aligned} \tag{266}$$

where the first identity makes use of (7b), the second line follows from (265). Invoking (7b) again to rewrite the $z(s_0)$ appearing in the first term of (266), we reach

$$\begin{aligned}
&\overline{V}_\tau^{(t)}(s_0) \\
&= \mathop{\mathbb{E}}_{a_0 \sim \overline{\pi}^{(t+1)}(\cdot|s_0)} \left[-\tau \log \overline{\pi}^{(t+1)}(a_0|s_0) + \overline{Q}_\tau^{(t)}(s_0,a_0) + \left(\tau - \frac{1-\gamma}{\eta}\right)\left(\log \overline{\pi}^{(t+1)}(a_0|s_0) - \log \overline{\pi}^{(t)}(a|s)\right)\right] \\
&\quad - \frac{1-\gamma}{\eta} \mathsf{KL}\big(\overline{\pi}^{(t)}(\cdot|s_0) \,\|\, \overline{\pi}^{(t+1)}(\cdot|s_0)\big) - \mathop{\mathbb{E}}_{a_0 \sim \overline{\pi}^{(t)}(\cdot|s_0)} \left[\delta^{(t)}(s_0,a_0)\right] + \mathop{\mathbb{E}}_{a_0 \sim \overline{\pi}^{(t+1)}(\cdot|s_0)} \left[\delta^{(t)}(s_0,a_0)\right] \\
&= \mathop{\mathbb{E}}_{\substack{a_0 \sim \overline{\pi}^{(t+1)}(\cdot|s_0), \\ s_1 \sim P(\cdot|s_0,a_0)}} \left[-\tau \log \overline{\pi}^{(t+1)}(a_0|s_0) + r(s_0,a_0) + \gamma \overline{V}_\tau^{(t)}(s_0)\right] \\
&\quad - \left(\frac{1-\gamma}{\eta} - \tau\right) \mathsf{KL}\big(\overline{\pi}^{(t+1)}(\cdot|s_0) \,\|\, \overline{\pi}^{(t)}(\cdot|s_0)\big) - \frac{1-\gamma}{\eta} \mathsf{KL}\big(\overline{\pi}^{(t)}(\cdot|s_0) \,\|\, \overline{\pi}^{(t+1)}(\cdot|s_0)\big) \\
&\quad - \mathop{\mathbb{E}}_{a_0 \sim \overline{\pi}^{(t)}(\cdot|s_0)} \left[\delta^{(t)}(s_0,a_0)\right] + \mathop{\mathbb{E}}_{a_0 \sim \overline{\pi}^{(t+1)}(\cdot|s_0)} \left[\delta^{(t)}(s_0,a_0)\right] .
\end{aligned} \tag{267}$$

Note that for any $(s_0,a_0) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}
&- \mathop{\mathbb{E}}_{a_0 \sim \overline{\pi}^{(t)}(\cdot|s_0)} \left[\delta^{(t)}(s_0,a_0)\right] + \mathop{\mathbb{E}}_{a_0 \sim \overline{\pi}^{(t+1)}(\cdot|s_0)} \left[\delta^{(t)}(s_0,a_0)\right] \\
&= \sum_{a_0 \in \mathcal{A}} \left(\overline{\pi}^{(t+1)}(a_0|s_0) - \overline{\pi}^{(t)}(a_0|s_0)\right) \delta^{(t)}(s_0,a_0) \\
&\le \left\|\overline{\pi}^{(t+1)}(\cdot|s_0) - \overline{\pi}^{(t)}(\cdot|s_0)\right\|_1 \left\|\delta^{(t)}\right\|_\infty \le 2\left\|\delta^{(t)}\right\|_\infty .
\end{aligned} \tag{268}$$

To finish up, applying (267) recursively to expand $\overline{V}_\tau^{(t)}(s_i)$, $i \ge 1$ and making use of (268), we arrive at

$$\begin{aligned}
&\overline{V}_\tau^{(t)}(s_0) \\
&\le \sum_{i=1}^{\infty} \gamma^i \cdot 2 \left\|\delta^{(t)}\right\|_\infty + \mathop{\mathbb{E}}_{\substack{a_i \sim \overline{\pi}^{(t+1)}(\cdot|s_i), \\ s_{i+1} \sim P(\cdot|s_i,a_i), \forall i \ge 0}} \left[\sum_{i=1}^{\infty} \gamma^i \left\{r(s_i,a_i) - \tau \log \overline{\pi}^{(t+1)}(a_i|s_i)\right\}\right.
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{i=1}^{\infty} \gamma^i \left\{ \left(\frac{1-\gamma}{\eta} - \tau \right) \mathsf{KL}\big(\overline{\pi}^{(t+1)}(\cdot|s_i) \,\|\, \overline{\pi}^{(t)}(\cdot|s_i)\big) + \frac{1-\gamma}{\eta} \mathsf{KL}\big(\overline{\pi}^{(t)}(\cdot|s_i) \,\|\, \overline{\pi}^{(t+1)}(\cdot|s_i)\big) \right\} \Bigg] \\
&= \frac{2}{1-\gamma} \left\| \delta^{(t)} \right\|_\infty + \overline{V}_\tau^{(t+1)}(s_0) \\
&\quad - \mathop{\mathbb{E}}_{s \sim d_{s_0}^{\overline{\pi}^{(t+1)}}} \left[\left(\frac{1}{\eta} - \frac{\tau}{1-\gamma} \right) \mathsf{KL}\big(\overline{\pi}^{(t+1)}(\cdot|s_i) \,\|\, \overline{\pi}^{(t)}(\cdot|s_i)\big) + \frac{1}{\eta} \mathsf{KL}\big(\overline{\pi}^{(t)}(\cdot|s_i) \,\|\, \overline{\pi}^{(t+1)}(\cdot|s_i)\big) \right], \qquad (269)
\end{aligned}
$$

where the third line follows since $\overline{V}_\tau^{(t+1)}$ can be viewed as the value function of $\overline{\pi}^{(t+1)}$ with adjusted rewards $\overline{r}^{(t+1)}(s,a) \coloneqq r(s,a) - \tau \log \overline{\pi}^{(t+1)}(s|a)$. And (199) follows immediately from the above inequality (269). By (7a) we can easily see that (200) is a consequence of (199).

C.9 Proof of Lemma 18

We first introduce the famous performance difference lemma which will be used in our proof.

Lemma 19 (Performance difference lemma)**.** *For any policy $\pi, \pi' \in \Delta(\mathcal{A})^{\mathcal{S}}$ and $\rho \in \Delta(\mathcal{S})$, we have*

$$
\begin{aligned}
V^\pi(\rho) - V^{\pi'}(\rho) &= \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim \bar{d}^\pi} \left[A^{\pi'}(s,a) \right] \qquad (270) \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^\pi} \left[\langle Q^{\pi'}(s), \pi(s) - \pi'(s) \rangle \right]. \qquad (271)
\end{aligned}
$$

Proof. See Lemma 3 in Yuan et al. (2022). □

For all $t \geq 0$, we define the advantage function $\overline{A}^{(t)}$ as:

$$
\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \overline{A}^{(t)}(s,a) \coloneqq \overline{Q}^{(t)}(s,a) - \overline{V}^{(t)}(s). \qquad (272)
$$

Then for Alg. 1, the update rule of $\overline{\pi}$ (Eq. (245)) can be written as

$$
\log \overline{\pi}^{(t+1)}(a|s) = \log \overline{\pi}^{(t)}(a|s) + \frac{\eta}{1-\gamma} \left(\overline{A}^{(t)}(s,a) + \delta^{(t)}(s,a) \right) - \log \widehat{z}^{(t)}(s), \qquad (273)
$$

where $\delta^{(t)}$ is defined in (264) and

$$
\begin{aligned}
\log \widehat{z}^{(t)}(s) &= \log \sum_{a' \in \mathcal{A}} \overline{\pi}^{(t)}(a'|s) \exp \left\{ \frac{\eta}{1-\gamma} \left(\overline{A}^{(t)}(s,a') + \delta^{(t)}(s,a') \right) \right\} \\
&\geq \sum_{a' \in \mathcal{A}} \overline{\pi}^{(t)}(a'|s) \log \exp \left\{ \frac{\eta}{1-\gamma} \left(\overline{A}^{(t)}(s,a') + \delta^{(t)}(s,a') \right) \right\} \\
&= \frac{\eta}{1-\gamma} \sum_{a' \in \mathcal{A}} \overline{\pi}^{(t)}(a'|s) \left(\overline{A}^{(t)}(s,a') + \delta^{(t)}(s,a') \right) \\
&= \frac{\eta}{1-\gamma} \sum_{a' \in \mathcal{A}} \overline{\pi}^{(t)}(a'|s) \delta^{(t)}(s,a') \geq -\frac{\eta}{1-\gamma} \left\| \delta^{(t)} \right\|_\infty, \qquad (274)
\end{aligned}
$$

where the first inequality follows by Jensen's inequality on the concave function $\log x$ and the last equality uses $\sum_{a' \in \mathcal{A}} \overline{\pi}^{(t)}(a'|s) \overline{A}^{(t)}(s,a') = 0$.

For all starting state distribution μ, we use $d^{(t+1)}$ as shorthand for $d_\mu^{\overline{\pi}^{(t+1)}}$, the performance difference lemma (Lemma 19) implies:

$$
\begin{aligned}
&\overline{V}^{(t+1)}(\mu) - \overline{V}^{(t)}(\mu) \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{(t+1)}} \sum_{a \in \mathcal{A}} \overline{\pi}^{(t+1)}(a|s) \left(\overline{A}^{(t)}(s,a) + \delta^{(t)}(s,a) \right) - \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{(t+1)}} \mathbb{E}_{a \sim \overline{\pi}^{(t+1)}(\cdot|s)} \left[\delta^{(t)}(s,a) \right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^{(t+1)}} \sum_{a\in\mathcal{A}} \overline{\pi}^{(t+1)}(a|s) \log \frac{\overline{\pi}^{(t+1)}(a|s)\widehat{z}^{(t)}(s)}{\overline{\pi}^{(t)}(a|s)} - \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{(t+1)}}\mathbb{E}_{a\sim\overline{\pi}^{(t+1)}(\cdot|s)}\left[\delta^{(t)}(s,a)\right] \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^{(t+1)}} \mathsf{KL}\big(\overline{\pi}^{(t+1)}(\cdot|s) \,\|\, \overline{\pi}^{(t)}(\cdot|s)\big) + \frac{1}{\eta}\mathbb{E}_{s\sim d^{(t+1)}} \log \widehat{z}^{(t)}(s) - \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{(t+1)}}\mathbb{E}_{a\sim\overline{\pi}^{(t+1)}(\cdot|s)}\left[\delta^{(t)}(s,a)\right] \\
&\geq \frac{1}{\eta}\mathbb{E}_{s\sim d^{(t+1)}}\left(\log \widehat{z}^{(t)}(s) + \frac{\eta}{1-\gamma}\big\|\delta^{(t)}\big\|_\infty\right) - \frac{2}{1-\gamma}\big\|\delta^{(t)}\big\|_\infty ,
\end{aligned}
$$

from which we can see that

$$
\overline{V}^{(t+1)}(\mu) - \overline{V}^{(t)}(\mu) \geq -\frac{2}{1-\gamma}\big\|\delta^{(t)}\big\|_\infty , \tag{275}
$$

where we use (274), and that

$$
\overline{V}^{(t+1)}(\mu) - \overline{V}^{(t)}(\mu) \geq \frac{1-\gamma}{\eta}\mathbb{E}_{s\sim\mu}\left(\log \widehat{z}^{(t)}(s) + \frac{\eta}{1-\gamma}\big\|\delta^{(t)}\big\|_\infty\right) - \frac{2}{1-\gamma}\big\|\delta^{(t)}\big\|_\infty , \tag{276}
$$

which follows from $d^{(t+1)} = d_\mu^{\overline{\pi}^{(t+1)}} \geq (1-\gamma)\mu$ and the fact that $\log \widehat{z}^{(t)}(s) + \frac{\eta}{1-\gamma}\big\|\delta^{(t)}\big\|_\infty \geq 0$ (by (274)).

For any fixed ρ, we use $d^\star$ as shorthand for $d_\rho^{\pi^\star}$. By the performance difference lemma (Lemma 19),

$$
\begin{aligned}
&V^\star(\rho) - \overline{V}^{(t)}(\rho) \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\star}\sum_{a\in\mathcal{A}} \pi^\star(a|s)\left(\overline{A}^{(t)}(s,a) + \delta^{(t)}(s,a)\right) - \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\star}\mathbb{E}_{a\sim\pi^\star(\cdot|s)}\left[\delta^{(t)}(s,a)\right] \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\sum_{a\in\mathcal{A}} \pi^\star(a|s)\log\frac{\overline{\pi}^{(t+1)}(a|s)\widehat{z}^{(t)}(s)}{\overline{\pi}^{(t)}(a|s)} - \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\star}\mathbb{E}_{a\sim\pi^\star(\cdot|s)}\left[\delta^{(t)}(s,a)\right] \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\left(\mathsf{KL}\big(\pi^\star(\cdot|s)\,\|\,\overline{\pi}^{(t)}(\cdot|s)\big) - \mathsf{KL}\big(\pi^\star(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\big) + \log\widehat{z}^{(t)}(s)\right) - \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\star}\mathbb{E}_{a\sim\pi^\star(\cdot|s)}\left[\delta^{(t)}(s,a)\right] \\
&\leq \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\left(\mathsf{KL}\big(\pi^\star(\cdot|s)\,\|\,\overline{\pi}^{(t)}(\cdot|s)\big) - \mathsf{KL}\big(\pi^\star(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\big) + \left(\log\widehat{z}^{(t)}(s) + \frac{\eta}{1-\gamma}\big\|\delta^{(t)}\big\|_\infty\right)\right) ,
\end{aligned} \tag{277}
$$

where we use (273) in the second equality.

By applying (276) with $\mu = d^\star$ as the initial state distribution, we have

$$
\frac{1}{\eta}\mathbb{E}_{s\sim\mu}\Big(\log\widehat{z}^{(t)}(s) + \frac{\eta}{1-\gamma}\big\|\delta^{(t)}\big\|_\infty\Big) \leq \frac{1}{1-\gamma}\Big(\overline{V}^{(t+1)}(d^\star) - \overline{V}^{(t)}(d^\star)\Big) + \frac{2}{(1-\gamma)^2}\big\|\delta^{(t)}\big\|_\infty .
$$

Plugging the above equation into (277), we obtain

$$
\begin{aligned}
V^\star(\rho) - \overline{V}^{(t)}(\rho) \leq{}& \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\Big(\mathsf{KL}\big(\pi^\star(\cdot|s)\,\|\,\overline{\pi}^{(t)}(\cdot|s)\big) - \mathsf{KL}\big(\pi^\star(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\big)\Big) \\
&+ \frac{1}{1-\gamma}\Big(\overline{V}^{(t+1)}(d^\star) - \overline{V}^{(t)}(d^\star)\Big) + \frac{2}{(1-\gamma)^2}\big\|\delta^{(t)}\big\|_\infty ,
\end{aligned}
$$

which gives Lemma 18.

D Proof of key lemmas for FedNAC

D.1 Proof of Lemma 10

For notational simplicity we let $V^\xi, V^{\xi'}$ denote $V^{f_\xi}, V^{f_{\xi'}}$, respectively. Same as in Lemma 15, we define $\xi^{(t)} = \xi + t(\xi' - \xi)$ and define $\boldsymbol{P}_t, \boldsymbol{M}_t, r_t$ by replacing $\pi_{\xi^{(t)}}$ with $f_{\xi^{(t)}}$ in (254),(253) and (255), respectively. Define

$$
\bar{\phi}_\xi(s,a) = \phi(s,a) - \mathbb{E}_{a'\sim f_{\xi^{(t)}}}[\phi(s,a')],
$$

then we have

$$\frac{\partial f_\xi(a|s)}{\partial \xi} = f_\xi(a|s)\bar{\phi}_\xi(s,a)\,. \tag{278}$$

Analogous to (263), we have

$$\begin{aligned}
\left\|\frac{d\boldsymbol{P}_t}{dt}x\right\|_\infty &\le \max_s \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a)\left|\frac{d\pi_{\xi^{(t)}}(a|s)}{dt}\right| \|x\|_\infty \\
&= \max_s \sum_{a\in\mathcal{A}} \left|\frac{d\pi_{\xi^{(t)}}(a|s)}{dt}\right| \|x\|_\infty \\
&\le 2C_\phi \|\xi'-\xi\|_2 \|x\|_\infty\,,
\end{aligned}$$

where the last line is due to

$$\begin{aligned}
\sum_{a\in\mathcal{A}}\left|\frac{df_{\xi^{(t)}}(a|s)}{dt}\right| &= \sum_{a\in\mathcal{A}}\left|\left\langle \frac{\partial f_{\xi^{(t)}}(a|s)}{\partial \xi^{(t)}}, \xi'-\xi\right\rangle\right| \\
&= \sum_{a\in\mathcal{A}} f_{\xi^{(t)}}(a|s)\left|\langle \bar{\phi}_\xi(s,a), \xi'-\xi\rangle\right| \\
&\le \sum_{a\in\mathcal{A}} f_{\xi^{(t)}}(a|s)\left\|\bar{\phi}_\xi(s,a)\right\|_2 \|\xi'-\xi\|_2 \\
&\le 2C_\phi \|\xi'-\xi\|_\infty\,.
\end{aligned}$$

Same as (256) in Lemma 15, we have

$$\frac{dV^{\xi^{(t)}}(s)}{dt} = \gamma\cdot e_s^\top \boldsymbol{M}_t \frac{d\boldsymbol{P}_t}{dt}\boldsymbol{M}_t r_t + e_s^\top \boldsymbol{M}_t \frac{dr_t}{dt}\,. \tag{279}$$

And similar to (260), we deduce

$$\begin{aligned}
\left\|\frac{dr_t}{dt}\right\|_\infty = \max_{s\in\mathcal{S}}\left|\frac{dr_t(s)}{dt}\right| &= \max_{s\in\mathcal{S}}\left|\left\langle \frac{\partial f_{\xi^{(t)}}(\cdot|s)^\top r(s,\cdot)}{\partial \xi^{(t)}}, \xi'-\xi\right\rangle\right| \\
&= \left|\langle \sum_{a\in\mathcal{A}} f_\xi(a|s)\bar{\phi}_\xi(s,a) r(s,a), \xi'-\xi\rangle\right| \\
&= \sum_{a\in\mathcal{A}} f_\xi(a|s) r(s,a)\left|\langle \bar{\phi}_\xi(s,a), \xi'-\xi\rangle\right| \\
&\le 2C_\phi \|\xi'-\xi\|_2\,,
\end{aligned}$$

which gives

$$\left|e_s^\top \boldsymbol{M}_t \frac{dr_t}{dt}\right| \le \frac{1}{1-\gamma}\left\|\frac{dr_t}{dt}\right\|_\infty \le \frac{2C_\phi}{1-\gamma}\|\xi'-\xi\|_2\,. \tag{280}$$

Following the same steps in (258), we deduce

$$\left|\gamma\cdot e_s^\top \boldsymbol{M}_t \frac{d\boldsymbol{P}_t}{dt}\boldsymbol{M}_t r_t\right| \le \frac{2\gamma C_\phi}{(1-\gamma)^2}\|\xi'-\xi\|_2\,. \tag{281}$$

Combining the above two expressions (280) and (281) with (279), we deduce

$$|V^\xi(s) - V^{\xi'}(s)| \le \frac{2C_\phi(1+\gamma)}{(1-\gamma)^2}\|\xi'-\xi\|_2\,, \tag{282}$$

which implies

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \quad |Q^\xi(s,a) - Q^{\xi'}(s,a)| \le \frac{2C_\phi\gamma(1+\gamma)}{(1-\gamma)^2}\|\xi'-\xi\|_2\,. \tag{283}$$

D.2 Proof of Lemma 11

This proof is inspired by the proof of Theorem 1 in Yuan et al. (2022). To give the proof, we first introduce the following three-point descent lemma.

Lemma 20 (Three-point descent lemma, Lemma 6 in Xiao (2022))**.** *Suppose that $\mathcal{C} \subset \mathbb{R}^m$ is a closed convex set, $g : \mathcal{C} \to \mathbb{R}$ is a proper, closed, convex function, $D_h(\cdot,\cdot)$ is the Bregman divergence generated by a function h of Lengendre type and rint dom$h \cap \mathcal{C} \neq \emptyset$. For any $x \in$ rintdom$h$, let*

$$x^+ \in \arg\min_{u \in domh \cap \mathcal{C}} \{f(u) + D_h(u,x)\},$$

then $x^+ \in domh \cap \mathcal{C}$ and for any $u \in domh \cap \mathcal{C}$, it holds that

$$f(x^+) + D_h(x^+, x) \leq f(u) + D_h(u,x) - D_h(u, x^+). \tag{284}$$

By the update rule (125) and the parameterization (38) we know know that

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \bar{f}^{(t+1)}(a|s) = \frac{1}{Z^{(t)}(s)} f^{(t)}(a|s) \exp\left(\alpha \phi^\top(s,a)\hat{w}^{(t)}\right),$$

where $Z^{(t)}(s)$ is a normalization coefficient to ensure $\sum_{a\in\mathcal{A}} f^{(t+1)}(s,a) = 1$ for each $s \in \mathcal{S}$. Note that the above $\pi^{(t+1)}$ could also be obtained by a mirror descent update:

$$\forall s \in \mathcal{S}: \quad f^{(t+1)}(\cdot|s) = \arg\min_{g \in \Delta(\mathcal{A})} \left\{-\alpha\langle \Phi(s)\hat{w}^{(t)}, g\rangle + \mathsf{KL}\big(g \,\|\, f^{(t)}(\cdot|s)\big)\right\}, \tag{285}$$

where $\Phi(s) \in \mathbb{R}^{|\mathcal{A}|\times p}$ is a matrix with rows $\phi^\top(s,a) \in \mathbb{R}^p$ for $a \in \mathcal{A}$.

We apply the three-point descent lemma (cf. Lemma 20) with $\mathcal{C} = \Delta(\mathcal{A})$, $f = -\alpha\langle \Phi(s)\hat{w}^{(t)}, \cdot\rangle$ and $h : \Delta(\mathcal{A}) \to \mathbb{R}$ is the negative entropy with $h(q) = \sum_{a\in\mathcal{A}} q(a)\log q(a)$ and deduce that for any $q \in \Delta(\mathcal{A})$, we have

$$-\alpha\langle \Phi(s)\hat{w}^{(t)}, \bar{f}^{(t+1)}(\cdot|s)\rangle + D\left(\bar{f}^{(t+1)}(\cdot|s), \bar{f}^{(t)}(\cdot|s)\right) \leq -\alpha\langle \Phi(s)\hat{w}^{(t)}, q\rangle + D\left(q, \bar{f}^{(t)}(\cdot|s)\right) - D\left(q, \bar{f}^{(t+1)}(\cdot|s)\right).$$

Rearranging terms and dividing both sides by $-\alpha$, we obtain

$$\langle \Phi(s)\hat{w}^{(t)}, \bar{f}^{(t+1)}(\cdot|s) - q\rangle - \frac{1}{\alpha} D\left(\bar{f}^{(t+1)}(\cdot|s), \bar{f}^{(t)}(\cdot|s)\right) \geq -\frac{1}{\alpha} D\left(q, \bar{f}^{(t)}(\cdot|s)\right) + \frac{1}{\alpha} D\left(q, \bar{f}^{(t+1)}(\cdot|s)\right). \tag{286}$$

Let $q = \bar{f}^{(t)}(\cdot|s)$ and $\pi^\star(\cdot|s)$,resp., we have the following two inequalities:

$$\langle \Phi(s)\hat{w}^{(t)}, \bar{f}^{(t+1)}(\cdot|s) - \bar{f}^{(t)}(\cdot|s)\rangle \geq \frac{1}{\alpha} D\left(\bar{f}^{(t+1)}(\cdot|s), \bar{f}^{(t)}(\cdot|s)\right) + \frac{1}{\alpha} D\left(\bar{f}^{(t)}(\cdot|s), \bar{f}^{(t+1)}(\cdot|s)\right) \geq 0. \tag{287}$$

$$\langle \Phi(s)\hat{w}^{(t)}, \bar{f}^{(t+1)}(\cdot|s) - \bar{f}^{(t)}(\cdot|s)\rangle + \langle \Phi(s)\hat{w}^{(t)}, \bar{f}^{(t)}(\cdot|s) - \pi^\star(\cdot|s)\rangle \geq -\frac{1}{\alpha} D\left(\pi^\star(\cdot|s), \bar{f}^{(t)}(\cdot|s)\right) + \frac{1}{\alpha} D\left(\pi^\star(\cdot|s), \bar{f}^{(t+1)}(\cdot|s)\right). \tag{288}$$

Taking expectation w.r.t. distribution $d^\star$ on both sides of (288), we arrive at

$$\mathbb{E}_{s\sim d^\star}\left[\langle \Phi(s)\hat{w}^{(t)}, \bar{f}^{(t+1)}(\cdot|s) - \bar{f}^{(t)}(\cdot|s)\rangle\right] + \mathbb{E}_{s\sim d^\star}\left[\langle \Phi(s)\hat{w}^{(t)}, \bar{f}^{(t)}(\cdot|s) - \pi^\star(\cdot|s)\rangle\right] \geq \frac{1}{\alpha}(D_\star^{(t+1)} - D_\star^{(t)}). \tag{289}$$

To simplify the notation we let $\bar{Q}^{(t)}$ and $\bar{V}^{(t)}$ denote $Q^{\bar{f}^{(t)}}$ and $V^{\bar{f}^{(t)}}$, respectively. Note that the first expectation in the above expression (289) could be upper bounded as follows:

$$\begin{aligned}
&\mathbb{E}_{s\sim d^\star}\left[\langle \Phi(s)\hat{w}^{(t)}, \bar{f}^{(t+1)}(\cdot|s) - \bar{f}^{(t)}(\cdot|s)\rangle\right] \\
&= \sum_{s\in\mathcal{S}} d^\star(s)\langle \Phi(s)\hat{w}^{(t)}, \bar{f}^{(t+1)}(\cdot|s) - \bar{f}^{(t)}(\cdot|s)\rangle
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{s\in\mathcal{S}} \frac{d^\star(s)}{d^{\bar{f}^{(k+1)}}(s)} d^{\bar{f}^{(k+1)}}(s) \langle \Phi(s)\hat{w}^{(t)}, \bar{f}^{(t+1)}(\cdot|s) - \bar{f}^{(t)}(\cdot|s)\rangle \\
&\leq \vartheta_\rho \sum_{s\in\mathcal{S}} d^{\bar{f}^{(k+1)}}(s) \langle \Phi(s)\hat{w}^{(t)}, \bar{f}^{(t+1)}(\cdot|s) - \bar{f}^{(t)}(\cdot|s)\rangle \\
&= \vartheta_\rho \sum_{s\in\mathcal{S}} d^{\bar{f}^{(k+1)}}(s) \langle \bar{Q}^{(t)}(s,\cdot), \bar{f}^{(t+1)}(\cdot|s) - \bar{f}^{(t)}(\cdot|s)\rangle + \vartheta_\rho \sum_{s\in\mathcal{S}} d^{\bar{f}^{(k+1)}}(s) \langle \bar{\Phi}(s)\hat{w}^{(t)} - \bar{Q}^{(t)}(s,\cdot), \bar{f}^{(t+1)}(\cdot|s) - \bar{f}^{(t)}(\cdot|s)\rangle \\
&= \vartheta_\rho (1-\gamma)\left(\bar{V}^{(t+1)}(\rho) - \bar{V}^{(t)}(\rho)\right) + \vartheta_\rho \sum_{s\in\mathcal{S}} d^{\bar{f}^{(k+1)}}(s) \langle \bar{\Phi}(s)\hat{w}^{(t)} - \bar{Q}^{(t)}(s,\cdot), \bar{f}^{(t+1)}(\cdot|s) - \bar{f}^{(t)}(\cdot|s)\rangle\,, \qquad (290)
\end{aligned}
$$

where the first inequality uses (10) and the definition of ϑ_ρ (118) and the last line follows from (271) in Lemma 19. We separate the second term of the last line into four terms as follows:

$$
\begin{aligned}
&\sum_{s\in\mathcal{S}} d^{\bar{f}^{(t+1)}}(s) \langle \bar{\Phi}(s)\hat{w}^{(t)} - \bar{Q}^{(t)}(s,\cdot), \bar{f}^{(t+1)}(\cdot|s) - \bar{f}^{(t)}(\cdot|s)\rangle \\
&= \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^{\bar{f}^{(t+1)}}(s) \bar{f}^{(t+1)}(a|s) \phi^\top(s,a)(\hat{w}^{(t)} - \hat{w}_\star^{(t)})}_{(I)} + \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^{\bar{f}^{(t+1)}}(s) \bar{f}^{(t+1)}(a|s) \left(\phi^\top(s,a)\hat{w}_\star^{(t)} - \bar{Q}^{(t)}(s,a)\right)}_{(II)} \\
&+ \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^{\bar{f}^{(t+1)}}(s) \bar{f}^{(t)}(a|s) \phi^\top(s,a)(\hat{w}_\star^{(t)} - \hat{w}^{(t)})}_{(III)} + \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^{\bar{f}^{(t+1)}}(s) \bar{f}^{(t)}(a|s) \left(\bar{Q}^{(t)}(s,a) - \phi^\top(s,a)\hat{w}_\star^{(t)}\right)}_{(IV)}\,. \qquad (291)
\end{aligned}
$$

Applying again Lemma 19, we deduce the equivalent form of the second expectation in (289) as follows:

$$
\begin{aligned}
&\mathbb{E}_{s\sim d^\star}\left[\langle \Phi(s)\hat{w}^{(t)}, \bar{f}^{(t)}(\cdot|s) - \pi^\star(\cdot|s)\rangle\right] \\
&= \mathbb{E}_{s\sim d^\star}\left[\langle \bar{Q}^{(t)}(s,\cdot), \bar{f}^{(t)}(\cdot|s) - \pi^\star(\cdot|s)\rangle\right] + \mathbb{E}_{s\sim d^\star}\left[\langle \Phi(s)\hat{w}^{(t)} - \bar{Q}^{(t)}(s,\cdot), \bar{f}^{(t)}(\cdot|s) - \pi^\star(\cdot|s)\rangle\right] \\
&= (1-\gamma)\left(\bar{V}^{(t)}(\rho) - V^{\pi^\star}(\rho)\right) + \mathbb{E}_{s\sim d^\star}\left[\langle \Phi(s)\hat{w}^{(t)} - \bar{Q}^{(t)}(s,\cdot), \bar{f}^{(t)}(\cdot|s) - \pi^\star(\cdot|s)\rangle\right]\,, \qquad (292)
\end{aligned}
$$

where the second term of the last line could be decomposed into the following terms:

$$
\begin{aligned}
&\mathbb{E}_{s\sim d^\star}\left[\langle \Phi(s)\hat{w}^{(t)} - \bar{Q}^{(t)}(s,\cdot), \bar{f}^{(t)}(\cdot|s) - \pi^\star(\cdot|s)\rangle\right] \\
&= \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^\star(s) \bar{f}^{(t)}(a|s) \phi^\top(s,a)(\hat{w}^{(t)} - \hat{w}_\star^{(t)})}_{(A)} + \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^\star(s) \bar{f}^{(t)}(a|s) \left(\phi^\top(s,a)\hat{w}_\star^{(t)} - \bar{Q}^{(t)}(s,a)\right)}_{(B)} \\
&+ \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^\star(s) \pi^\star(a|s) \phi^\top(s,a)(\hat{w}_\star^{(t)} - \hat{w}^{(t)})}_{(C)} + \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^\star(s) \pi^\star(a|s) \left(\bar{Q}^{(t)}(s,a) - \phi^\top(s,a)\hat{w}_\star^{(t)}\right)}_{(D)}\,. \qquad (293)
\end{aligned}
$$

Plugging (291), (293) into (290) and (292), resp., and making use of (289), we have

$$
\begin{aligned}
&\vartheta_\rho(1-\gamma)\left(\bar{V}^{(t+1)}(\rho) - \bar{V}^{(t)}(\rho)\right) + (1-\gamma)\left(\bar{V}^{(t)}(\rho) - V^{\pi^\star}(\rho)\right) \\
&+ \vartheta_\rho\Bigg(\underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^{\bar{f}^{(t+1)}}(s) \bar{f}^{(t+1)}(a|s) \phi^\top(s,a)(\hat{w}^{(t)} - \hat{w}_\star^{(t)})}_{\text{(I)}} + \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^{\bar{f}^{(t+1)}}(s) \bar{f}^{(t+1)}(a|s) \left(\phi^\top(s,a)\hat{w}_\star^{(t)} - \bar{Q}^{(t)}(s,a)\right)}_{\text{(II)}} \\
&+ \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^{\bar{f}^{(t+1)}}(s) \bar{f}^{(t)}(a|s) \phi^\top(s,a)(\hat{w}_\star^{(t)} - \hat{w}^{(t)})}_{\text{(III)}} + \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^{\bar{f}^{(t+1)}}(s) \bar{f}^{(t)}(a|s) \left(\bar{Q}^{(t)}(s,a) - \phi^\top(s,a)\hat{w}_\star^{(t)}\right)}_{\text{(IV)}}\Bigg)
\end{aligned}
$$

$$+\underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^\star(s)\bar{f}^{(t)}(a|s)\phi^\top(s,a)(\hat{w}^{(t)}-\hat{w}_\star^{(t)})}_{(A)}+\underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^\star(s)\bar{f}^{(t)}(a|s)\left(\phi^\top(s,a)\hat{w}_\star^{(t)}-\bar{Q}^{(t)}(s,a)\right)}_{(B)}$$

$$+\underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^\star(s)\pi^\star(a|s)\phi^\top(s,a)(\hat{w}_\star^{(t)}-\hat{w}^{(t)})}_{(C)}+\underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^\star(s)\pi^\star(a|s)\left(\bar{Q}^{(t)}(s,a)-\phi^\top(s,a)\hat{w}_\star^{(t)}\right)}_{(D)}$$

$$\geq \frac{1}{\alpha}(D_\star^{(t+1)}-D_\star^{(t)}). \tag{294}$$

Below we upper bound the terms |(I)|-|(IV)| and $|(A)|$-$|(D)|$.

For any $t\in\mathbb{N}$ and $n\in[N]$, we define matrix $\Sigma_{\tilde{d}_n^{(t)}}\in\mathbb{R}^{p\times p}$ as

$$\Sigma_{\tilde{d}_n^{(t)}} \coloneqq \mathbb{E}_{(s,a)\sim\tilde{d}_n^{(t)}}\left[\phi(s,a)\phi^\top(s,a)\right], \tag{295}$$

and

$$\varepsilon_{\text{stat},n}^{(t)} \coloneqq \ell\left(w_n^{(t)},Q_n^{(t)},\tilde{d}_n^{(t)}\right)-\ell\left(w_{\star,n}^{(t)},Q_n^{(t)},\tilde{d}_n^{(t)}\right), \tag{296}$$

$$\varepsilon_{\text{approx},n}^{(t)} \coloneqq \ell\left(w_{\star,n}^{(t)},Q_n^{(t)},\tilde{d}_n^{(t)}\right), \tag{297}$$

then for all $n\in[N]$, by Assumption 5 and Assumption 3 we have

$$\mathbb{E}\left[\varepsilon_{\text{stat},n}^{(t)}\right]\leq\varepsilon_{\text{stat}}^n, \quad \text{and} \quad \mathbb{E}\left[\varepsilon_{\text{approx},n}^{(t)}\right]\leq\varepsilon_{\text{approx}}^n. \tag{298}$$

We let $\bar{\varepsilon}_{\text{stat}}^{(t)}\coloneqq\frac{1}{N}\sum_{n=1}^N\varepsilon_{\text{stat},n}^{(t)}$ and $\bar{\varepsilon}_{\text{approx}}^{(t)}\coloneqq\frac{1}{N}\sum_{n=1}^N\varepsilon_{\text{approx},n}^{(t)}$. By Cauchy-Schwartz's inequality we have

$$\begin{aligned}
|(\mathrm{I})| &\leq \frac{1}{N}\sum_{n=1}^N\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^{\bar{f}^{(t+1)}}(s)\bar{f}^{(t+1)}(a|s)|\phi^\top(s,a)(w_n^{(t)}-w_{\star,n}^{(t)})| \\
&\leq \frac{1}{N}\sum_{n=1}^N\sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\left(d^{\bar{f}^{(t+1)}}(s)\right)^2\left(\bar{f}^{(t+1)}(a|s)\right)^2}{\tilde{d}_n^{(t)}(s,a)}\cdot\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\tilde{d}_n^{(t)}(s,a)\left(\phi^\top(s,a)(w_n^{(t)}-w_{\star,n}^{(t)})\right)^2} \\
&= \frac{1}{N}\sum_{n=1}^N\sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}_n^{(t)}}\left[\left(\frac{\left(d^{\bar{f}^{(t+1)}}(s)\right)\left(\bar{f}^{(t+1)}(a|s)\right)}{\tilde{d}_n^{(t)}(s,a)}\right)^2\right]\left\|w_n^{(t)}-w_{\star,n}^{(t)}\right\|^2_{\Sigma_{\tilde{d}_n^{(t)}}}} \\
&\leq \frac{1}{N}\sum_{n=1}^N\sqrt{C_\nu\left\|w_n^{(t)}-w_{\star,n}^{(t)}\right\|^2_{\Sigma_{\tilde{d}_n^{(t)}}}} \\
&\leq \frac{1}{N}\sum_{n=1}^N\sqrt{C_\nu\varepsilon_{\text{stat},n}^{(t)}}\leq\sqrt{C_\nu\bar{\varepsilon}_{\text{stat}}^{(t)}},
\end{aligned} \tag{299}$$

where the third inequality follows from Assumption 4, the last inequality uses Jensen's inequality, and the penultimate inequality by Assumption 5 and by noticing that for all $w\in\mathbb{R}^p$, we have

$$\begin{aligned}
&\ell\left(w,Q_n^{(t)},\tilde{d}_n^{(t)}\right)-\ell\left(w_{\star,n}^{(t)},Q_n^{(t)},\tilde{d}_n^{(t)}\right) \\
&=\mathbb{E}_{(s,a)\sim\tilde{d}_n^{(t)}}\left[\left(\phi^\top(s,a)w-\phi^\top(s,a)w_{\star,n}^{(t)}+\phi^\top(s,a)w_{\star,n}^{(t)}-Q_n^{(t)}(s,a)\right)^2\right]-\ell\left(w_{\star,n}^{(t)},Q_n^{(t)},\tilde{d}_n^{(t)}\right) \\
&=\mathbb{E}_{(s,a)\sim\tilde{d}_n^{(t)}}\left[\left(\phi^\top(s,a)w-\phi^\top(s,a)w_{\star,n}^{(t)}\right)^2\right]+2\left(w-w_{\star,n}^{(t)}\right)^\top\mathbb{E}_{(s,a)\sim\tilde{d}_n^{(t)}}\left[\left(\phi^\top(s,a)w_{\star,n}^{(t)}-Q_n^{(t)}(s,a)\right)\phi(s,a)\right] \\
&=\left\|w-w_{\star,n}^{(t)}\right\|_{\Sigma_{\tilde{d}_n^{(t)}}}+\left(w-w_{\star,n}^{(t)}\right)^\top\nabla_w\ell\left(w_{\star,n}^{(t)},Q_n^{(t)},\tilde{d}_n^{(t)}\right)
\end{aligned}$$

$$\geq \left\| w - w_{\star,n}^{(t)} \right\|_{\Sigma_{\tilde{d}_n^{(t)}}}, \tag{300}$$

where the last line follows from the first-order optimality condition $w_{\star,n}^{(t)} \in \arg\min_w \ell\left(w, Q_n^{(t)}, \tilde{d}_n^{(t)}\right)$:

$$\forall w \in \mathbb{R}^p: \quad \left(w - w_{\star,n}^{(t)}\right)^\top \nabla_w \ell\left(w_{\star,n}^{(t)}, Q_n^{(t)}, \tilde{d}_n^{(t)}\right) \geq 0.$$

Analogous to bounding $|(\mathrm{I})|$, by simply substituting $\bar{f}^{(t+1)}$ with $\bar{f}^{(t)}$ or $\pi^\star$ or substituting $d^{\bar{f}^{(t+1)}}$ into $d^\star$, we obtain the same upper bound for $|(\mathrm{III})|$, $|(A)|$ and $|(C)|$, i.e.,

$$|(\mathrm{III})|, |(A)|, |(C)| \leq \sqrt{C_\nu \bar{\varepsilon}_{\mathrm{stat}}^{(t)}}. \tag{301}$$

Now we upper bound $|(\mathrm{II})|$ as follows:

$$\begin{aligned}
|(\mathrm{II})| &\leq \frac{1}{N} \sum_{n=1}^N \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^{\bar{f}^{(t+1)}}(s) \bar{f}^{(t+1)}(a|s) \left(|\phi^\top(s,a) w_{\star,n}^{(t)} - Q_n^{(t)}(s,a)| + |Q_n^{(t)}(s,a) - \bar{Q}^{(t)}(s,a)| \right) \\
&\leq \frac{1}{N} \sum_{n=1}^N \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left(d^{\bar{f}^{(t+1)}}(s)\right)^2 \left(\bar{f}^{(t+1)}(a|s)\right)^2}{\tilde{d}_n^{(t)}(s,a)}} \cdot \\
&\quad \cdot \sqrt{2 \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \tilde{d}_n^{(t)}(s,a) \left(\left(\phi^\top(s,a) w_{\star,n}^{(t)} - Q_n^{(t)}(s,a)\right)^2 + \left(Q_n^{(t)}(s,a) - \bar{Q}^{(t)}(s,a)\right)^2 \right)} \\
&= \frac{1}{N} \sum_{n=1}^N \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}_n^{(t)}} \left[\left(\frac{\left(d^{\bar{f}^{(t+1)}}(s)\right) \left(\bar{f}^{(t+1)}(a|s)\right)}{\tilde{d}_n^{(t)}(s,a)} \right)^2 \right] \cdot 2 \left(\varepsilon_{\mathrm{approx},n}^{(t)} + L_Q^2 \left\| \xi_n^{(t)} - \bar{\xi}^{(t)} \right\|_2^2 \right)} \\
&\leq \sqrt{2C_\nu \left({}_{\mathrm{F}}\bar{\varepsilon}_{\mathrm{approx}}^{(t)} + \frac{L_Q^2}{N} \left\| \boldsymbol{\xi}^{(t)} - \mathbf{1}\bar{\xi}^{(t)} \right\|_{\mathrm{F}}^2 \right)},
\end{aligned} \tag{302}$$

where L_Q is defined in Lemma 10, the second line uses Cauchy-Schwartz's inequality and Young's inequality (128) and the last inequality uses Assumption 4 and Jensen's inequality.

Analogous to bounding $|(\mathrm{II})|$, by simply substituting $\bar{f}^{(t+1)}$ with $\bar{f}^{(t)}$ or $\pi^\star$ or substituting $d^{\bar{f}^{(t+1)}}$ into $d^\star$, we obtain the same upper bound for $|(\mathrm{IV})|$, $|(B)|$ and $|(D)|$, i.e.,

$$|(\mathrm{IV})|, |(B)|, |(D)| \leq \sqrt{2C_\nu \left(\bar{\varepsilon}_{\mathrm{approx}}^{(t)} + \frac{L_Q^2}{N} \left\| \boldsymbol{\xi}^{(t)} - \mathbf{1}\bar{\xi}^{(t)} \right\|_{\mathrm{F}}^2 \right)}. \tag{303}$$

Plugging (299),(301),(302),(303) into (294) and dividing both sides by $(1-\gamma)$ yield

$$\vartheta_\rho \left(\delta^{(t+1)} - \delta^{(t)}\right) + \delta^{(t)} \leq \frac{D_\star^{(t)}}{(1-\gamma)\alpha} - \frac{D_\star^{(t+1)}}{(1-\gamma)\alpha} + \frac{2\sqrt{C_\nu}(\vartheta+1)}{1-\gamma} \left(\sqrt{\bar{\varepsilon}_{\mathrm{stat}}^{(t)}} + \sqrt{2\left(\bar{\varepsilon}_{\mathrm{approx}}^{(t)} + \frac{L_Q^2}{N} \left\| \boldsymbol{\xi}^{(t)} - \mathbf{1}\bar{\xi}^{(t)} \right\|_{\mathrm{F}}^2 \right)} \right).$$

Taking expectation on both sides of the above expression and making use of the simple fact that

$$\mathbb{E}\left[\sqrt{x}\right] \leq \sqrt{\mathbb{E}[x]},$$

we reach the conclusion (134).

D.3 Proof of Lemma 12

For any $\zeta > 0$, by the actor update rule (41) and (125) we have that

$$\begin{aligned}\left\|\boldsymbol{\xi}^{(t+1)} - \mathbf{1}_N \bar{\xi}^{(t+1)\top}\right\|_{\mathrm{F}}^2 &= \left\|\boldsymbol{W}(\boldsymbol{\xi}^{(t)} + \alpha \boldsymbol{h}^{(t)}) - \mathbf{1}_N(\bar{\xi}^{(t)} + \alpha \hat{w}^{(t)})^\top\right\|_{\mathrm{F}}^2 \\ &\leq (1+\zeta)\sigma^2 \left\|\boldsymbol{\xi}^{(t)} - \mathbf{1}_N \bar{\xi}^{(t)\top}\right\|_{\mathrm{F}}^2 + \alpha^2(1+1/\zeta)\sigma^2 \left\|\boldsymbol{h}^{(t)} - \mathbf{1}_N \hat{w}^{(t)\top}\right\|_{\mathrm{F}}^2, \end{aligned} \tag{304}$$

where the last line follows from Young's inequality (127) and (21). By the gradient tracking step (40) , Young's inequality (127) and (21), we have

$$\begin{aligned}\left\|\boldsymbol{h}^{(t+1)} - \mathbf{1}\hat{w}^{(t+1)\top}\right\|_{\mathrm{F}}^2 &= \left\|\boldsymbol{W}(\boldsymbol{h}^{(t)} + \boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)}) - \mathbf{1}\hat{w}^{(t)\top} + \mathbf{1}(\hat{w}^{(t)\top} - \hat{w}^{(t+1)\top})\right\|_{\mathrm{F}}^2 \\ &= \left\|\boldsymbol{W}\boldsymbol{h}^{(t)} - \mathbf{1}\hat{w}^{(t)\top} + \boldsymbol{W}(\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)}) - \mathbf{1}(\hat{w}^{(t+1)\top} - \hat{w}^{(t)\top})\right\|_{\mathrm{F}}^2 \\ &\leq (1+\zeta)\sigma^2 \left\|\boldsymbol{h}^{(t)} - \mathbf{1}_N\hat{w}^{(t)\top}\right\| + (1+1/\zeta)\sigma^2 \left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)} - \mathbf{1}(\hat{w}^{(t+1)\top} - \hat{w}^{(t)\top})\right\|_{\mathrm{F}}^2 \\ &\leq (1+\zeta)\sigma^2 \left\|\boldsymbol{h}^{(t)} - \mathbf{1}_N\hat{w}^{(t)\top}\right\| + (1+1/\zeta)\sigma^2 \left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)}\right\|_{\mathrm{F}}^2, \end{aligned} \tag{305}$$

where the last inequality follows from the fact

$$\begin{aligned}&\left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)} - \mathbf{1}(\hat{w}^{(t+1)\top} - \hat{w}^{(t)\top})\right\|_{\mathrm{F}}^2 \\ &= \left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)}\right\|_{\mathrm{F}}^2 + N\left\|\hat{w}^{(t+1)} - \hat{w}^{(t)}\right\|_2^2 - 2\sum_{n=1}^N \langle w_n^{(t+1)} - w_n^{(t)}, \hat{w}^{(t+1)} - \hat{w}^{(t)}\rangle \\ &= \left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)}\right\|_{\mathrm{F}}^2 - N\left\|\hat{w}^{(t+1)} - \hat{w}^{(t)}\right\|_2^2 \\ &\leq \left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)}\right\|_{\mathrm{F}}^2. \end{aligned} \tag{306}$$

Then for any $n \in [N]$, $t \in \mathbb{N}$ and $w \in \mathbb{R}^p$, we have

$$\begin{aligned}&\ell(w, Q_n^{(t)}, \tilde{d}_n^{(t)}) - \ell(w_{\star,n}^{(t)}, Q_n^{(t)}, \tilde{d}_n^{(t)}) \\ &= \mathbb{E}_{(s,a)\sim\tilde{d}_n^{(t)}}\left[\left(\phi^\top(s,a)w - \phi^\top(s,a)w_{\star,n}^{(t)} + \phi^\top(s,a)w_{\star,n}^{(t)} - Q_n^{(t)}(s,a)\right)^2\right] - \ell(w_{\star,n}^{(t)}, Q_n^{(t)}, \tilde{d}_n^{(t)}) \\ &= \mathbb{E}_{(s,a)\sim\tilde{d}_n^{(t)}}\left[\left(\phi^\top(s,a)w - \phi^\top(s,a)w_{\star,n}^{(t)}\right)^2\right] + 2(w - w_{\star,n}^{(t)})^\top \mathbb{E}_{(s,a)\sim\tilde{d}_n^{(t)}}\left[\left(\phi^\top(s,a)w_{\star,n}^{(t)} - Q_n^{(t)}(s,a)\right)\phi(s,a)\right] \\ &= \left\|w - w_{\star,n}^{(t)}\right\|_{\Sigma_{\tilde{d}_n^{(t)}}}^2 + (w - w_{\star,n}^{(t)})^\top \nabla_w \ell(w_{\star,n}^{(t)}, Q_n^{(t)}, \tilde{d}_n^{(t)}) \\ &\geq \left\|w - w_{\star,n}^{(t)}\right\|_{\Sigma_{\tilde{d}_n^{(t)}}}^2 \\ &\geq (1-\gamma)\mu\left\|w - w_{\star,n}^{(t)}\right\|_2^2, \end{aligned} \tag{307}$$

where the penultimate line follows from the first-order optimality conditions for the optima $w_{\star,n}^{(t)}$:

$$\forall w \in \mathbb{R}^p: \quad (w - w_{\star,n}^{(t)})^\top \nabla_w \ell(w_{\star,n}^{(t)}, Q_n^{(t)}, \tilde{d}_n^{(t)}) \geq 0 \tag{308}$$

and the last line is by Assumption 2 and (10).

Note that

$$\ell(w_{\star,n}^{(t)}, Q_n^{(t+1)}, \tilde{d}_n^{(t+1)})$$

$$
\begin{aligned}
&= \mathbb{E}_{(s,a)\sim \tilde{d}_n^{(t+1)}}\left[(\phi^\top(s,a)w_{\star,n}^{(t)} - Q_n^{(t+1)}(s,a))^2\right] \\
&\le 2\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \tilde{d}_n^{(t)}(s,a)\frac{\tilde{d}_n^{(t+1)}(s,a)}{\tilde{d}_n^{(t)}(s,a)}(\phi^\top(s,a)w_{\star,n}^{(t)} - Q_n^{(t)}(s,a))^2 + 2\mathbb{E}_{(s,a)\sim\tilde{d}_n^{(t+1)}}(Q_n^{(t+1)}(s,a) - Q_n^{(t)}(s,a))^2 \\
&\le 2C_\nu \mathbb{E}_{(s,a)\sim\tilde{d}_n^{(t)}}(\phi^\top(s,a)w_{\star,n}^{(t)} - Q_n^{(t)}(s,a))^2 + 2L_Q\left\|\xi_n^{(t+1)} - \xi_n^{(t)}\right\|_2^2 \\
&\le 2C_\nu \varepsilon_{\text{approx}}^n + 2L_Q^2\left\|\xi_n^{(t+1)} - \xi_n^{(t)}\right\|_2^2,
\end{aligned} \tag{309}
$$

where the second inequality uses Assumption 4 and Lemma 10, and the last line uses Assumption 3.

The above equation (309) together with (307) gives

$$
\begin{aligned}
\left\|\boldsymbol{w}_\star^{(t+1)} - \boldsymbol{w}_\star^{(t)}\right\|_{\text{F}}^2 = \sum_{n=1}^N \left\|w_{\star,n}^{(t+1)} - w_{\star,n}^{(t)}\right\|_2^2 &\le \frac{1}{(1-\gamma)\mu}\sum_{n=1}^N\left(\ell(w_{\star,n}^{(t)}, Q_n^{(t+1)}, \tilde{d}_n^{(t+1)}) - \ell(w_{\star,n}^{(t+1)}, Q_n^{(t+1)}, \tilde{d}_n^{(t+1)})\right) \\
&\le \frac{2}{(1-\gamma)\mu}\left(C_\nu\sum_{n=1}^N \varepsilon_{\text{approx}}^n + L_Q^2\left\|\boldsymbol{\xi}^{(t+1)} - \boldsymbol{\xi}^{(t)}\right\|_{\text{F}}^2\right).
\end{aligned} \tag{310}
$$

where $\boldsymbol{w}_\star^{(t)} \coloneqq (w_1^{(t)}, \cdots, w_N^{(t)})^\top$, $\forall t$.

Also note that by Assumption 5 and (307) we have

$$
\forall t \in \mathbb{N}: \quad \left\|\boldsymbol{w}^{(t)} - \boldsymbol{w}_\star^{(t)}\right\|_{\text{F}}^2 \le \frac{\sum_{n=1}^N \varepsilon_{\text{stat}}^n}{(1-\gamma)\mu}. \tag{311}
$$

Therefore, by (309) and (311) we have

$$
\begin{aligned}
\left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)}\right\|_{\text{F}}^2 &\le 3\left(\left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}_\star^{(t+1)}\right\|_{\text{F}}^2 + \left\|\boldsymbol{w}_\star^{(t+1)} - \boldsymbol{w}_\star^{(t)}\right\|_{\text{F}}^2 + \left\|\boldsymbol{w}^{(t)} - \boldsymbol{w}_\star^{(t)}\right\|_{\text{F}}^2\right) \\
&\le \frac{6}{(1-\gamma)\mu}\left(N(C_\nu\bar{\varepsilon}_{\text{approx}} + \bar{\varepsilon}_{\text{stat}}) + L_Q^2\left\|\boldsymbol{\xi}^{(t+1)} - \boldsymbol{\xi}^{(t)}\right\|_{\text{F}}^2\right).
\end{aligned} \tag{312}
$$

where the first inequality uses Young's inequality (128).

Note that by the update rule (41), the double stochasticity of the mixing matrix $\boldsymbol{W}$ and the consensus property (21) we have

$$
\begin{aligned}
\left\|\boldsymbol{\xi}^{(t+1)} - \boldsymbol{\xi}^{(t)}\right\|_{\text{F}}^2 &= \left\|\boldsymbol{W}(\boldsymbol{\xi}^{(t)} + \alpha\boldsymbol{h}^{(t)}) - \boldsymbol{\xi}^{(t)}\right\|_{\text{F}}^2 \\
&= \left\|(\boldsymbol{W} - \boldsymbol{I})(\boldsymbol{\xi}^{(t)} - \mathbf{1}_N\bar{\xi}^{(t)\top}) + \alpha(\boldsymbol{W}\boldsymbol{h}^{(t)} - \mathbf{1}_N\hat{w}^{(t)\top}) + \alpha\mathbf{1}(\hat{w}^{(t)} - \hat{w}_\star^{(t)})^\top + \mathbf{1}(\hat{w}_\star^{(t)})^\top\right\|_{\text{F}}^2 \\
&\le 16\left\|\boldsymbol{\xi}^{(t)} - \mathbf{1}_N\bar{\xi}^{(t)\top}\right\|_{\text{F}}^2 + 4\alpha^2\sigma^2\left\|\boldsymbol{h}^{(t)} - \mathbf{1}_N\hat{w}^{(t)\top}\right\|_{\text{F}}^2 + 4\alpha^2 N\left\|\hat{w}^{(t)} - \hat{w}_\star^{(t)}\right\|_2^2 + 4\alpha^2 N\left\|\hat{w}_\star^{(t)}\right\|_2^2 \\
&\le 16\left\|\boldsymbol{\xi}^{(t)} - \mathbf{1}_N\bar{\xi}^{(t)\top}\right\|_{\text{F}}^2 + 4\alpha^2\sigma^2\left\|\boldsymbol{h}^{(t)} - \mathbf{1}_N\hat{w}^{(t)\top}\right\|_{\text{F}}^2 + 4\alpha^2\sum_{n=1}^N\left\|w_n^{(t)} - w_{\star,n}^{(t)}\right\|_2^2 + 4\alpha^2\sum_{n=1}^N\left\|w_{\star,n}^{(t)}\right\|_2^2 \\
&\le 16\left\|\boldsymbol{\xi}^{(t)} - \mathbf{1}_N\bar{\xi}^{(t)\top}\right\|_{\text{F}}^2 + 4\alpha^2\sigma^2\left\|\boldsymbol{h}^{(t)} - \mathbf{1}_N\hat{w}^{(t)\top}\right\|_{\text{F}}^2 + \frac{4\alpha^2 N\bar{\varepsilon}_{\text{stat}}}{(1-\gamma)\mu} + \frac{4\alpha^2 N C_\phi^2}{\mu^2(1-\gamma)^4},
\end{aligned} \tag{313}
$$

where the penultimate line uses Jensen's inequality and the last line follows from (307), Assumption 5 and (167).

Combining (313) and (312) with (305), we deduce

$$
\left\|\boldsymbol{h}^{(t+1)} - \mathbf{1}\hat{w}^{(t+1)\top}\right\|_{\text{F}}^2 \le (1+1/\zeta)\frac{96\sigma^2 L_Q^2}{(1-\gamma)\mu}\left\|\boldsymbol{\xi}^{(t)} - \mathbf{1}\bar{\xi}^{(t)\top}\right\|_{\text{F}}^2 + \sigma^2\left(1+\zeta+(1+1/\zeta)\frac{24L_Q^2\alpha^2}{(1-\gamma)\mu}\right)\left\|\boldsymbol{h}^{(t)} - \mathbf{1}\hat{w}^{(t)\top}\right\|_{\text{F}}^2
$$

$$+ (1 + 1/\zeta)\frac{6\sigma^2}{(1-\gamma)\mu}\left(N(\bar{\varepsilon}_{\text{stat}} + C_\nu \bar{\varepsilon}_{\text{approx}}) + 4L_Q^2\left(\frac{\alpha^2 N \bar{\varepsilon}_{\text{stat}}}{(1-\gamma)\mu} + \frac{\alpha^2 N C_\phi^2}{\mu^2(1-\gamma)^2}\right)\right). \tag{314}$$

Finally, (135) follows from taking expectations on both sides of (304) and (314).