

# Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization

Shicong Cen\*    Chen Cheng†    Yuxin Chen‡    Yuting Wei§    Yuejie Chi¶  
CMU            Stanford            Princeton            CMU            CMU

July 13, 2020;    Revised April 8, 2021

## Abstract

Natural policy gradient (NPG) methods are among the most widely used policy optimization algorithms in contemporary reinforcement learning. This class of methods is often applied in conjunction with entropy regularization — an algorithmic scheme that encourages exploration — and is closely related to soft policy iteration and trust region policy optimization. Despite the empirical success, the theoretical underpinnings for NPG methods remain limited even for the tabular setting.

This paper develops *non-asymptotic* convergence guarantees for entropy-regularized NPG methods under softmax parameterization, focusing on discounted Markov decision processes (MDPs). Assuming access to exact policy evaluation, we demonstrate that the algorithm converges linearly — even quadratically once it enters a local region around the optimal policy — when computing optimal value functions of the regularized MDP. Moreover, the algorithm is provably stable vis-à-vis inexactness of policy evaluation. Our convergence results accommodate a wide range of learning rates, and shed light upon the role of entropy regularization in enabling fast convergence.

**Keywords:** natural policy gradient methods, entropy regularization, global convergence, soft policy iteration, conservative policy iteration, trust region policy optimization

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Background and motivation	3
1.2	This paper	3
1.3	Other related works	5
1.4	Notation	6
<b>2</b>	<b>Model and algorithms</b>	<b>6</b>
2.1	Problem settings	6
2.2	Algorithm: NPG methods with entropy regularization	8
2.3	A warm-up example: the bandit case	9
<b>3</b>	<b>Main results</b>	<b>10</b>
3.1	Exact entropy-regularized NPG methods	10
3.2	Approximate entropy-regularized NPG methods	12
3.3	Quadratic convergence in the small- $\epsilon$ regime	13

---

\*Department of Electrical and Computer Engineering, Carnegie Mellon University; email: [shicongc@andrew.cmu.edu](mailto:shicongc@andrew.cmu.edu).

†Department of Statistics, Stanford University; email: [chencheng@stanford.edu](mailto:chencheng@stanford.edu).

‡Department of Electrical and Computer Engineering, Princeton University; email: [yuxin.chen@princeton.edu](mailto:yuxin.chen@princeton.edu).

§Department of Statistics and Data Science, Carnegie Mellon University; email: [ytwei@cmu.edu](mailto:ytwei@cmu.edu).

¶Department of Electrical and Computer Engineering, Carnegie Mellon University; email: [yuejiechi@cmu.edu](mailto:yuejiechi@cmu.edu).

<b>4</b>	<b>Analysis</b>	<b>14</b>
4.1	Main pillars for the convergence analysis . . . . .	14
4.2	Analysis of exact entropy-regularized NPG methods . . . . .	15
4.2.1	The SPI case (i.e. $\eta = (1 - \gamma)/\tau$ ) . . . . .	15
4.2.2	The case with general learning rates . . . . .	16
4.3	Analysis of approximate entropy-regularized NPG methods . . . . .	18
4.4	Analysis of local quadratic convergence . . . . .	19
<b>5</b>	<b>Discussions</b>	<b>20</b>
<b>A</b>	<b>Preliminaries</b>	<b>24</b>
A.1	Derivation of entropy-regularized NPG methods . . . . .	24
A.2	Basic facts about the function $\log(\ \exp(\theta)\ _1)$ . . . . .	24
<b>B</b>	<b>Proof for the bandit case (Proposition 1)</b>	<b>25</b>
<b>C</b>	<b>Proof for key lemmas</b>	<b>25</b>
C.1	Proof of Lemma 1 . . . . .	25
C.2	Proof of Lemma 2 . . . . .	27
C.3	Proof of Lemma 3 . . . . .	28
C.4	Proof of Lemma 4 . . . . .	29
C.5	Proof of Lemma 5 . . . . .	30
C.6	Proof of Lemma 6 . . . . .	31
<b>D</b>	<b>Convergence guarantees for CPI-style policy updates</b>	<b>33</b>
D.1	Proof of Lemma 7 . . . . .	34
<b>E</b>	<b>Proof for approximate entropy-regularized NPG (Theorem 2)</b>	<b>35</b>
<b>F</b>	<b>Proof for local quadratic convergence (Theorem 3)</b>	<b>38</b>
F.1	Proof of Lemma 8 . . . . .	38

# 1 Introduction

Policy gradient (PG) methods and their variants (Williams, 1992; Sutton et al., 2000; Kakade, 2002; Peters and Schaal, 2008; Konda and Tsitsiklis, 2000), which aim to optimize (parameterized) policies via gradient-type methods, lie at the heart of recent advances in reinforcement learning (RL) (e.g. Mnih et al. (2015); Schulman et al. (2015); Silver et al. (2016); Schulman et al. (2017b)). Perhaps most appealing is their flexibility in adopting various kinds of policy parameterizations (e.g. a class of policies parameterized via deep neural networks), which makes them remarkably powerful and versatile in contemporary RL.

As an important and widely used extension of PG methods, *natural policy gradient* (NPG) methods propose to employ natural policy gradients (Amari, 1998) as search directions, in order to achieve faster convergence than the update rules based on policy gradients (Kakade, 2002; Peters and Schaal, 2008; Bhatnagar et al., 2009; Even-Dar et al., 2009). Informally speaking, NPG methods precondition the gradient directions by Fisher information matrices (which are the Hessians of a certain divergence metric), and fall under the category of quasi second-order policy optimization methods. In fact, a variety of mainstream RL algorithms, such as *trust region policy optimization* (TRPO) (Schulman et al., 2015) and *proximal policy optimization* (PPO) (Schulman et al., 2017b), can be viewed as generalizations of NPG methods (Shani et al., 2019). In this paper, we pursue in-depth theoretical understanding about this popular class of methods — in conjunction with entropy regularization to be introduced momentarily.

## 1.1 Background and motivation

Despite the enormous empirical success, the theoretical underpinnings of policy gradient type methods have been limited even until recently, primarily due to the intrinsic non-concavity underlying the value maximization problem of interest (Bhandari and Russo, 2019; Agarwal et al., 2020b). To further exacerbate the situation, an abundance of problem instances contain suboptimal policies residing in regions with flat curvatures (namely, vanishingly small gradients and high-order derivatives) (Agarwal et al., 2020b). Such plateaus in the optimization landscape could, in principle, be difficult to escape once entered, thereby necessitating a higher degree of exploration in order to accelerate policy optimization.

In practice, a strategy that has been frequently adopted to encourage exploration and improve convergence is to enforce entropy regularization (Williams and Peng, 1991; Peters et al., 2010; Mnih et al., 2016; Duan et al., 2016; Haarnoja et al., 2017; Hazan et al., 2019; Vieillard et al., 2020; Xiao et al., 2019). By inserting an additional penalty term to the objective function, this strategy penalizes policies that are not stochastic/exploratory enough, in the hope of preventing a policy optimization algorithm from being trapped in an undesired local region. Through empirical visualization, Ahmed et al. (2019) suggested that entropy regularization induces a smoother landscape that allows for the use of larger learning rates, and hence, faster convergence. However, the theoretical support for regularization-based policy optimization remains highly inadequate.

Motivated by this, a very recent line of works set out to elucidate, in a theoretically sound manner, the efficiency of entropy-regularized policy gradient methods. Assuming access to exact policy gradients, Agarwal et al. (2020b) and Mei et al. (2020) developed convergence guarantees for regularized PG methods (with relative entropy regularization considered in Agarwal et al. (2020b) and entropy regularization in Mei et al. (2020)). Encouragingly, both papers suggested the positive role of regularization in guaranteeing faster convergence for the tabular setting. However, these works fell short of explaining the role of entropy regularization for other policy optimization algorithms like NPG methods, which we seek to understand in this paper.

## 1.2 This paper

Inspired by recent theoretical progress towards understanding PG methods (Agarwal et al., 2020b; Bhandari and Russo, 2019; Mei et al., 2020), we aim to develop non-asymptotic convergence guarantees for entropy-regularized NPG methods in conjunction with softmax parameterization. We focus attention on studying tabular discounted Markov decision processes (MDPs), which is an important first step and a stepping stone towards demystifying the effectiveness of entropy-regularized policy optimization in more complex settings.

**Settings.** Consider a  $\gamma$ -discounted infinite-horizon MDP with state space  $\mathcal{S}$  and action space  $\mathcal{A}$ . Assuming availability of exact policy evaluation, the update rule of entropy-regularized NPG methods with softmax parameterization admits a simple update rule in the policy space (see Section 2 for precise descriptions)

$$\pi^{(t+1)}(a|s) \propto (\pi^{(t)}(a|s))^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta Q_\tau^\pi(s, a)}{1-\gamma}\right) \quad (1)$$

for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , where  $\tau > 0$  is the regularization parameter,  $0 < \eta \leq \frac{1-\gamma}{\tau}$  is the learning rate (or stepsize),  $\pi^{(t)}$  indicates the  $t$ -th policy iterate, and  $Q_\tau^\pi$  is the soft Q-function under policy  $\pi$  (to be defined in (11a)). The update rule (1) is closely connected to several popular algorithms in practice. For instance, the *trust region policy optimization* (TRPO) algorithm (Schulman et al., 2015), when instantiated in the tabular setting, can be viewed as implementing (1) with line search. In addition, by setting the learning rate as  $\eta = \frac{1-\gamma}{\tau}$ , the update rule (1) coincides with *soft policy iteration* (SPI) studied in Haarnoja et al. (2017).

**Our contributions.** The results of this paper deliver fully non-asymptotic convergence rates of entropy-regularized NPG methods without any hidden constants, which are previewed as follows (in an orderwise manner). The definition of  $\epsilon$ -optimality can be found in Table 1.

- **Linear convergence of exact entropy-regularized NPG methods.** We establish linear convergence of entropy-regularized NPG methods for finding the optimal policy of the entropy-regularized

MDP, assuming access to exact policy evaluation. To yield an  $\epsilon$ -optimal policy for the regularized MDP (cf. Table 1), the algorithm (1) with a general learning rate  $0 < \eta \leq \frac{1-\gamma}{\tau}$  needs no more than an order of

$$\frac{1}{\eta\tau} \log \left( \frac{1}{\epsilon} \right)$$

iterations, where we hide the dependencies that are logarithmic on salient problem parameters (see Theorem 1). Some highlights of our convergence results are (i) their near dimension-free feature and (ii) their applicability to a wide range of learning rates (including small learning rates).

- **Linear convergence of approximate entropy-regularized NPG methods.** We demonstrate the stability of the regularized NPG method with a general learning rate  $0 < \eta \leq \frac{1-\gamma}{\tau}$  even when the soft Q-functions of interest are only available approximately. This paves the way for future investigations that involve finite-sample analysis. Informally speaking, the algorithm exhibits the same convergence behavior as in the exact gradient case before an error floor is hit, where the error floor scales linearly in the entrywise error of the soft Q-function estimates (see Theorem 2).
- **Quadratic convergence in the small- $\epsilon$  regime.** In the high-accuracy regime where the target level  $\epsilon$  is *very small*, the algorithm (1) with  $\eta = \frac{1-\gamma}{\tau}$  converges super-linearly, in the sense that the iteration complexity to reach  $\epsilon$ -accuracy for the regularized MDP is at most on the order of

$$\log \log \left( \frac{1}{\epsilon} \right),$$

*after* entering a small local neighborhood surrounding the optimal policy. Here, we again hide the dependencies that are logarithmic on salient problem parameters (see Theorem 3).

**Comparisons with prior art.** Agarwal et al. (2020b) proved that unregularized NPG methods with softmax parameterization attain an  $\epsilon$ -accuracy within  $O(1/\epsilon)$  iterations. In contrast, our results assert that  $O(\log(1/\epsilon))$  iterations suffice with the assistance of entropy regularization, which hints at the potential benefit of entropy regularization in accelerating the convergence of NPG methods. Shortly after the initial posting of our paper, Bhandari and Russo (2020) posted a note that proves linear convergence of unregularized NPG methods with exact line search, by exploiting a clever connection to policy iteration. Their convergence rate is governed by a quantity  $\min_{s \in \mathcal{S}} \rho(s)$ , resulting in an iteration complexity at least  $|\mathcal{S}|$  times larger than ours. In comparison, our results cover a broad range of fixed learning rates (including small stepsizes that are of particular interest in practice), and accommodate the scenario with inexact gradient evaluation. See Table 1 for a quantitative comparison. Moreover, we note that the entropy-regularized NPG method with general learning rates is closely related to TRPO in the tabular setting (see Shani et al. (2019)). The recent work Shani et al. (2019) demonstrated that TRPO converges with an iteration complexity  $O(1/\epsilon)$  in entropy-regularized MDPs. The analysis therein is inspired by the mirror descent theory in generic optimization literature, which characterizes sublinear convergence under properly decaying stepsizes and accommodates various choices of divergence metrics. In comparison, our analysis strengthens the performance guarantees by carefully exploiting properties specific to the current version of the NPG method. In particular, we identify the delicate interplay between the crucial operational quantities  $Q_\tau^* - Q_\tau^{(t)}$  and  $Q_\tau^* - \tau \log \xi^{(t)}$  (to be defined later), and invoke the linear system theory to establish appealing contraction, which allow for the use of more aggressive constant stepsizes and hence improved convergence.

It is also helpful to compare our results with the state-of-the-art theory for PG methods with softmax parameterization (Agarwal et al., 2020b; Mei et al., 2020). Specifically, Agarwal et al. (2020b) established the asymptotic convergence of unregularized PG methods with softmax parameterization, while an iteration complexity of  $O(1/\epsilon)$  was recently pinned down by Mei et al. (2020). In the presence of entropy regularization, Agarwal et al. (2020b) showed that PG with relative entropy regularization and softmax parameterization enjoys an iteration complexity of  $O(1/\epsilon^2)$ , while Mei et al. (2020) showed that the entropy-regularized softmax PG method converges linearly in  $O(\log(1/\epsilon))$  iterations. However, the dependencies of the iteration complexity in Mei et al. (2020) on other salient parameters like  $|\mathcal{S}|$ ,  $|\mathcal{A}|$  and  $\frac{1}{1-\gamma}$  are not fully specified. Very recently, Li et al. (2021b) delivered a negative message demonstrating that these dependencies can be highly pessimistic; in fact, one can find an MDP instance which takes softmax PG methods (super)-exponential

paper	iteration complexity	regularization	learning rates
Agarwal et al. (2020b)	$\frac{2}{(1-\gamma)^2\epsilon} + \frac{2}{\eta\epsilon}$	unregularized	constant: $(0, \infty)$
Bhandari and Russo (2020)	$\frac{1}{(1-\gamma)\min_{s \in \mathcal{S}} \rho(s)} \log\left(\frac{1}{\epsilon}\right)$	unregularized	exact line search
<b>this work</b>	$\frac{1}{1-\gamma} \log\left(\frac{1}{\epsilon}\right)$	regularized	constant: $\frac{1-\gamma}{\tau}$
<b>this work</b>	$\frac{1}{\eta\tau} \log\left(\frac{1}{\epsilon}\right)$	regularized	constant: $(0, \frac{1-\gamma}{\tau})$

Table 1: The iteration complexities of NPG methods to reach  $\epsilon$ -accuracy *in terms of optimization error*, where the unregularized (resp. regularized) version is given by (13) (cf. (15)) with  $\eta$  the learning rate. We assume exact gradient evaluation and softmax parameterization, and hide the dependencies that are logarithmic on problem parameters. Here,  $\epsilon$ -accuracy or  $\epsilon$ -optimality for the unregularized (resp. regularized) case mean  $V^*(s) - V^{\pi^{(t)}}(s) \leq \epsilon$  (resp.  $V_\tau^*(s) - V_\tau^{\pi^{(t)}}(s) \leq \epsilon$ ) holds simultaneously for all  $s \in \mathcal{S}$ ;  $\rho$  denotes the initial state distribution, which clearly obeys  $\frac{1}{\min_{s \in \mathcal{S}} \rho(s)} \geq |\mathcal{S}|$ .

time (in terms of  $|\mathcal{S}|$  and  $\frac{1}{1-\gamma}$ ) to converge. In contrast, the bounds derived in the current paper are fully non-asymptotic, delineating clear dependencies on all salient problem parameters, which clearly demonstrate the algorithmic advantages of NPG methods. Fig. 1 depicts the policy paths of PG and NPG methods with entropy regularization for a simple bandit problem with three actions. It is evident from the plots that the NPG method follows a more direct path to the global optimum compared to the PG counterpart and hence converges faster. In addition, both algorithms converge more rapidly as the regularization parameter  $\tau$  increases.

### 1.3 Other related works

There has been a flurry of recent activities in studying theoretical behaviors of policy optimization methods. For example, Fazel et al. (2018); Jansch-Porto et al. (2020); Tu and Recht (2019); Zhang et al. (2019a); Mohammadi et al. (2019) established the global convergence of policy optimization methods for a couple of control problems; Bhandari and Russo (2019) identified structural properties that guarantee the global optimality of PG methods without parameterization; Karimi et al. (2019) studied the convergence of PG methods to an approximate first-order stationary point, and Zhang et al. (2019b) proposed a variant of PG methods that converges to locally optimal policies leveraging saddle-point escaping algorithms in nonconvex optimization. Beyond the tabular setting, the convergence of PG methods with function approximations has been studied in Agarwal et al. (2020b); Wang et al. (2019); Liu et al. (2019). In particular, Cai et al. (2019) developed an optimistic variant of NPG that incorporates linear function approximation. We do not elaborate on this line of works since our focus is on understanding the performance of entropy-regularized NPG in the tabular setting; we also do not elaborate on PG methods that involve sample-based estimates, since we primarily consider exact gradients or black-box gradient estimators.

Regarding entropy regularization, Neu et al. (2017); Geist et al. (2019) provided unified views of entropy-regularized MDPs from an optimization perspective by connecting them to algorithms such as mirror descent (Nemirovsky and Yudin, 1983) and dual averaging (Nesterov, 2009). The soft policy iteration algorithm has been identified as a special case of entropy-regularized NPG, highlighting again the link between policy gradient methods and soft Q-learning (Schulman et al., 2017a). The asymptotic convergence of soft policy iteration was established in Haarnoja et al. (2017), which fell short of providing explicit convergence rate guarantees. Additionally, Grill et al. (2019) developed planning algorithms for entropy-regularized MDPs, and Mei et al. (2020) showed that the sub-optimality gap of soft policy iteration is small if the policy improvement is small in consecutive iterations.

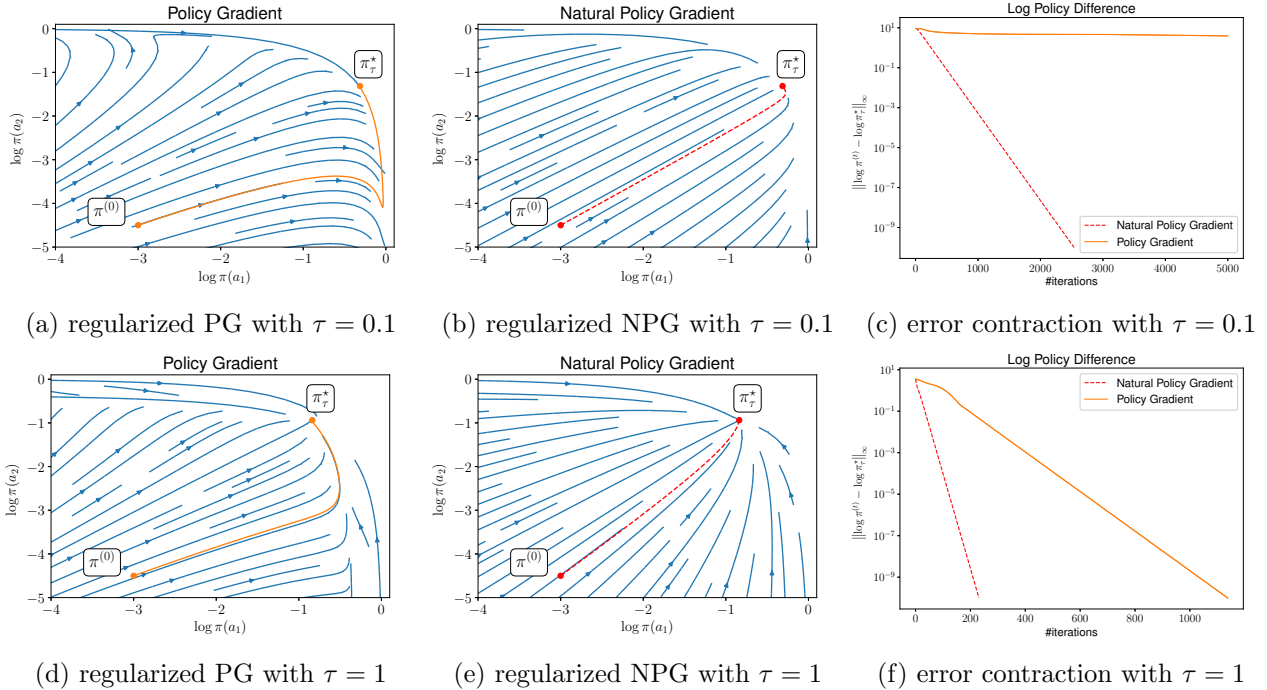


Figure 1: Comparisons of PG and NPG methods with entropy regularization for a bandit problem ( $\gamma = 0$ ) with 3 actions, whose corresponding rewards are 1.0, 0.9 and 0.1, respectively. The regularization parameter is set as  $\tau = 0.1$  for the first row and  $\tau = 1$  for the second row. In (a) and (d), the policy paths of  $(\log \pi(a_1), \log \pi(a_2))$  following the PG method are plotted in orange, with the blue lines indicating the gradient flow; in (b) and (e), the policy paths of  $(\log \pi(a_1), \log \pi(a_2))$  following the NPG method are depicted in red, with the blue lines indicating the natural gradient flow. The error contractions of both PG and NPG methods with  $\eta = 0.1$  are shown in (c) and (f).

## 1.4 Notation

We denote by  $\Delta(\mathcal{S})$  (resp.  $\Delta(\mathcal{A})$ ) the probability simplex over the set  $\mathcal{S}$  (resp.  $\mathcal{A}$ ). When scalar functions such as  $|\cdot|$ ,  $\exp(\cdot)$  and  $\log(\cdot)$  are applied to vectors, their applications should be understood in an entry-wise fashion. For instance, given any vector  $z = [z_i]_{1 \leq i \leq n} \in \mathbb{R}^n$ , the notation  $\|\cdot\|$  denotes  $\|z\| := \|[z_i]_{1 \leq i \leq n}\|$ ; other functions are defined analogously. For any vectors  $z = [z_i]_{1 \leq i \leq n}$  and  $w = [w_i]_{1 \leq i \leq n}$ , the notation  $z \geq w$  (resp.  $z \leq w$ ) means  $z_i \geq w_i$  (resp.  $z_i \leq w_i$ ) for all  $1 \leq i \leq n$ . The softmax function  $\text{softmax} : \mathbb{R}^n \mapsto \mathbb{R}^n$  is defined such that  $[\text{softmax}(\theta)]_i := \exp(\theta_i) / (\sum_i \exp(\theta_i))$  for a vector  $\theta = [\theta_i]_{1 \leq i \leq n} \in \mathbb{R}^n$ . Given two probability distributions  $\pi_1$  and  $\pi_2$  over  $\mathcal{A}$ , the Kullback-Leibler (KL) divergence from  $\pi_2$  to  $\pi_1$  is defined by  $\text{KL}(\pi_1 \parallel \pi_2) := \sum_{a \in \mathcal{A}} \pi_1(a) \log \frac{\pi_1(a)}{\pi_2(a)}$ . Given two probability distributions  $p$  and  $q$  over  $\mathcal{S}$ , we introduce the notation  $\|p\|_q^\infty := \max_{s \in \mathcal{S}} \frac{p(s)}{q(s)}$  and  $\|\frac{1}{q}\|_\infty := \max_{s \in \mathcal{S}} \frac{1}{q(s)}$ .

## 2 Model and algorithms

### 2.1 Problem settings

**Markov decision processes.** The current paper studies a discounted Markov decision process (MDP) (Puterman, 2014) denoted by  $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ , where  $\mathcal{S}$  is the state space,  $\mathcal{A}$  is the action space,  $\gamma \in (0, 1)$  indicates the discount factor,  $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$  is the transition kernel, and  $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$  stands for

the reward function.<sup>1</sup> 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  $s$  due to action  $a$ . A policy  $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$  represents a (randomized) action selection rule, namely,  $\pi(a|s)$  specifies the probability of executing action  $a$  in state  $s$  for each  $(s, a) \in \mathcal{S} \times \mathcal{A}$ .

**Value functions and Q-functions.** For any given policy  $\pi$ , we denote by  $V^\pi : \mathcal{S} \rightarrow \mathbb{R}$  the corresponding value function, namely, the expected discounted cumulative reward with an initial state  $s_0 = s$ , given by

$$\forall s \in \mathcal{S} : \quad V^\pi(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right], \quad (2)$$

where the action  $a_t \sim \pi(\cdot|s_t)$  follows the policy  $\pi$  and  $s_{t+1} \sim P(\cdot|s_t, a_t)$  is generated by the MDP  $\mathcal{M}$  for all  $t \geq 0$ . We also overload the notation  $V^\pi(\rho)$  to indicate the expected value function of a policy  $\pi$  when the initial state is drawn from a distribution  $\rho$  over  $\mathcal{S}$ , namely,

$$V^\pi(\rho) := \mathbb{E}_{s \sim \rho} [V^\pi(s)]. \quad (3)$$

Additionally, the Q-function  $Q^\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  of a policy  $\pi$  — namely, the expected discounted cumulative reward with an initial state  $s_0 = s$  and an initial action  $a_0 = a$  — is defined by

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad Q^\pi(s, a) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \right], \quad (4)$$

where the action  $a_t \sim \pi(\cdot|s_t)$  follows the policy  $\pi$  for all  $t \geq 1$ , and  $s_{t+1} \sim P(\cdot|s_t, a_t)$  is generated by the MDP  $\mathcal{M}$  for all  $t \geq 0$ .

**Discounted state visitation distributions.** A type of marginal distributions — commonly dubbed as *discounted state visitation distributions* — plays an important role in our theoretical development. To be specific, the discounted state visitation distribution  $d_{s_0}^\pi$  of a policy  $\pi$  given the initial state  $s_0 \in \mathcal{S}$  is defined by

$$\forall s \in \mathcal{S} : \quad d_{s_0}^\pi(s) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \mid s_0), \quad (5)$$

where the trajectory  $(s_0, s_1, \dots)$  is generated by the MDP  $\mathcal{M}$  under policy  $\pi$  starting from state  $s_0$ . In words,  $d_{s_0}^\pi(\cdot)$  captures the state occupancy probabilities when each state visitation is properly discounted depending on the time stamp. Further, for any distribution  $\rho$  over  $\mathcal{S}$ , we define the distribution  $d_\rho^\pi$  as follows

$$\forall s \in \mathcal{S} : \quad d_\rho^\pi(s) := \mathbb{E}_{s_0 \sim \rho} [d_{s_0}^\pi(s)], \quad (6)$$

which describes the discounted state visitation distribution when the initial state  $s_0$  is randomly drawn from a prescribed initial distribution  $\rho$ .

**Softmax parameterization.** It is common practice to parameterize the class of feasible policies in a way that is amenable to policy optimization. The focal point of this paper is softmax parameterization — a widely adopted scheme which naturally ensures that the policy lies in the probability simplex. Specifically, for any  $\theta : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  (called “logic values”), the corresponding softmax policy  $\pi_\theta$  is generated through the softmax transform

$$\pi_\theta := \text{softmax}(\theta) \quad \text{or} \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad \pi_\theta(a|s) := \frac{\exp(\theta(s, a))}{\sum_{a' \in \mathcal{A}} \exp(\theta(s, a'))}. \quad (7)$$

In what follows, we shall often abuse the notation to treat  $\pi_\theta$  and  $\theta$  as vectors in  $\mathbb{R}^{|\mathcal{S}| \cdot |\mathcal{A}|}$ , and suppress the subscript  $\theta$  from  $\pi_\theta$ , whenever it is clear from the context.

<sup>1</sup>For the sake of simplicity, we assume throughout that the reward resides within  $[0, 1]$ . Our results can be generalized in a straightforward manner to other ranges of bounded rewards.

**Entropy-regularized value maximization.** To promote exploration and discourage premature convergence to suboptimal policies, a widely used strategy is entropy regularization, which searches for a policy that maximizes the following entropy-regularized value function

$$V_\tau^\pi(\rho) := V^\pi(\rho) + \tau \cdot \mathcal{H}(\rho, \pi). \quad (8)$$

Here, the quantity  $\tau \geq 0$  denotes the regularization parameter, and  $\mathcal{H}(\rho, \pi)$  stands for a sort of *discounted entropy* defined as follows

$$\mathcal{H}(\rho, \pi) := \mathbb{E}_{\substack{s_0 \sim \rho, a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim P(\cdot|s_t, a_t), \forall t \geq 0}} \left[ \sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t) \right] = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\rho^\pi} \left[ \sum_{a \in \mathcal{A}} \pi(a|s) \log \frac{1}{\pi(a|s)} \right]. \quad (9)$$

Equivalently,  $V_\tau^\pi$  can be viewed as the value function of  $\pi$  by adjusting the instantaneous reward to be policy-dependent regularized version as follows

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad r_\tau(s, a) := r(s, a) - \tau \log \pi(a|s). \quad (10)$$

We also define  $V_\tau^\pi(s)$  analogously when the initial state is fixed to be any given state  $s \in \mathcal{S}$ . The regularized Q-function  $Q_\tau^\pi$  of a policy  $\pi$ , also known as the soft Q-function,<sup>2</sup> is related to  $V_\tau^\pi$  as

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad Q_\tau^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} [V_\tau^\pi(s')], \quad (11a)$$

$$\forall s \in \mathcal{S}: \quad V_\tau^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} [-\tau \log \pi(a|s) + Q_\tau^\pi(s, a)]. \quad (11b)$$

**Optimal policies and stationary distributions.** Denote by  $\pi^*$  (resp.  $\pi_\tau^*$ ) the policy that maximizes the value function (resp. regularized value function with regularization parameter  $\tau$ ), and let  $V^*$  (resp.  $V_\tau^*$ ) represent the resulting optimal value function (resp. regularized value function). Importantly, the optimal policies  $\pi^*$  and  $\pi_\tau^*$  of the MDP do not depend on the initial distribution  $\rho$  (Mei et al., 2020). In addition,  $\pi^*$  and  $\pi_\tau^*$  maximize the Q-function and the soft Q-function, respectively (which is self-evident from (11a)). A simple yet crucial connection between  $\pi^*$  and  $\pi_\tau^*$  can be demonstrated via the following sandwich bound<sup>3</sup>

$$V^{\pi_\tau^*}(\rho) \leq V^{\pi^*}(\rho) \leq V^{\pi_\tau^*}(\rho) + \frac{\tau}{1-\gamma} \log |\mathcal{A}|, \quad (12)$$

which holds for all initial distributions  $\rho$ . The key takeaway message is that: the optimal policy  $\pi_\tau^*$  of the regularized problem could also be nearly optimal in terms of the unregularized value function, as long as the regularization parameter  $\tau$  is chosen to be sufficiently small.

## 2.2 Algorithm: NPG methods with entropy regularization

**Natural policy gradient methods.** Towards computing the optimal policy (in the parameterized form), perhaps the first strategy that comes into mind is to run gradient ascent w.r.t. the parameter  $\theta$  until convergence — a first-order method commonly referred to as the *policy gradient* (PG) algorithm (e.g. Sutton et al. (2000)). In comparison, the *natural policy gradient* (NPG) method (Kakade, 2002) adopts a pre-conditioned gradient update rule

$$\theta \leftarrow \theta + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V^{\pi_\theta}(\rho), \quad (13)$$

in the hope of searching along a direction independent of the policy parameterization in use. Here,  $\eta$  is the learning rate or stepsize,  $\mathcal{F}_\rho^\theta$  denotes the Fisher information matrix given by

$$\mathcal{F}_\rho^\theta := \mathbb{E}_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \left[ (\nabla_\theta \log \pi_\theta(a|s)) (\nabla_\theta \log \pi_\theta(a|s))^\top \right], \quad (14)$$

<sup>2</sup>In this paper, we use the terms “regularized” value (resp. Q) functions and “soft” value (resp. Q) functions interchangeably.

<sup>3</sup>To see this, invoke the optimality of  $\pi_\tau^*$  and the elementary entropy bound  $0 \leq \mathcal{H}(\rho, \pi) \leq \frac{1}{1-\gamma} \log |\mathcal{A}|$  to obtain

$$V^{\pi_\tau^*}(\rho) + \frac{\tau}{1-\gamma} \log |\mathcal{A}| \geq V^{\pi_\tau^*}(\rho) + \tau \mathcal{H}(\rho, \pi_\tau^*) = V_\tau^*(\rho) \geq V^{\pi^*}(\rho) \geq V^{\pi^*}(\rho).$$



and we use  $B^\dagger$  to indicate the Moore-Penrose pseudoinverse of a matrix  $B$ . It has been understood that the NPG method essentially attempts to monitor/control the policy changes approximately in terms of the Kullback-Leibler (KL) divergence (see e.g. Schulman et al. (2015), Section 7)).

**NPG methods with entropy regularization.** Equipped with entropy regularization, the NPG update rule can be written as

$$\theta \leftarrow \theta + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V_\tau^{\pi_\theta}(\rho), \quad (15)$$

where  $\mathcal{F}_\rho^\theta$  is defined in (14) and  $V_\tau^\pi(\rho)$  is defined in (8). Under softmax parameterization, this update rule admits a fairly simple form in the policy space (see Appendix A.1 for detailed derivations), which, interestingly, is invariant to the choice of  $\rho$ . More precisely, if we let  $\theta^{(t)}$  denote the  $t$ -th iterate and  $\pi^{(t)} = \text{softmax}(\theta^{(t)})$  the associated policy, then the entropy-regularized NPG updates satisfy

$$\pi^{(t+1)}(a|s) = \frac{1}{Z^{(t)}(s)} (\pi^{(t)}(a|s))^{1 - \frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta Q_\tau^{\pi^{(t)}}(s, a)}{1 - \gamma}\right), \quad (16)$$

where  $Q_\tau^{\pi^{(t)}}$  is the soft Q-function of policy  $\pi^{(t)}$ , and  $Z^{(t)}(s)$  is some normalization factor. This can alternatively be viewed as an instantiation/variant of the *trust region policy optimization* (TRPO) algorithm (see Schulman et al. (2015); Shani et al. (2019)). As an important special case, the update rule (16) reduces to

$$\pi^{(t+1)}(\cdot|s) = \frac{1}{Z^{(t)}(s)} \exp\left(\frac{Q_\tau^{\pi^{(t)}}(s, \cdot)}{\tau}\right) \quad \text{when } \eta = \frac{1 - \gamma}{\tau} \quad (17)$$

for some normalization factor  $Z^{(t)}(s)$ . The procedure (17) can be interpreted as a “soft” version of the classical policy iteration algorithm (Bertsekas, 2017) (as it employs a softmax function to approximate the max operator) w.r.t. the soft Q-function, and is often dubbed as *soft policy iteration* (SPI) (see Haarnoja et al. (2018), Section 4.1)).

To simplify notation, we shall use  $V_\tau^{(t)}$ ,  $Q_\tau^{(t)}$  and  $d_\rho^{(t)}$  throughout to denote  $V_\tau^{\pi^{(t)}}$ ,  $Q_\tau^{\pi^{(t)}}$  and  $d_\rho^{\pi^{(t)}}$ , respectively. The complete procedure is summarized in Algorithm 1.

---

**Algorithm 1:** Entropy-regularized NPG with exact policy evaluation

---

1 **inputs:** learning rate  $\eta$ , initialization  $\pi^{(0)}$ .

2 **for**  $t = 0, 1, 2, \dots$  **do**

3     Compute the regularized Q-function  $Q_\tau^{(t)}$  (defined in (11a)) of policy  $\pi^{(t)}$ .

4     Update the policy:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad \pi^{(t+1)}(a|s) = \frac{1}{Z^{(t)}(s)} (\pi^{(t)}(a|s))^{1 - \frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta Q_\tau^{(t)}(s, a)}{1 - \gamma}\right), \quad (18)$$

$$\text{where } Z^{(t)}(s) = \sum_{a' \in \mathcal{A}} (\pi^{(t)}(a'|s))^{1 - \frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta Q_\tau^{(t)}(s, a')}{1 - \gamma}\right).$$


---

### 2.3 A warm-up example: the bandit case

Inspired by Schulman et al. (2017a); Mei et al. (2020), we look at a toy example — the bandit case — before proceeding to general MDPs. To be more precise, this is concerned with an MDP with only a single state and discount factor  $\gamma = 0$ . Despite its simplicity, the exposition of this example sheds light upon the convergence behavior of the regularized NPG methods of interest.

In this single-state example with  $\gamma = 0$ , the aim reduces to computing a policy  $\pi_\theta : \mathcal{A} \rightarrow \Delta(\mathcal{A})$  that solves the following optimization problem

$$\underset{\theta}{\text{maximize}} \quad \mathbb{E}_{a \sim \pi_\theta} [r(a) - \tau \log \pi_\theta(a)], \quad (19)$$

where  $r(a)$  is the instantaneous reward of taking action  $a$  (i.e. pulling arm  $a$  in the bandit language). As demonstrated in Mei et al. (2020, Proposition 1), this toy case is already non-concave and hence nontrivial to solve. As it turns out, direct calculation reveals that the optimal policy of (19) is given by

$$\pi_\tau^* = \text{softmax}(r/\tau), \quad (20)$$

which is in general a randomized policy. When applied to this example, the entropy-regularized NPG update rule (18) simplifies to (up to normalization)

$$\pi^{(t+1)}(a) \propto \pi^{(t)}(a) \exp\left(\eta r(a) - \eta \tau \log \pi^{(t)}(a)\right) = (\pi^{(t)}(a))^{1-\eta\tau} \exp(\eta r(a)), \quad (21)$$

with  $\eta$  the learning rate. The following proposition, whose proof is fairly elementary and can be found in Appendix B, reveals that the above procedure converges (at least) linearly to the optimal policy  $\pi_\tau^*$ .

**Proposition 1** (The bandit case). *The algorithm (21) converges linearly to  $\pi_\tau^*$  (cf. (20)) in an entrywise fashion, namely,*

$$\|\log \pi^{(t)} - \log \pi_\tau^*\|_\infty \leq 2(1 - \tau\eta)^t \|\log \pi^{(0)} - \log \pi_\tau^*\|_\infty.$$

While this result concentrates only on a toy example, it hints at the potential capability of entropy-regularized NPG methods in achieving rapid convergence. In particular, by setting the learning rate to be  $\eta = 1/\tau$ , the algorithm converges in *a single iteration*. This special choice corresponds to the SPI update (17), which will be singled out in our general theory due to its appealing convergence properties.

### 3 Main results

Given its appealing convergence behavior when applied to the preceding warm-up example (the bandit case), it is natural to ask whether the entropy-regularized NPG method is fast-convergent for general MDPs. This section answers this question in the affirmative.

#### 3.1 Exact entropy-regularized NPG methods

We first study the convergence behavior of entropy-regularized NPG methods (18) assuming access to exact policy evaluation in every iteration (namely, we assume the soft Q-function  $Q_\tau^{(t)}$  can be evaluated accurately in all  $t$ ). Remarkably, this algorithm converges linearly — in terms of computing both the optimal soft Q-function  $Q_\tau^*$  and the associated log policy  $\log \pi_\tau^*$  — as asserted by the following theorem. The proof of this result is provided in Section 4.2.

**Theorem 1** (Linear convergence of exact entropy-regularized NPG). *For any learning rate  $0 < \eta \leq (1-\gamma)/\tau$ , the entropy-regularized NPG updates (18) satisfy*

$$\|Q_\tau^* - Q_\tau^{(t+1)}\|_\infty \leq C_1 \gamma (1 - \eta\tau)^t \quad (22a)$$

$$\|\log \pi_\tau^* - \log \pi^{(t+1)}\|_\infty \leq 2C_1 \tau^{-1} (1 - \eta\tau)^t \quad (22b)$$

for all  $t \geq 0$ , where

$$C_1 := \|Q_\tau^* - Q_\tau^{(0)}\|_\infty + 2\tau \left(1 - \frac{\eta\tau}{1-\gamma}\right) \|\log \pi_\tau^* - \log \pi^{(0)}\|_\infty. \quad (23)$$

It is worth emphasizing that Theorem 1 is stated in a completely non-asymptotic form containing *no* hidden constants, and that our result covers any learning rate  $\eta$  in the range  $(0, (1-\gamma)/\tau]$ . A few implications of this theorem are in order.

- **Linear convergence of soft Q-functions.** To reach  $\|Q_\tau^* - Q_\tau^{(t)}\|_\infty \leq \epsilon$ , the entropy-regularized NPG method needs at most  $\frac{1}{\eta\tau} \log\left(\frac{C_1\gamma}{\epsilon}\right)$  iterations. Remarkably, the iteration complexity almost does not depend on the dimensions of the MDP (except for some very weak dependency embedded in

$\log C_1$ ) — this inherits a dimension-free feature of NPG methods that has been highlighted in [Agarwal et al. \(2020b\)](#) for the unregularized case. When the learning rate  $\eta$  is fixed in the admissible range, the iteration complexity scales inverse proportionally with  $\tau$ , suggesting a higher level of entropy regularization might accelerate convergence, albeit to the solution of a regularized problem that is further away from the original MDP.

- **Linear convergence of log policies.** In contrast to the unregularized case, entropy regularization ensures uniqueness of the optimal policy and, therefore, makes it possible to study the convergence of the policy directly. Our theorem reveals that the entropy-regularized NPG method needs at most  $\frac{1}{\eta\tau} \log\left(\frac{2C_1}{\epsilon\tau}\right)$  iterations to yield  $\|\log \pi_\tau^* - \log \pi^{(t+1)}\|_\infty \leq \epsilon$ .

- **Linear convergence of soft value functions.** As a byproduct, Theorem 1 implies that the iterates of soft value functions also converge linearly, namely,

$$\|V_\tau^* - V_\tau^{(t+1)}\|_\infty \leq 3C_1\gamma(1 - \eta\tau)^t. \quad (24)$$

To see this, we make note of the following relation previously established in [Nachum et al. \(2017\)](#):

$$\begin{aligned} \forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad & V_\tau^*(s) = -\tau \log \pi_\tau^*(a|s) + Q_\tau^*(s, a), \\ \implies \quad & V_\tau^*(s) = \mathbb{E}_{a \sim \pi^{(t+1)}(\cdot|s)} \left[ -\tau \log \pi_\tau^*(a|s) + Q_\tau^*(s, a) \right]. \end{aligned}$$

Consequently, combining this with the definition (11b) yields

$$\begin{aligned} |V_\tau^*(s) - V_\tau^{(t+1)}(s)| &= \mathbb{E}_{a \sim \pi^{(t+1)}(\cdot|s)} \left[ \left( -\tau \log \pi_\tau^*(a|s) + Q_\tau^*(s, a) \right) - \left( -\tau \log \pi_\tau^{(t+1)}(a|s) + Q_\tau^{(t+1)}(s, a) \right) \right] \\ &\leq \tau \|\log \pi_\tau^* - \log \pi_\tau^{(t+1)}\|_\infty + \|Q_\tau^* - Q_\tau^{(t+1)}\|_\infty, \end{aligned}$$

which together with (22) immediately establishes (24).

- **Convergence rate of SPI.** The best convergence guarantee is achieved when  $\eta = (1 - \gamma)/\tau$  (i.e. the SPI case), where the iteration complexity to reach  $\|Q_\tau^* - Q_\tau^{(t)}\|_\infty \leq \epsilon$  reduces to

$$\frac{1}{1 - \gamma} \log \left( \frac{\gamma \|Q_\tau^* - Q_\tau^{(0)}\|_\infty}{\epsilon} \right),$$

which is proportional to the effective horizon  $\frac{1}{1 - \gamma}$  modulo some log factor. This means the iteration complexity of SPI recovers that of policy iteration ([Puterman, 2014](#)). Interestingly, the contraction rate in this case (which is  $\gamma$ ) is independent of the choice of the regularization parameter  $\tau$ . Similarly, the iteration complexity of SPI to reach  $\|\log \pi_\tau^* - \log \pi^{(t+1)}\|_\infty \leq \epsilon$  becomes  $\frac{1}{1 - \gamma} \log \left( \frac{2\|Q_\tau^* - Q_\tau^{(0)}\|_\infty}{\epsilon\tau} \right)$ , and the contraction rate is again independent of  $\tau$ .

**Comparison with entropy-regularized policy gradient methods.** [Mei et al. \(2020, Theorem 6\)](#) proved that the entropy-regularized policy gradient method achieves<sup>4</sup>

$$\begin{aligned} V_\tau^*(\rho) - V_\tau^{(t)}(\rho) &\leq \left( V_\tau^*(\rho) - V_\tau^{(0)}(\rho) \right) \\ &\quad \cdot \exp \left( - \frac{(1 - \gamma)^4 t}{(8/\tau + 4 + 8 \log |\mathcal{A}|) |\mathcal{S}|} \left\| \frac{d_\rho^{\pi_\tau^*}}{\rho} \right\|_\infty^{-1} \min_s \rho(s) \left( \inf_{0 \leq k \leq t-1} \min_{s,a} \pi^{(k)}(a|s) \right)^2 \right), \end{aligned}$$

and they further showed that  $\inf_{k \geq 0} \min_{s,a} \pi^{(k)}(a|s)$  is non-vanishing in  $t$ . It remains unclear, however, how  $\inf_{t \geq 0} \min_{s,a} \pi^{(t)}(a|s)$  scales with other potentially large salient parameters like  $(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1 - \gamma}, \frac{1}{\tau})$ . In truth, existing theory does not rule out the possibility of exponential dependency on these salient parameters. It would thus be of great interest to establish algorithm-dependent lower bounds to uncover the right scaling with these important parameters. In contrast, our convergence guarantees for entropy-regularized NPG methods unveil concrete dependencies on all problem parameters.

<sup>4</sup>Here, we have assumed the exact policy gradient is computed with respect to  $V_\tau^{(t)}(\rho)$ .

**Computing an  $\epsilon$ -optimal policy for the original MDP.** Thus far, we have established an intriguing convergence behavior of the entropy-regularized NPG method. However, caution needs to be exercised when interpreting the efficacy of this method: the preceding results are concerned with convergence to the optimal regularized value function  $V_\tau^*$ , as opposed to finding the optimal value function  $V^*$  of the original MDP. Fortunately, by choosing the regularization parameter  $\tau$  to be sufficiently small (in accordance with the target accuracy level  $\epsilon$ ), we can guarantee that  $V_\tau^* \approx V^*$  (cf. (12)), thus ensuring the relevance and applicability of our results for solving the original MDP. To be specific, let us adopt the following choice of  $\tau$ :

$$\tau = \frac{(1-\gamma)\epsilon}{4\log|\mathcal{A}|}, \quad (25)$$

and assume the error of the regularized value function satisfies  $\|V_\tau^* - V_\tau^{(t)}\|_\infty < \epsilon/2$ . By virtue of Theorem 1, this optimization accuracy can be achieved via no more than  $\frac{4\log|\mathcal{A}|}{(1-\gamma)\eta\epsilon} \log\left(\frac{2C_{1\gamma}}{\epsilon}\right)$  iterations of entropy-regularized NPG updates with a general learning rate,<sup>5</sup> or no more than  $\frac{1}{1-\gamma} \log\left(\frac{\gamma\|Q_\tau^* - Q_\tau^{(0)}\|_\infty}{\epsilon}\right)$  iterations with the specific choice  $\eta = \frac{1-\gamma}{\tau}$ . It then follows that

$$\begin{aligned} V^*(s) - V^{(t)}(s) &= V^*(s) - V_\tau^*(s) + V_\tau^*(s) - V_\tau^{(t)}(s) + V_\tau^{(t)}(s) - V^{(t)}(s) \\ &\leq (V^*(s) - V_\tau^*(s)) + \|V_\tau^* - V_\tau^{(t)}\|_\infty + (V_\tau^{(t)}(s) - V^{(t)}(s)) \\ &\leq \frac{2\tau \log|\mathcal{A}|}{1-\gamma} + \frac{\epsilon}{2} = \epsilon \end{aligned}$$

for any  $s \in \mathcal{S}$ , where we have used our choice of  $\tau$  in (25). Here, the second inequality arises from (12) as well as the fact that for any policy  $\pi$ ,

$$\|V_\tau^\pi - V^\pi\|_\infty = \tau \max_s |\mathcal{H}(s, \pi)| \leq \frac{\tau \log|\mathcal{A}|}{1-\gamma},$$

given the elementary entropy bound  $0 \leq \mathcal{H}(s, \pi) \leq \frac{1}{1-\gamma} \log|\mathcal{A}|$ .

**Convergence guarantee for conservative policy iteration (CPI).** Our analysis framework also leads to a similar convergence guarantee for a type of policy updates adopted in *conservative policy iteration* (Kakade and Langford, 2002), where the policy is updated as a convex combination of the previous policy and an improved one. We refer the interested reader to Appendix D for details.

### 3.2 Approximate entropy-regularized NPG methods

There is no shortage of scenarios where the soft Q-function  $Q_\tau^{(t)}(s, a)$  is available only in an approximate fashion, e.g. the cases when the value function has to be evaluated using finite samples. To account for inexactness of policy evaluation, we extend our theory to accommodate the following approximate update rule: for any  $s \in \mathcal{S}$  and any  $t \geq 0$ ,

$$\pi^{(t+1)}(\cdot|s) \propto (\pi^{(t)}(\cdot|s))^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta\widehat{Q}_\tau^{(t)}(s, \cdot)}{1-\gamma}\right), \quad \text{where } \|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\|_\infty \leq \delta. \quad (26)$$

Here,  $\delta$  is some quantity that captures the size of approximation errors. We do not specify the estimator for the soft Q-function (as long as it satisfies the entrywise estimation bound), thus allowing one to plug in both model-based and model-free value function estimators designed for a variety of sampling mechanisms (e.g. Azar et al. (2013); Li et al. (2020b)). Encouragingly, the algorithm (26) is robust vis-à-vis inexactness of value function estimates, as it still converges linearly until an error floor is hit. This is formalized in the following theorem, with the proof postponed to Section 4.3.

<sup>5</sup>This result is in fact better than the iteration complexity  $\frac{2}{(1-\gamma)^2\epsilon}$  of the unregularized NPG method established in Agarwal et al. (2020b) as soon as  $\eta \geq 2(1-\gamma) \log|\mathcal{A}| \log\left(\frac{2C_{1\gamma}}{\epsilon}\right)$ . Consequently, our finding hints at the potential advantage of entropy-regularized NPG methods over the unregularized counterpart even when solving the original MDP.

**Theorem 2** (Linear convergence of approximate entropy-regularized NPG). *When  $0 < \eta \leq (1 - \gamma)/\tau$ , the inexact entropy-regularized NPG updates (26) satisfy*

$$\|Q_\tau^* - Q_\tau^{(t+1)}\|_\infty \leq \gamma \left[ (1 - \eta\tau)^t C_1 + C_2 \right] \quad (27a)$$

$$\|\log \pi_\tau^* - \log \pi^{(t+1)}\|_\infty \leq 2\tau^{-1} \left[ (1 - \eta\tau)^t C_1 + C_2 \right] \quad (27b)$$

for all  $t \geq 0$ , where  $C_1$  is the same as defined in (23) and  $C_2$  is given by

$$C_2 := \frac{2\delta}{1 - \gamma} \left( 1 + \frac{\gamma}{\eta\tau} \right) = \frac{2\delta}{(1 - \gamma)^2} \left[ 1 + \gamma \left( \frac{1 - \gamma}{\eta\tau} - 1 \right) \right]. \quad (28)$$

Apparently, Theorem 2 reduces to Theorem 1 when  $\delta = 0$ . As implied by this theorem, if the  $\ell_\infty$  error of the soft-Q function estimates does not exceed

$$\delta \leq \frac{(1 - \gamma)^2 \epsilon}{2\gamma \left[ 1 + \gamma \left( \frac{1 - \gamma}{\eta\tau} - 1 \right) \right]},$$

then the algorithm (26) achieves  $2\epsilon$ -accuracy (i.e.  $\|Q_\tau^* - Q_\tau^{(t)}\|_\infty \leq 2\epsilon$ ) within  $\frac{1}{\eta\tau} \log \left( \frac{C_1\gamma}{\epsilon} \right)$  iterations. In particular, in the case of soft policy iteration (i.e.  $\eta = \frac{1 - \gamma}{\tau}$ ), the tolerance level  $\delta$  can be up to  $\frac{(1 - \gamma)^2 \epsilon}{2\gamma}$ , which matches the theory of approximate policy iteration in Agarwal et al. (2019).

**Remark 1.** *It is straightforward to combine Theorem 2 with known sample complexities for approximate policy evaluation to obtain a crude sample complexity bound. For instance, assuming access to a generative model, Li et al. (2020a) asserts that for any fixed policy  $\pi$ , model-based policy evaluation achieves  $\|\hat{Q}_\tau^\pi - Q_\tau^\pi\|_\infty \leq \delta$  with high probability, as long as the number of samples per state-action pair exceeds the order of*

$$\frac{1}{(1 - \gamma)^3 \delta^2}$$

*up to some logarithmic factor. By employing fresh samples for each policy evaluation, we can set  $\delta = \frac{(1 - \gamma)^2 \epsilon}{2\gamma}$  and invoke the union bound over  $\tilde{O}\left(\frac{1}{1 - \gamma}\right)$  iterations to demonstrate that: SPI with model-based policy evaluation needs at most*

$$\tilde{O} \left( \frac{|\mathcal{S}| |\mathcal{A}|}{(1 - \gamma)^8 \epsilon^2} \right)$$

*samples to find an  $\epsilon$ -optimal policy. Here,  $\tilde{O}(\cdot)$  hides any logarithmic factor. We note, however, that the above sample analysis is extremely crude and might be improvable by, say, allowing sample reuses across iterations. It remains an interesting open question as to whether NPG with entropy regularization is minimax-optimal with a generative model, where the minimax lower bound is on the order of  $\frac{|\mathcal{S}| |\mathcal{A}|}{(1 - \gamma)^3 \epsilon^2}$  (Azar et al., 2013) and achievable by model-based plug-in estimators (Agarwal et al., 2020a; Li et al., 2020a) but not by vanilla Q-learning (Li et al., 2021a).*

### 3.3 Quadratic convergence in the small- $\epsilon$ regime

Somewhat remarkably, the regularized NPG method with  $\eta = \frac{1 - \gamma}{\tau}$  achieves super-linear convergence in computing  $V_\tau^*$ , once the algorithm enters a sufficiently small local neighborhood surrounding the optimizer.

Before presenting the result, we need to introduce the stationary distribution over  $\mathcal{S}$  of the MDP  $\mathcal{M}$  under policy  $\pi_\tau^*$ , denoted by  $\mu_\tau^* \in \Delta(\mathcal{S})$ . It is straightforward to verify the following basic property

$$d_{\mu_\tau^*}^{\pi_\tau^*} = \mu_\tau^*, \quad (29)$$

given that the state visitation distribution remains unchanged if the initial state is already in a steady state. Throughout this paper, we assume that  $\min_s \mu_\tau^*(s) > 0$ . Our finding is stated in the following theorem, with the proof deferred to Section 4.4.

**Theorem 3** (Quadratic convergence of exact regularized NPG). *Suppose that the algorithm (17) with  $\eta = \frac{1-\gamma}{\tau}$  (or SPI) satisfies*

$$\|\log \pi^{(t)} - \log \pi_\tau^*\|_\infty \leq 1. \quad (30)$$

for all  $t \geq 0$ , then one has

$$V_\tau^*(\rho) - V_\tau^{(t)}(\rho) \leq \left\| \frac{\rho}{\mu_\tau^*} \right\|_\infty \frac{(1-\gamma)\tau}{4\gamma^2} \left\| \frac{1}{\mu_\tau^*} \right\|_\infty^{-1} \left( \frac{4\gamma^2}{(1-\gamma)\tau} \left\| \frac{1}{\mu_\tau^*} \right\|_\infty \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(0)}(\mu_\tau^*) \right) \right)^{2^t}.$$

**Remark 2.** *In view of the convergence guarantees in Theorem 2, a suitable initialization of  $\pi^{(0)}$  and  $V_\tau^{(0)}$  (such that  $\frac{4\gamma^2}{(1-\gamma)\tau} \left\| \frac{1}{\mu_\tau^*} \right\|_\infty (V_\tau^*(\mu_\tau^*) - V_\tau^{(0)}(\mu_\tau^*)) < 1$ ) can be obtained by running SPI for sufficiently many iterations; further, all subsequent iterations are then guaranteed to satisfy (30) according to Theorem 2.*

Under the assumptions of Theorem 3, our result indicates that: when  $\epsilon$  is sufficiently small, the iteration complexity for SPI to yield an  $\epsilon$  optimization accuracy — that is,  $V_\tau^*(\rho) - V_\tau^{(t)}(\rho) \leq \epsilon$  — is at most on the order of

$$\log \log \left( \frac{(1-\gamma)\tau}{4\gamma^2} \left\| \frac{1}{\mu_\tau^*} \right\|_\infty^{-1} \left\| \frac{\rho}{\mu_\tau^*} \right\|_\infty \frac{1}{\epsilon} \right). \quad (31)$$

This uncovers the faster-than-linear convergence behavior of regularized NPG methods in the high-accuracy regime, accommodating a range of *optimization* accuracy and all possible choices of the regularization parameter  $\tau$ . It is worth noting, however, that our quadratic convergence result is stated in terms of the optimization accuracy (namely, convergence to the soft value function  $V_\tau^*(\rho)$ ) as opposed to the accuracy w.r.t. the original unregularized MDP. Thus, interpreting Theorem 3 in practice requires caution, since the approximation error  $V_\tau^*(\rho) - V^*(\rho)$  might sometimes dominate the optimization error in this regime.

## 4 Analysis

### 4.1 Main pillars for the convergence analysis

Before proceeding, we isolate a few ingredients that provide the main pillars for our theoretical development.

**Performance improvement and monotonicity.** This lemma is a sort of *ascent lemma*, which quantifies the progress made over each iteration — measured in terms of the soft value function.

**Lemma 1** (Performance improvement). *Suppose that  $0 < \eta \leq (1-\gamma)/\tau$ . For any distribution  $\rho$ , one has*

$$V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) = \mathbb{E}_{s \sim d_p^{(t+1)}} \left[ \left( \frac{1}{\eta} - \frac{\tau}{1-\gamma} \right) \text{KL} \left( \pi^{(t+1)}(\cdot|s) \parallel \pi^{(t)}(\cdot|s) \right) + \frac{1}{\eta} \text{KL} \left( \pi^{(t)}(\cdot|s) \parallel \pi^{(t+1)}(\cdot|s) \right) \right]. \quad (32)$$

*Proof.* See Appendix C.1. □

In a nutshell, Lemma 1 asserts that each iteration of the entropy-regularized NPG method is guaranteed to improve the estimates of the soft value function, with the improvement depending on the KL divergence between the current policy  $\pi^{(t)}$  and the updated one  $\pi^{(t+1)}$ . In fact, the arbitrary choice of  $\rho$  readily reveals a sort of *pointwise* monotonicity for the above range of learning rates, in the sense that  $V_\tau^{(t+1)}(s) \geq V_\tau^{(t)}(s)$  for all  $s \in \mathcal{S}$ . Indeed, this lemma can be viewed as the counterpart of the performance difference lemma in [Kakade and Langford \(2002\)](#) for the unregularized form. Lemma 1 also implies the monotonicity of the soft Q-function in  $t$ , since for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$  one has

$$Q_\tau^{(t+1)}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ V_\tau^{(t+1)}(s') \right] \geq r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ V_\tau^{(t)}(s') \right] = Q_\tau^{(t)}(s, a), \quad (33)$$

where the equalities follow from the definition (11a), and the inequality follows since  $V_\tau^{(t+1)}(s) \geq V_\tau^{(t)}(s)$  for all  $s \in \mathcal{S}$  — a consequence of Lemma 1 and the non-negativity of the KL divergence.

**A key contraction operator: the soft Bellman optimality operator.** An operator that plays a pivotal role in the theory of dynamic programming (Bellman, 1952) is the renowned Bellman optimality operator  $\mathcal{T} : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \rightarrow \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ , defined as follows

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad \mathcal{T}(Q)(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \max_{a'} Q(s', a') \right]. \quad (34)$$

In order to facilitate analysis for entropy-regularized MDPs, we find it particularly fruitful to introduce a “soft” Bellman optimality operator  $\mathcal{T}_\tau : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \rightarrow \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$  as follows

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad \mathcal{T}_\tau(Q)(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \max_{\pi(\cdot|s') \in \Delta(\mathcal{A})} \mathbb{E}_{a' \sim \pi(\cdot|s')} \left[ Q(s', a') - \tau \log \pi(a'|s') \right] \right], \quad (35)$$

which reduces to  $\mathcal{T}$  when  $\tau = 0$ . To see this, observe that

$$\begin{aligned} \mathcal{T}_0(Q)(s, a) &= r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \max_{\pi(\cdot|s') \in \Delta(\mathcal{A})} \mathbb{E}_{a' \sim \pi(\cdot|s')} [Q(s', a')] \right] \\ &= r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \max_{a'} Q(s', a') \right] = \mathcal{T}(Q)(s, a), \end{aligned}$$

where the last line follows since the optimal policy is exactly the greedy policy w.r.t.  $Q$  (Puterman, 2014). The operator  $\mathcal{T}_\tau$  plays a similar role as does the Bellman optimality operator for the unregularized case, whose key properties are summarized below. Similar results have been derived in Dai et al. (2018, Section 3.1).

**Lemma 2** (Soft Bellman optimality operator). *The operator  $\mathcal{T}_\tau$  defined in (35) satisfies the properties below.*

- $\mathcal{T}_\tau$  admits the following closed-form expression:

$$\mathcal{T}_\tau(Q)(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \tau \log \left( \left\| \exp(Q(s', \cdot)/\tau) \right\|_1 \right) \right]. \quad (36)$$

- The optimal soft  $Q$ -function  $Q_\tau^*$  is a fixed point of  $\mathcal{T}_\tau$ , namely,

$$\mathcal{T}_\tau(Q_\tau^*) = Q_\tau^*. \quad (37)$$

- $\mathcal{T}_\tau$  is a  $\gamma$ -contraction in the  $\ell_\infty$  norm, namely, for any  $Q_1, Q_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$  one has

$$\|\mathcal{T}_\tau(Q_1) - \mathcal{T}_\tau(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty. \quad (38)$$

*Proof.* See Appendix C.2. □

For those familiar with dynamic programming, it should become evident that  $\mathcal{T}_\tau$  inherits many appealing features of the original Bellman optimality operator  $\mathcal{T}$ . For example, as an immediate application of the  $\gamma$ -contraction property (38) and the fixed-point property (37), the following soft  $Q$ -value iteration

$$Q_{\text{svi}}^{(t+1)} = \mathcal{T}_\tau(Q_{\text{svi}}^{(t)}), \quad t \geq 0$$

is guaranteed to converge linearly to the optimal  $Q_\tau^*$  with a contraction rate  $\gamma$  — a simple observation consistent with the behavior of value iteration designed for unregularized MDPs.

## 4.2 Analysis of exact entropy-regularized NPG methods

### 4.2.1 The SPI case (i.e. $\eta = (1 - \gamma)/\tau$ )

With the help of the soft Bellman optimality operator, we have

$$Q_\tau^{(t+1)}(s, a) \stackrel{(i)}{=} r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ V_\tau^{(t+1)}(s') \right]$$

$$\begin{aligned}
&\stackrel{\text{(ii)}}{=} r(s, a) + \gamma \mathbb{E}_{\substack{s' \sim P(\cdot|s, a), \\ a' \sim \pi^{(t+1)}(\cdot|s')}} \left[ -\tau \log \pi^{(t+1)}(a'|s') + Q_\tau^{(t+1)}(s', a') \right] \\
&\stackrel{\text{(iii)}}{\geq} r(s, a) + \gamma \mathbb{E}_{\substack{s' \sim P(\cdot|s, a), \\ a' \sim \pi^{(t+1)}(\cdot|s')}} \left[ -\tau \log \pi^{(t+1)}(a'|s') + Q_\tau^{(t)}(s', a') \right] \\
&\stackrel{\text{(iv)}}{=} r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \tau \log \left( \left\| \exp(Q^{(t)}(s', \cdot)/\tau) \right\|_1 \right) \right] \\
&\stackrel{\text{(v)}}{=} \mathcal{T}_\tau(Q_\tau^{(t)})(s, a). \tag{39}
\end{aligned}$$

Here, (i) comes from the definition (11a) of the soft Q-function, (ii) follows from the relation (11b), (iii) relies on the monotonicity of the soft Q-function (see (33)), (iv) uses the form of  $\pi^{(t+1)}$  in (17), whereas (v) makes use of the expression (36). The inequality (39) further leads to  $0 \leq Q_\tau^* - Q_\tau^{(t+1)} \leq Q_\tau^* - \mathcal{T}_\tau(Q_\tau^{(t+1)})$ , and hence

$$\begin{aligned}
\|Q_\tau^* - Q_\tau^{(t+1)}\|_\infty &\leq \|Q_\tau^* - \mathcal{T}_\tau(Q_\tau^{(t)})\|_\infty = \|\mathcal{T}_\tau(Q_\tau^*) - \mathcal{T}_\tau(Q_\tau^{(t)})\|_\infty \leq \gamma \|Q_\tau^* - Q_\tau^{(t)}\|_\infty \\
&\leq \gamma^{t+1} \|Q_\tau^* - Q_\tau^{(0)}\|_\infty, \tag{40}
\end{aligned}$$

where the first equality follows from the fixed-point property (37), and the second inequality is due to the contraction property (38). We have thus established linear convergence of  $Q_\tau^{(t)}$  in  $\|\cdot\|_\infty$  for this case.

Turning to the log policies, recall that

$$\pi^{(t+1)}(\cdot|s) \propto \exp(Q_\tau^{(t)}(s, \cdot)/\tau) \quad \text{and} \quad \pi_\tau^*(\cdot|s) \propto \exp(Q_\tau^*(s, \cdot)/\tau),$$

where the second relation comes from Nachum et al. (2017, Eqn. (12)). It then follows from an elementary property of the softmax function (see (66) in Appendix A.2) that

$$\|\log \pi^{(t+1)} - \log \pi_\tau^*\|_\infty \leq \frac{2}{\tau} \|Q_\tau^{(t)} - Q_\tau^*\|_\infty \leq \frac{2}{\tau} \gamma^t \|Q_\tau^* - Q_\tau^{(0)}\|_\infty,$$

thus concluding the proof for this case.

#### 4.2.2 The case with general learning rates

We now move to the case with a general learning rate. For the sake of brevity, we shall denote

$$\alpha := 1 - \frac{\eta\tau}{1 - \gamma}. \tag{41}$$

Additionally, it is helpful to introduce an auxiliary sequence  $\{\xi^{(t)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}\}$  constructed recursively by

$$\xi^{(0)}(s, a) := \left\| \exp(Q_\tau^*(s, \cdot)/\tau) \right\|_1 \cdot \pi^{(0)}(a|s), \tag{42a}$$

$$\xi^{(t+1)}(s, a) := [\xi^{(t)}(s, a)]^\alpha \exp\left((1 - \alpha) \frac{Q_\tau^{(t)}(s, a)}{\tau}\right), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}, t \geq 0. \tag{42b}$$

It is easily seen from the construction (42b) that

$$\begin{aligned}
Q_\tau^* - \tau \log \xi^{(t+1)} &= Q_\tau^* - \tau \alpha \log \xi^{(t)} - (1 - \alpha) Q_\tau^{(t)} \\
&= \alpha (Q_\tau^* - \tau \log \xi^{(t)}) + (1 - \alpha) (Q_\tau^* - Q_\tau^{(t)})
\end{aligned} \tag{43}$$

and, consequently,

$$\|Q_\tau^* - \tau \log \xi^{(t+1)}\|_\infty \leq \alpha \|Q_\tau^* - \tau \log \xi^{(t)}\|_\infty + (1 - \alpha) \|Q_\tau^* - Q_\tau^{(t)}\|_\infty. \tag{44}$$



**Step 1: a linear system that describes the error recursions.** In the case with general learning rates, the estimation error  $\|Q_\tau^* - Q_\tau^{(t)}\|_\infty$  does not contract in the same form as that of soft policy iteration; instead, it is more succinctly controlled with the aid of an auxiliary quantity  $\|Q_\tau^* - \tau \log \xi^{(t)}\|_\infty$ . In what follows, we leverage a simple yet powerful technique by describing the dynamics concerning  $\|Q_\tau^* - Q_\tau^{(t)}\|_\infty$  and  $\|Q_\tau^* - \tau \log \xi^{(t)}\|_\infty$  via a linear system, whose spectral properties dictate the convergence rate. Towards this, we start with the following key observation, whose proof is deferred to Appendix C.3.

**Lemma 3.** *For any learning rate  $0 < \eta \leq (1 - \gamma)/\tau$ , the entropy-regularized NPG updates (18) satisfy*

$$\|Q_\tau^* - Q_\tau^{(t+1)}\|_\infty \leq \gamma \|Q_\tau^* - \tau \log \xi^{(t+1)}\|_\infty + \gamma \alpha^{t+1} \|Q_\tau^{(0)} - \tau \log \xi^{(0)}\|_\infty, \quad (45)$$

where  $\alpha$  is defined in (41).

If we substitute (43) into (45), it is straightforwardly seen that Lemma 3 is a generalization of the contraction property (40) of soft policy iteration (the case corresponding to  $\alpha = 0$ ). Given that Lemma 3 involves the interaction of more than one quantities, it is convenient to combine (44) and (45) into the following linear system

$$x_{t+1} \leq Ax_t + \gamma \alpha^{t+1} y, \quad (46)$$

where

$$A := \begin{bmatrix} \gamma(1 - \alpha) & \gamma\alpha \\ 1 - \alpha & \alpha \end{bmatrix}, \quad x_t := \begin{bmatrix} \|Q_\tau^* - Q_\tau^{(t)}\|_\infty \\ \|Q_\tau^* - \tau \log \xi^{(t)}\|_\infty \end{bmatrix} \quad \text{and} \quad y := \begin{bmatrix} \|Q_\tau^{(0)} - \tau \log \xi^{(0)}\|_\infty \\ 0 \end{bmatrix}. \quad (47)$$

We shall make note of the following appealing features of the rank-1 system matrix  $A$ :

$$A = \begin{bmatrix} \gamma \\ 1 \end{bmatrix} [1 - \alpha, \alpha], \quad \text{and} \quad A^t = (1 - \eta\tau)^{t-1} A \quad \forall t \geq 0, \quad (48)$$

which relies on the identity  $(1 - \alpha)\gamma + \alpha = 1 - \eta\tau$  (according to the definition (41) of  $\alpha$ ).

**Remark 3.** *By left multiplying both sides of (46) by  $[1 - \alpha, \alpha]$ , we obtain*

$$L^{(t+1)} \leq (1 - \eta\tau)L^{(t)} + \gamma(1 - \alpha)\alpha^{t+1} \|Q_\tau^{(0)} - \tau \log \xi^{(0)}\|_\infty,$$

where  $L^{(t)} := (1 - \alpha)\|Q_\tau^* - Q_\tau^{(t)}\|_\infty + \alpha\|Q_\tau^* - \tau \log \xi^{(t)}\|_\infty$  can be viewed as a sort of Lyapunov function. This hints at the intimate connection between our proof and the Lyapunov-type analysis used in system theory.

**Step 2: characterizing the contraction rate from the linear system.** In view of the recursion formula (46) and the non-negativity of  $(A, x_t, y)$ , it is immediate to deduce that

$$\begin{aligned} x_{t+1} &\leq A(Ax_{t-1} + \gamma\alpha^t y) + \gamma\alpha^{t+1} y \\ &\leq A^{t+1}x_0 + \gamma(\alpha^{t+1}I + \alpha^t A + \cdots + \alpha A^t)y \\ &= A^{t+1}x_0 + \gamma(A^{t+1} - \alpha^{t+1}I)(\alpha^{-1}A - I)^{-1}y. \end{aligned} \quad (49)$$

Here, the last line follows from the elementary relation

$$(\alpha^{t+1}I + \alpha^t A + \cdots + \alpha A^t)(\alpha^{-1}A - I) = A^{t+1} - \alpha^{t+1}I$$

and the invertibility of  $\alpha^{-1}A - I$  (since  $\alpha^{-1}A$  is a rank-1 matrix whose non-zero singular value is larger than 1). In addition, the Woodbury matrix inversion formula together with the decomposition (48) yields

$$\gamma(\alpha^{-1}A - I)^{-1}y = \gamma \left\{ \begin{bmatrix} 1 & \frac{\alpha}{1-\alpha} \\ \frac{1}{\gamma} & \frac{\alpha}{(1-\alpha)\gamma} \end{bmatrix} - I \right\} y = \begin{bmatrix} 0 & \frac{\gamma\alpha}{1-\alpha} \\ 1 & \frac{\gamma\alpha + \alpha - \gamma}{1-\alpha} \end{bmatrix} y = \begin{bmatrix} 0 \\ \|Q_\tau^{(0)} - \tau \log \xi^{(0)}\|_\infty \end{bmatrix}, \quad (50)$$

which is a non-negative vector. Consequently, this taken together with (49) gives

$$\begin{aligned}
x_{t+1} &\leq A^{t+1} \left[ x_0 + \gamma (\alpha^{-1} A - I)^{-1} y \right] - \alpha^{t+1} \left\{ \gamma (\alpha^{-1} A - I)^{-1} y \right\} \\
&\leq A^{t+1} \left[ x_0 + \gamma (\alpha^{-1} A - I)^{-1} y \right] \\
&= (1 - \eta\tau)^t \left( \begin{bmatrix} \gamma \\ 1 \end{bmatrix} [1 - \alpha, \alpha] \right) \left[ \begin{array}{c} \|Q_\tau^* - Q_\tau^{(0)}\|_\infty \\ \|Q_\tau^* - \tau \log \xi^{(0)}\|_\infty + \|Q_\tau^{(0)} - \tau \log \xi^{(0)}\|_\infty \end{array} \right] \\
&= (1 - \eta\tau)^t \left\{ (1 - \alpha) \|Q_\tau^* - Q_\tau^{(0)}\|_\infty + \alpha \left( \|Q_\tau^* - \tau \log \xi^{(0)}\|_\infty + \|Q_\tau^{(0)} - \tau \log \xi^{(0)}\|_\infty \right) \right\} \begin{bmatrix} \gamma \\ 1 \end{bmatrix}, \quad (51)
\end{aligned}$$

where the third line follows from (48), (50) and the definition of  $x_t$ . Further, observe that

$$\begin{aligned}
&\|Q_\tau^* - \tau \log \xi^{(0)}\|_\infty + \|Q_\tau^{(0)} - \tau \log \xi^{(0)}\|_\infty - \|Q_\tau^* - Q_\tau^{(0)}\|_\infty \\
&\leq 2 \|Q_\tau^* - \tau \log \xi^{(0)}\|_\infty = 2\tau \|\log \pi_\tau^* - \log \pi^{(0)}\|_\infty, \quad (52)
\end{aligned}$$

where the inequality comes from the triangle inequality, and the last identity follows from (42a). Substituting this back into (51), we obtain

$$x_{t+1} \leq (1 - \eta\tau)^t \left\{ \|Q_\tau^* - Q_\tau^{(0)}\|_\infty + 2\alpha\tau \|\log \pi_\tau^* - \log \pi^{(0)}\|_\infty \right\} \begin{bmatrix} \gamma \\ 1 \end{bmatrix}. \quad (53)$$

To finish up, recall that  $\pi^{(t)}$  is related to  $\xi^{(t)}$  as follows

$$\forall s \in \mathcal{S} : \quad \pi^{(t)}(\cdot | s) = \frac{1}{\|\xi^{(t)}(s, \cdot)\|_1} \xi^{(t)}(s, \cdot), \quad (54)$$

which can be seen by comparing (42) with (18). Therefore, invoking the elementary property of the softmax function (see (66) in Appendix A.2), we arrive at

$$\|\log \pi_\tau^* - \log \pi^{(t+1)}\|_\infty \leq 2 \|Q_\tau^*/\tau - \log \xi^{(t+1)}\|_\infty.$$

This combined with (53) as well as the definition (47) of  $x_{t+1}$  immediately establishes Theorem 1.

### 4.3 Analysis of approximate entropy-regularized NPG methods

We now turn to the convergence properties of approximate entropy-regularized NPG methods — as claimed in Theorem 2 — when only inexact policy evaluation  $\widehat{Q}_\tau^{(t)}$  is available (in the sense of (26)).

**Step 1: performance difference accounting for inexact policy evaluation.** We first bound the quality of the policy updates (26) by examining the difference between  $V_\tau^{(t+1)}$  and  $V_\tau^{(t)}$  and how it is impacted by the imperfectness of policy evaluation. This is made precise by the following lemma.

**Lemma 4** (Performance difference of approximate entropy-regularized NPG). *Suppose that  $0 < \eta \leq (1 - \gamma)/\tau$ . For any state  $s_0 \in \mathcal{S}$ , one has*

$$V_\tau^{(t)}(s_0) \leq V_\tau^{(t+1)}(s_0) + \frac{2}{1 - \gamma} \|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\|_\infty. \quad (55)$$

*Proof.* See Appendix C.4. □

The careful reader might already realize that the above lemma is a relaxation of Lemma 1; in particular, the last term of (55) quantifies the effect of the approximation error (i.e. the difference between  $\widehat{Q}_\tau^{(t)}$  and  $Q_\tau^{(t)}$ ) upon performance improvement. Under the assumption  $\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\|_\infty \leq \delta$ , repeating the argument of (33) reveals that the soft  $Q$ -function estimates are not far from being monotone in  $t$ , in the sense that

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad Q_\tau^{(t)}(s, a) - Q_\tau^{(t+1)}(s, a) = \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[ V_\tau^{(t)}(s') - V_\tau^{(t+1)}(s') \right] \leq \frac{2\gamma\delta}{1 - \gamma}. \quad (56)$$

**Step 2: a linear system accounting for inexact policy evaluation.** With the assistance of (56), it is possible to construct a linear system — similar to the one built in Section 4.2 — that takes into account inexact policy evaluation. Towards this end, we adopt a similar approach as in (42) by introducing the following auxiliary sequence  $\widehat{\xi}^{(t)}$  defined recursively using  $\widehat{Q}_\tau^{(t)}$ :

$$\widehat{\xi}^{(0)}(s, a) := \left\| \exp(Q_\tau^*(s, \cdot)/\tau) \right\|_1 \cdot \pi^{(0)}(s, a), \quad (57a)$$

$$\widehat{\xi}^{(t+1)}(s, a) := \left[ \widehat{\xi}^{(t)}(s, a) \right]^\alpha \exp\left( (1 - \alpha) \frac{\widehat{Q}_\tau^{(t)}(s, a)}{\tau} \right), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}, t \geq 0, \quad (57b)$$

where  $\alpha := 1 - \frac{\eta\tau}{1-\gamma}$  as before.

We claim that the following linear system tracks the error dynamics of the policy updates:

$$z_{t+1} \leq Bz_t + b, \quad (58)$$

where

$$B := \begin{bmatrix} \gamma(1-\alpha) & \gamma\alpha & \gamma\alpha \\ 1-\alpha & \alpha & 0 \\ 0 & 0 & \alpha \end{bmatrix}, \quad z_t := \begin{bmatrix} \|Q_\tau^* - Q_\tau^{(t)}\|_\infty \\ \|Q_\tau^* - \tau \log \widehat{\xi}^{(t)}\|_\infty \\ -\min_{s,a} (Q_\tau^{(t)}(s, a) - \tau \log \widehat{\xi}^{(t)}(s, a)) \end{bmatrix}, \quad b := (1-\alpha)\delta \begin{bmatrix} \gamma \left( 2 + \frac{2\gamma}{\eta\tau} \right) \\ 1 \\ 1 + \frac{2\gamma}{\eta\tau} \end{bmatrix}. \quad (59)$$

Here, the system matrix  $B$  (in particular its eigenvalues) governs the contraction rate, while the term  $b$  captures the error introduced by inexact policy evaluation. Theorem 2 then follows by carrying out a similar analysis argument as in Section 4.2 to characterize the error dynamics. Details are postponed to Appendix E.

#### 4.4 Analysis of local quadratic convergence

We now sketch the proof of Theorem 3, which establishes local quadratic convergence of SPI.

**Step 1: characterization of the sub-optimality gap.** Lemma 1 bounds the performance improvement of SPI by the KL divergence between the current policy  $\pi^{(t)}$  and the updated policy  $\pi^{(t+1)}$ . Interestingly, the type of KL divergence can be further employed to bound the sub-optimality gap for each iteration.

**Lemma 5** (Sub-optimality gap). *Suppose that  $\eta = (1 - \gamma)/\tau$ . For any distribution  $\rho$ , one has*

$$V_\tau^*(\rho) - V_\tau^{(t)}(\rho) \leq \frac{1}{\eta} \mathbb{E}_{s \sim d_\rho^{\pi_\tau^*}} \left[ \text{KL} \left( \pi^{(t)}(\cdot|s) \parallel \pi^{(t+1)}(\cdot|s) \right) \right].$$

*Proof.* This result has appeared in Mei et al. (2020, Eqn. (486)). For completeness we include a proof in Appendix C.5.  $\square$

In words, Lemma 5 formalizes the connection between the sub-optimality gap (w.r.t. the optimal soft value function) and the proximity of the two consecutive policy iterates. As reflected by this lemma, if the current and the updated policies do not differ by much (which indicates that the algorithm might be close to convergence), then the current estimate of the soft value function is close to optimal.

**Step 2: a contraction property.** The importance of the above two lemmas is made apparent by the following contraction property when  $\eta = (1 - \gamma)/\tau$ :

$$\begin{aligned} V_\tau^*(\rho) - V_\tau^{(t+1)}(\rho) &= V_\tau^*(\rho) - V_\tau^{(t)}(\rho) + \left( V_\tau^{(t)}(\rho) - V_\tau^{(t+1)}(\rho) \right) \\ &\stackrel{(i)}{=} V_\tau^*(\rho) - V_\tau^{(t)}(\rho) - \frac{1}{\eta} \mathbb{E}_{s \sim d_\rho^{(\pi_\tau^{(t+1)})}} \left[ \text{KL} \left( \pi^{(t)}(\cdot|s) \parallel \pi^{(t+1)}(\cdot|s) \right) \right] \\ &\stackrel{(ii)}{\leq} V_\tau^*(\rho) - V_\tau^{(t)}(\rho) - \frac{1}{\eta} \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\rho^{(\pi_\tau^{(t+1)})}} \right\|_\infty^{-1} \mathbb{E}_{s \sim d_\rho^{\pi_\tau^*}} \left[ \text{KL} \left( \pi^{(t)}(\cdot|s) \parallel \pi^{(t+1)}(\cdot|s) \right) \right] \end{aligned}$$

$$\begin{aligned}
&\stackrel{\text{(iii)}}{\leq} V_\tau^*(\rho) - V_\tau^{(t)}(\rho) - \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\rho^{(t+1)}} \right\|_\infty^{-1} \left( V_\tau^*(\rho) - V_\tau^{(t)}(\rho) \right) \\
&= \left( 1 - \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\rho^{(t+1)}} \right\|_\infty^{-1} \right) \left( V_\tau^*(\rho) - V_\tau^{(t)}(\rho) \right). \tag{60}
\end{aligned}$$

Here, (i) arises from Lemma 1, (ii) employs the pre-factor  $\|d_\rho^{\pi_\tau^*}/d_\rho^{(t+1)}\|_\infty^{-1}$  to accommodate the change of distributions, whereas (iii) follows from Lemma 5.

**Step 3: super-linear convergence in the small- $\epsilon$  regime.** The contraction property (60) implies that  $V_\tau^{(t+1)}(\rho)$  converges *super-linearly* to  $V_\tau^*$ , once  $\pi_\tau^{(t)}$  gets sufficiently close to  $\pi_\tau^*$ . In fact, once the ratio  $d_\rho^{(t+1)}/d_\rho^{\pi_\tau^*}$  becomes sufficiently close to 1, the contraction factor  $1 - \|d_\rho^{\pi_\tau^*}/d_\rho^{(t+1)}\|_\infty^{-1}$  in (60) is approaching 0, thereby accelerating convergence. This observation underlies Theorem 3, whose complete analysis is postponed to Appendix F.

## 5 Discussions

This paper establishes non-asymptotic convergence of entropy-regularized natural policy gradient methods, providing theoretical footings for the role of entropy regularization in guaranteeing fast convergence. Our analysis opens up several directions for future research; we close the paper by sampling a few of them.

- *Extended analysis of policy gradient methods with inexact gradients.* It would be of interest to see whether our analysis framework can be applied to improve the theory of policy gradient methods (Mei et al., 2020) to accommodate the case with inexact policy gradients.
- *Finite-sample analysis in the presence of sample-based policy evaluation.* Another natural extension is towards understanding the sample complexity of entropy-regularized NPG methods when the value functions are estimated using rollout trajectories (see e.g. Kakade and Langford (2002); Agarwal et al. (2020b); Shani et al. (2019)), or using bootstrapping (see e.g. Xu et al. (2020); Haarnoja et al. (2018); Wu et al. (2020)).
- *Function approximation.* The current work has been limited to the tabular setting. It would certainly be interesting, and fundamentally important, to understand entropy-regularized NPG methods in conjunction with function approximation; see Sutton et al. (2000); Agarwal et al. (2019, 2020b) for a few representative scenarios.
- *Beyond softmax parameterization.* The current paper has been devoted to softmax parameterization, which enables a concise and NPG update rule. A couple of other parameterization schemes have been proposed for (vanilla) PG methods as well (Agarwal et al., 2019, 2020b; Bhandari and Russo, 2019, 2020), e.g. vanilla parameterization (paired with proper projection onto the probability simplex in each iteration), log-linear parameterization, and neural softmax parameterization. Unfortunately, the analysis in our paper relies heavily on the softmax NPG update rule, and does not immediately extend to other parameterization. It would be of great importance to establish convergence guarantees that accommodate other parameterizations of practical interest.

## Acknowledgments

The authors are grateful to anonymous reviewers for helpful suggestions, particularly for bringing Dai et al. (2018) to our attention. S. Cen and Y. Chi are supported in part by the grants ONR N00014-18-1-2142 and N00014-19-1-2404, ARO W911NF-18-1-0303, NSF CCF-1806154, CCF-1901199 and CCF-2007911. C. Cheng is supported by the William R. Hewlett Stanford graduate fellowship. Y. Wei is supported in part by the NSF grants CCF-2007911 and DMS-2015447. Y. Chen is supported in part by the grants AFOSR YIP award FA9550-19-1-0030, ONR N00014-19-1-2120, ARO YIP award W911NF-20-1-0097, ARO W911NF-18-1-0303, NSF CCF-1907661, IIS-1900140 and DMS-2014279, and the Princeton SEAS Innovation Award.

## References

- Agarwal, A., Jiang, N., and Kakade, S. M. (2019). Reinforcement learning: Theory and algorithms. Technical report.
- Agarwal, A., Kakade, S., and Yang, L. F. (2020a). Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR.
- Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2020b). Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR.
- 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.
- Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349.
- Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716.
- Bertsekas, D. P. (2017). *Dynamic programming and optimal control (4th edition)*. Athena Scientific.
- Bhandari, J. and Russo, D. (2019). Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*.
- Bhandari, J. and Russo, D. (2020). A note on the linear convergence of policy gradient methods. *arXiv preprint arXiv:2007.11120*.
- Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. (2009). Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482.
- Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2019). Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*.
- Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.
- Dai, B., Shaw, A., Li, L., Xiao, L., He, N., Liu, Z., Chen, J., and Song, L. (2018). SBEEED: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134. PMLR.
- Duan, Y., Chen, X., Houthoofd, R., Schulman, J., and Abbeel, P. (2016). Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338.
- Even-Dar, E., Kakade, S. M., and Mansour, Y. (2009). Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736.
- Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476.
- Geist, M., Scherrer, B., and Pietquin, O. (2019). A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169.
- Grill, J.-B., Darwiche Domingues, O., Menard, P., Munos, R., and Valko, M. (2019). Planning in entropy-regularized markov decision processes and games. In *Advances in Neural Information Processing Systems*, volume 32.
- Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361.

- Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*.
- Hazan, E., Kakade, S., Singh, K., and Van Soest, A. (2019). Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691.
- Jansch-Porto, J. P., Hu, B., and Dullerud, G. (2020). Convergence guarantees of policy optimization methods for Markovian jump linear systems. *arXiv preprint arXiv:2002.04090*.
- Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274.
- Kakade, S. M. (2002). A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538.
- Karimi, B., Miasojedow, B., Moulines, É., and Wai, H.-T. (2019). Non-asymptotic analysis of biased stochastic approximation scheme. *arXiv preprint arXiv:1902.00629*.
- Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014.
- Li, G., Cai, C., Chen, Y., Gu, Y., Wei, Y., and Chi, Y. (2021a). Is Q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*.
- Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020a). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *arXiv preprint arXiv:2005.12900*.
- Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020b). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *arXiv preprint arXiv:2006.03041*.
- Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2021b). Softmax policy gradient methods can take exponential time to converge. *arXiv preprint arXiv:2102.11270*.
- Liu, B., Cai, Q., Yang, Z., and Wang, Z. (2019). Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, pages 10565–10576.
- Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020). On the global convergence rates of softmax policy gradient methods. *arXiv preprint arXiv:2005.06392*.
- 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.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.
- Mohammadi, H., Zare, A., Soltanolkotabi, M., and Jovanović, M. R. (2019). Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem. *arXiv preprint arXiv:1912.11899*.
- 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.
- Nemirovsky, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization.
- Nesterov, Y. (2009). Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259.
- Neu, G., Jonsson, A., and Gómez, V. (2017). A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*.

- Peters, J., Mulling, K., and Altun, Y. (2010). Relative entropy policy search. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.
- Peters, J. and Schaal, S. (2008). Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190.
- Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.
- Schulman, J., Chen, X., and Abbeel, P. (2017a). Equivalence between policy gradients and soft Q-learning. *arXiv preprint arXiv:1704.06440*.
- 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. (2017b). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Shani, L., Efroni, Y., and Mannor, S. (2019). Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. *arXiv preprint arXiv:1909.02769*.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489.
- Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063.
- Tu, S. and Recht, B. (2019). The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory*, pages 3036–3083.
- Vieillard, N., Kozuno, T., Scherrer, B., Pietquin, O., Munos, R., and Geist, M. (2020). Leverage the average: an analysis of regularization in RL. *arXiv preprint arXiv:2003.14089*.
- 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. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.
- Williams, R. J. and Peng, J. (1991). Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268.
- Wu, Y., Zhang, W., Xu, P., and Gu, Q. (2020). A finite time analysis of two time-scale actor critic methods.
- Xiao, C., Huang, R., Mei, J., Schuurmans, D., and Müller, M. (2019). Maximum entropy Monte-Carlo planning. In *Advances in Neural Information Processing Systems*, pages 9520–9528.
- Xu, T., Wang, Z., and Liang, Y. (2020). Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*.
- Zhang, K., Hu, B., and Basar, T. (2019a). Policy optimization for  $\mathcal{H}_2$  linear control with  $\mathcal{H}_\infty$  robustness guarantee: Implicit regularization and global convergence. *arXiv preprint arXiv:1910.09496*.
- Zhang, K., Koppel, A., Zhu, H., and Başar, T. (2019b). Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383*.

## A Preliminaries

### A.1 Derivation of entropy-regularized NPG methods

This subsection establishes the equivalence between the update rules (15) and (18). Such derivations are inherently similar to the ones for the NPG update rule (without entropy regularization) (see, e.g., Agarwal et al. (2019)); we provide the proof here for pedagogical reasons.

First of all, let us follow the convention to introduce the advantage function  $A_\tau^\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  of a policy  $\pi$  w.r.t. the entropy-regularized value function:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad A_\tau^\pi(s, a) := Q_\tau^\pi(s, a) - \tau \log \pi(a|s) - V_\tau^\pi(s) \quad (61)$$

with  $Q_\tau^\pi$  defined in (11a), which reflects the gain one can harvest by executing action  $a$  instead of following the policy  $\pi$  in state  $s$ . This advantage function plays a crucial role in the calculation of policy gradients, due to the following fundamental relation (see Appendix C.6 for the proof):

**Lemma 6.** *Under softmax parameterization (7), the gradient of the regularized value function satisfies*

$$\frac{\partial V_\tau^{\pi_\theta}(\rho)}{\partial \theta(s, a)} = \frac{1}{1 - \gamma} d_\rho^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A_\tau^{\pi_\theta}(s, a); \quad (62a)$$

$$\left[ (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V_\tau^{\pi_\theta}(\rho) \right] (s, a) = \frac{1}{1 - \gamma} A_\tau^{\pi_\theta}(s, a) + c(s) \quad (62b)$$

for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , where  $c(s) := \sum_a \pi_\theta(a|s) w_{s,a}$  is some function depending only on  $s$ .

It is worth highlighting that the search direction of NPG, given in (62b), is invariant to the choice of  $\rho$ . With the above calculations in place, it is seen that for any  $s \in \mathcal{S}$ , the regularized NPG update rule (15) results in a policy update as follows

$$\begin{aligned} \pi^{(t+1)}(a|s) &\stackrel{(i)}{\propto} \exp\left(\theta^{(t+1)}(s, a)\right) \stackrel{(ii)}{=} \exp\left(\theta^{(t)}(s, a) + \eta \left[ (\mathcal{F}_\rho^{\theta^{(t)}})^\dagger \nabla_\theta V_\tau^{(t)}(\rho) \right] (s, a)\right) \\ &\stackrel{(iii)}{\propto} \exp\left(\theta^{(t)}(s, a) + \frac{\eta}{1 - \gamma} A_\tau^{(t)}(s, a)\right) \\ &\stackrel{(iv)}{\propto} \pi^{(t)}(a|s) \exp\left(\frac{\eta}{1 - \gamma} Q_\tau^{(t)}(s, a) - \frac{\eta\tau}{1 - \gamma} \log \pi^{(t)}(a|s)\right) \\ &= \left(\pi^{(t)}(a|s)\right)^{1 - \frac{\eta\tau}{1 - \gamma}} \exp\left(\frac{\eta}{1 - \gamma} Q_\tau^{(t)}(s, a)\right). \end{aligned}$$

where we use  $A_\tau^{(t)}$  to abbreviate  $A_\tau^{\pi^{(t)}}$ . Here, (i) uses the definition of the softmax policy, (ii) comes from the update rule (15), (iii) is a consequence of (62b) (since  $c(\cdot)$  does not depend on  $a$ ), whereas (iv) results from the definition (61) and the fact that  $V_\tau^\pi(\cdot)$  is not dependent on  $a$ . This validates the equivalence between (15) and (18).

### A.2 Basic facts about the function $\log(\|\exp(\theta)\|_1)$

In the current paper, we often encounter the function  $\log(\|\exp(\theta)\|_1) := \log(\sum_{1 \leq a \leq |\mathcal{A}|} \exp(\theta_a))$  for any vector  $\theta = [\theta_a]_{1 \leq a \leq |\mathcal{A}|} \in \mathbb{R}^{|\mathcal{A}|}$ . To facilitate analysis, we single out several basic properties concerning this function, which will be used multiple times when establishing our main results. For notational convenience, we denote by  $\pi_\theta \in \mathbb{R}^{|\mathcal{A}|}$  the softmax transform of  $\theta$  such that

$$\pi_\theta(a) = \frac{\exp(\theta_a)}{\sum_{1 \leq j \leq |\mathcal{A}|} \exp(\theta_j)}, \quad 1 \leq a \leq |\mathcal{A}|. \quad (63)$$

By straightforward calculations, the gradient of the function  $\log(\|\exp(\theta)\|_1)$  is given by

$$\nabla_\theta \log(\|\exp(\theta)\|_1) = \frac{1}{\|\exp(\theta)\|_1} \exp(\theta) = \pi_\theta; . \quad (64)$$



**Difference of log policies.** In the analysis, we often need to control the difference of two policies, towards which the following bounds prove useful. To begin with, the mean value theorem reveals a Lipschitz continuity property (w.r.t. the  $\ell_\infty$  norm): for any  $\theta_1, \theta_2 \in \mathbb{R}^{|\mathcal{A}|}$ ,

$$\begin{aligned} |\log(\|\exp(\theta_1)\|_1) - \log(\|\exp(\theta_2)\|_1)| &= |\langle \theta_1 - \theta_2, \nabla_\theta \log(\|\exp(\theta)\|_1)|_{\theta=\theta_c} \rangle| \\ &\leq \|\theta_1 - \theta_2\|_\infty \|\nabla_\theta \log(\|\exp(\theta)\|_1)|_{\theta=\theta_c}\|_1 = \|\theta_1 - \theta_2\|_\infty, \end{aligned} \quad (65)$$

where  $\theta_c$  is a certain convex combination of  $\theta_1$  and  $\theta_2$ , and the second line relies on (64). In addition, for any two vectors  $\pi_{\theta_1}$  and  $\pi_{\theta_2}$  defined w.r.t.  $\theta_1, \theta_2 \in \mathbb{R}^{|\mathcal{A}|}$  (see (63)), one has

$$\|\log \pi_{\theta_1} - \log \pi_{\theta_2}\|_\infty \leq 2 \|\theta_1 - \theta_2\|_\infty, \quad (66)$$

where  $\log(\cdot)$  denotes entrywise operation. To justify (66), we observe from the definition (63) that

$$\|\log \pi_{\theta_1} - \log \pi_{\theta_2}\|_\infty \leq \|\theta_1 - \theta_2\|_\infty + \left| \log(\|\exp(\theta_1)\|_1) - \log(\|\exp(\theta_2)\|_1) \right| \leq 2 \|\theta_1 - \theta_2\|_\infty,$$

where the last inequality is a consequence of (65).

## B Proof for the bandit case (Proposition 1)

We start by defining an auxiliary sequence  $\xi^{(t)} \in \mathbb{R}^{|\mathcal{A}|}$  ( $t \geq 0$ ) recursively as follows

$$\begin{aligned} \xi^{(0)} &:= \|\exp(r/\tau)\|_1 \cdot \pi^{(0)}, \\ \xi^{(t+1)}(a) &:= (\xi^{(t)}(a))^{1-\tau\eta} \exp(\eta r(a)), \quad a \in \mathcal{A}. \end{aligned}$$

When combined with (21), it is easily seen that  $\pi^{(t)}(\cdot) \propto \xi^{(t)}(\cdot)$  and, as a result,  $\pi^{(t)} = \xi^{(t)} / \|\xi^{(t)}\|_1$ .

By construction, the auxiliary sequence satisfies the following property

$$\begin{aligned} \log(\xi^{(t+1)}(a)) - r(a)/\tau &= (1 - \tau\eta) \log(\xi^{(t)}(a)) + \eta r(a) - r(a)/\tau \\ &= (1 - \tau\eta) \left( \log(\xi^{(t)}(a)) - r(a)/\tau \right), \end{aligned}$$

thus indicating that

$$\left\| \log \xi^{(t)} - r/\tau \right\|_\infty \leq (1 - \tau\eta)^t \left\| \log \xi^{(0)} - r/\tau \right\|_\infty. \quad (67)$$

This taken together with the optimal policy  $\pi_\tau^* = \text{softmax}(r/\tau) \propto \exp(r/\tau)$  leads to

$$\begin{aligned} \left\| \log \pi^{(t)} - \log \pi_\tau^* \right\|_\infty &\leq 2 \left\| \log \xi^{(t)} - r/\tau \right\|_\infty \leq 2(1 - \tau\eta)^t \left\| \log \xi^{(0)} - r/\tau \right\|_\infty \\ &= 2(1 - \tau\eta)^t \left\| \log \pi^{(0)} + (\log \|\exp(r/\tau)\|_1) \cdot \mathbf{1} - r/\tau \right\|_\infty \\ &= 2(1 - \tau\eta)^t \left\| \log \pi^{(0)} - \log \pi_\tau^* \right\|_\infty, \end{aligned}$$

where the first line follows from the inequality (66), the second line follows from the expression (67), whereas the last line follows from the form of  $\pi_\tau^*$ . We have thus completed the proof of Proposition 1.

## C Proof for key lemmas

### C.1 Proof of Lemma 1

To begin with, the regularized NPG update rule (see (18) in Algorithm 1) indicates that

$$\log \pi^{(t+1)}(a|s) = \left(1 - \frac{\eta\tau}{1-\gamma}\right) \log \pi^{(t)}(a|s) + \frac{\eta}{1-\gamma} Q_\tau^{(t)}(s, a) - \log Z^{(t)}(s), \quad (68)$$

where  $Z^{(t)}$  is some quantity depending only on the state  $s$  (but not the action  $a$ ). Rearranging terms gives

$$-\tau \log \pi^{(t)}(a|s) + Q_\tau^{(t)}(s, a) = \frac{1-\gamma}{\eta} \left( \log \pi^{(t+1)}(a|s) - \log \pi^{(t)}(a|s) \right) + \frac{1-\gamma}{\eta} \log Z^{(t)}(s). \quad (69)$$

This in turn allows us to express  $V_\tau^{(t)}(s_0)$  for any  $s_0 \in \mathcal{S}$  as follows

$$\begin{aligned} V_\tau^{(t)}(s_0) &= \mathbb{E}_{a_0 \sim \pi^{(t)}(\cdot|s_0)} \left[ -\tau \log \pi^{(t)}(a_0|s_0) + Q_\tau^{(t)}(s_0, a_0) \right] \\ &= \mathbb{E}_{a_0 \sim \pi^{(t)}(\cdot|s_0)} \left[ \frac{1-\gamma}{\eta} \log Z^{(t)}(s_0) \right] + \mathbb{E}_{a_0 \sim \pi^{(t)}(\cdot|s_0)} \left[ \frac{1-\gamma}{\eta} \left( \log \pi^{(t+1)}(a_0|s_0) - \log \pi^{(t)}(a_0|s_0) \right) \right] \\ &= \frac{1-\gamma}{\eta} \log Z^{(t)}(s_0) - \frac{1-\gamma}{\eta} \text{KL} \left( \pi^{(t)}(\cdot|s_0) \parallel \pi^{(t+1)}(\cdot|s_0) \right) \\ &= \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ \frac{1-\gamma}{\eta} \log Z^{(t)}(s_0) \right] - \frac{1-\gamma}{\eta} \text{KL} \left( \pi^{(t)}(\cdot|s_0) \parallel \pi^{(t+1)}(\cdot|s_0) \right), \end{aligned} \quad (70)$$

where the first identity makes use of the definitions (8) and (11a), the second line follows from (69), the third line relies on the definition of the KL divergence, and the last line follows since  $Z^{(t)}(s)$  does not depend on  $a$ . Invoking (69) again to rewrite  $\log Z^{(t)}(s_0)$  appearing in the first term of (70), we reach

$$\begin{aligned} V_\tau^{(t)}(s_0) &= \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ -\tau \log \pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0, a_0) + \left( \tau - \frac{1-\gamma}{\eta} \right) \left( \log \pi^{(t+1)}(a_0|s_0) - \log \pi^{(t)}(a_0|s_0) \right) \right] \\ &\quad - \frac{1-\gamma}{\eta} \text{KL} \left( \pi^{(t)}(\cdot|s_0) \parallel \pi^{(t+1)}(\cdot|s_0) \right) \\ &= \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ -\tau \log \pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0, a_0) \right] + \left( \tau - \frac{1-\gamma}{\eta} \right) \text{KL} \left( \pi^{(t+1)}(\cdot|s_0) \parallel \pi^{(t)}(\cdot|s_0) \right) \\ &\quad - \frac{1-\gamma}{\eta} \text{KL} \left( \pi^{(t)}(\cdot|s_0) \parallel \pi^{(t+1)}(\cdot|s_0) \right) \\ &= \mathbb{E}_{\substack{a_0 \sim \pi^{(t+1)}(\cdot|s_0), \\ s_1 \sim P(\cdot|s_0, a_0)}} \left[ -\tau \log \pi^{(t+1)}(a_0|s_0) + r(s_0, a_0) + \gamma V_\tau^{(t)}(s_1) \right] \\ &\quad - \left( \frac{1-\gamma}{\eta} - \tau \right) \text{KL} \left( \pi^{(t+1)}(\cdot|s_0) \parallel \pi^{(t)}(\cdot|s_0) \right) - \frac{1-\gamma}{\eta} \text{KL} \left( \pi^{(t)}(\cdot|s_0) \parallel \pi^{(t+1)}(\cdot|s_0) \right), \end{aligned} \quad (71)$$

where the second line uses the definition of the KL divergence, and the third line expands  $Q_\tau^{(t)}$  using the definition (11a).

To finish up, applying the above relation (71) recursively to expand  $V_\tau^{(t)}(s_i)$  ( $i \geq 1$ ), we arrive at

$$\begin{aligned} V_\tau^{(t)}(s_0) &= \mathbb{E}_{\substack{a_i \sim \pi^{(t+1)}(\cdot|s_i), \\ s_{i+1} \sim P(\cdot|s_i, a_i), \forall i \geq 0}} \left[ \sum_{i=0}^{\infty} \gamma^i \left\{ r(s_i, a_i) - \tau \log \pi^{(t+1)}(a_i|s_i) \right\} \right. \\ &\quad \left. - \sum_{i=0}^{\infty} \gamma^i \left\{ \left( \frac{1-\gamma}{\eta} - \tau \right) \text{KL} \left( \pi^{(t+1)}(\cdot|s_i) \parallel \pi^{(t)}(\cdot|s_i) \right) + \frac{1-\gamma}{\eta} \text{KL} \left( \pi^{(t)}(\cdot|s_i) \parallel \pi^{(t+1)}(\cdot|s_i) \right) \right\} \right] \\ &= V_\tau^{(t+1)}(s_0) - \mathbb{E}_{s \sim d_{s_0}^{(t+1)}} \left[ \left( \frac{1}{\eta} - \frac{\tau}{1-\gamma} \right) \text{KL} \left( \pi^{(t+1)}(\cdot|s) \parallel \pi^{(t)}(\cdot|s) \right) + \frac{1}{\eta} \text{KL} \left( \pi^{(t)}(\cdot|s) \parallel \pi^{(t+1)}(\cdot|s) \right) \right], \end{aligned} \quad (72)$$

where the second line follows since the regularized value function  $V_\tau^{(t+1)}$  can be viewed as the value function of  $\pi^{(t+1)}$  with adjusted rewards  $r_\tau^{(t+1)}(s, a) := r(s, a) - \tau \log \pi^{(t+1)}(a|s)$ . Averaging the initial state  $s_0$  over the distribution  $\rho$  concludes the proof.

## C.2 Proof of Lemma 2

In the sequel, we prove each claim in Lemma 2 in order.

**Proof of Eqn. (36).** Jensen's inequality tells us that: for any  $s \in \mathcal{S}$  one has

$$\begin{aligned} \mathbb{E}_{a \sim \pi(\cdot|s)} \left[ Q(s, a) - \tau \log \pi(a|s) \right] &= \tau \sum_a \pi(a|s) \log \left( \frac{\exp(Q(s, a)/\tau)}{\pi(a|s)} \right) \\ &\leq \tau \log \left( \sum_a \pi(a|s) \frac{\exp(Q(s, a)/\tau)}{\pi(a|s)} \right) \\ &= \tau \log \left( \sum_a \exp(Q(s, a)/\tau) \right) = \tau \log (\|\exp(Q(s, \cdot)/\tau)\|_1), \end{aligned} \quad (73)$$

where in the second line, equality is attained if  $\pi(\cdot|s) \propto \exp(Q(s, \cdot)/\tau)$ . This immediately gives rise to

$$\begin{aligned} \mathcal{T}_\tau(Q)(s, a) &= r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \max_{\pi(\cdot|s') \in \Delta(\mathcal{A})} \mathbb{E}_{a' \sim \pi(\cdot|s')} \left[ Q(s', a') - \tau \log \pi(a'|s') \right] \right] \\ &= r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \tau \log (\|\exp(Q(s', \cdot)/\tau)\|_1) \right]. \end{aligned}$$

**Proof of Eqn. (37).** Recall the characterization of  $\pi_\tau^*$  and  $V_\tau^*$  established in Nachum et al. (2017):

$$\pi_\tau^*(a|s) = \exp \left( \frac{Q_\tau^*(s, a) - V_\tau^*(s)}{\tau} \right), \quad (74a)$$

$$V_\tau^*(s) = \tau \log (\|\exp(Q_\tau^*(s, \cdot)/\tau)\|_1). \quad (74b)$$

Substitution into the expression (36) tells us that for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ ,

$$\begin{aligned} \mathcal{T}_\tau(Q_\tau^*)(s, a) &= r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \tau \log (\|\exp(Q_\tau^*(s', \cdot)/\tau)\|_1) \right] \\ &= r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ V_\tau^*(s') \right] \\ &= Q_\tau^*(s, a), \end{aligned}$$

where the second line results from (74b), and the last line follows from the definition of the soft Q-function.

**Proof of Eqn. (38).** Invoking again the expression (36), we can demonstrate that for any  $Q_1$  and  $Q_2$ ,

$$\begin{aligned} &\left| \mathcal{T}_\tau(Q_1)(s, a) - \mathcal{T}_\tau(Q_2)(s, a) \right| \\ &= \left| \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \tau \log (\|\exp(Q_1(s', \cdot)/\tau)\|_1) \right] - \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \tau \log (\|\exp(Q_2(s', \cdot)/\tau)\|_1) \right] \right| \\ &= \gamma \tau \left| \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \log (\|\exp(Q_1(s', \cdot)/\tau)\|_1) - \log (\|\exp(Q_2(s', \cdot)/\tau)\|_1) \right] \right| \\ &\leq \gamma \tau \|Q_1/\tau - Q_2/\tau\|_\infty \\ &= \gamma \|Q_1 - Q_2\|_\infty \end{aligned}$$

holds for all  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , where the inequality follows from the Lipschitz property (65).

### C.3 Proof of Lemma 3

For any state-action pair  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , we observe that

$$\begin{aligned}
& Q_\tau^*(s, a) - Q_\tau^{(t+1)}(s, a) \\
&= r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} [V_\tau^*(s')] - \left( r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} [V_\tau^{(t+1)}(s')] \right) \\
&= \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \tau \log \left( \left\| \exp \left( \frac{Q_\tau^*(s', \cdot)}{\tau} \right) \right\|_1 \right) \right] - \gamma \mathbb{E}_{\substack{s' \sim P(\cdot|s, a), \\ a' \sim \pi^{(t+1)}(\cdot|s')}} \left[ Q_\tau^{(t+1)}(s', a') - \tau \log \pi^{(t+1)}(a'|s') \right], \quad (75)
\end{aligned}$$

where the first step invokes the definition (11a) of  $Q_\tau$ , and the second step is due to the expression (74b) of  $V_\tau^*$ . To continue, recall that  $\pi^{(t)}$  is related to  $\xi^{(t)}$  as

$$\forall s \in \mathcal{S}: \quad \pi^{(t)}(\cdot|s) = \frac{1}{\|\xi^{(t)}(s, \cdot)\|_1} \xi^{(t)}(s, \cdot) \quad (76)$$

which can be seen by comparing (42) with (18). This in turn leads to

$$\begin{aligned}
\log \pi^{(t+1)}(a|s) &= \log \xi^{(t+1)}(s, a) - \log (\|\xi^{(t+1)}(s, \cdot)\|_1) \\
&= \alpha \log \xi^{(t)}(s, a) + (1 - \alpha) \frac{Q_\tau^{(t)}(s, a)}{\tau} - \log (\|\xi^{(t+1)}(s, \cdot)\|_1), \quad (77)
\end{aligned}$$

where the second line comes from (42b). By plugging (77) into (75) we obtain

$$\begin{aligned}
Q_\tau^*(s, a) - Q_\tau^{(t+1)}(s, a) &= \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ \tau \log (\|\exp(Q_\tau^*(s', \cdot)/\tau)\|_1) - \tau \log (\|\xi^{(t+1)}(s', \cdot)\|_1) \right] \\
&\quad - \gamma \mathbb{E}_{\substack{s' \sim P(\cdot|s, a), \\ a' \sim \pi^{(t+1)}(\cdot|s')}} \left[ \underbrace{Q_\tau^{(t+1)}(s', a') - \tau \left( \alpha \log \xi^{(t)}(s', a') + (1 - \alpha) \frac{Q_\tau^{(t)}(s', a')}{\tau} \right)}_{= \log \xi^{(t+1)}(s', a')} \right] \quad (78)
\end{aligned}$$

for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . In the sequel, we bound each term on the right-hand side of (78) separately.

- In view of the property (65), the first term on the right-hand side of (78) can be bounded by  $\tau \log (\|\exp(Q_\tau^*(s', \cdot)/\tau)\|_1) - \tau \log (\|\xi^{(t+1)}(s', \cdot)\|_1) \leq \|Q_\tau^* - \tau \log \xi^{(t+1)}\|_\infty$ .
- Regarding the second term, the monotonicity (33) of the soft Q-function allows us to derive

$$\begin{aligned}
& Q_\tau^{(t+1)}(s, a) - \tau \left( \alpha \log \xi^{(t)}(s, a) + (1 - \alpha) \frac{Q_\tau^{(t)}(s, a)}{\tau} \right) \\
&\geq Q_\tau^{(t)}(s, a) - \tau \left( \alpha \log \xi^{(t)}(s, a) + (1 - \alpha) \frac{Q_\tau^{(t)}(s, a)}{\tau} \right) \\
&= \alpha \left( Q_\tau^{(t)}(s, a) - \tau \log \xi^{(t)}(s, a) \right) \\
&\stackrel{(i)}{=} \alpha \left( \alpha \left( Q_\tau^{(t-1)}(s, a) - \tau \log \xi^{(t-1)}(s, a) \right) + Q_\tau^{(t)}(s, a) - Q_\tau^{(t-1)}(s, a) \right) \\
&\stackrel{(ii)}{\geq} \alpha^2 \left( Q_\tau^{(t-1)}(s, a) - \tau \log \xi^{(t-1)}(s, a) \right) \\
&\stackrel{(iii)}{\geq} \alpha^{t+1} \left( Q_\tau^{(0)}(s, a) - \tau \log \xi^{(0)}(s, a) \right) \\
&\stackrel{(iv)}{\geq} -\alpha^{t+1} \|Q_\tau^{(0)} - \tau \log \xi^{(0)}\|_\infty
\end{aligned}$$

for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . Here, (i) follows by construction (42b), (ii) invokes the monotonicity property (33) (so that  $Q_\tau^{(t)} \geq Q_\tau^{(t-1)}$ ), and (iii) follows by repeating the arguments (i) and (ii) recursively.

Combining the preceding two bounds with the expression (78), we conclude that

$$0 \leq Q_\tau^*(s, a) - Q_\tau^{(t+1)}(s, a) \leq \gamma \|Q_\tau^* - \tau \log \xi^{(t+1)}\|_\infty + \gamma \alpha^{t+1} \|Q_\tau^{(0)} - \tau \log \xi^{(0)}\|_\infty \quad (79)$$

for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , thus concluding the proof.

## C.4 Proof of Lemma 4

Recall that, in this scenario, the policies are updated using inexact policy evaluation via (26), namely,

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}, \quad \pi^{(t+1)}(a|s) = \frac{(\pi^{(t)}(a|s))^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta}{1-\gamma} \widehat{Q}_\tau^{(t)}(s, a)\right)}{\widehat{Z}^{(t)}(s)}, \quad (80)$$

where  $\widehat{Z}^{(t)}(s) := \sum_{a'} \pi^{(t)}(a'|s)^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta}{1-\gamma} \widehat{Q}_\tau^{(t)}(s, a')\right)$ . To facilitate analysis, we further introduce another auxiliary policy sequence  $\{\check{\pi}^{(t)}\}$ , which corresponds to the policy update as if we had access to exact soft Q-function of  $\pi^{(t)}$  in the  $t$ -th iteration; this is defined as

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}, \quad \check{\pi}^{(t+1)}(a|s) = \frac{(\pi^{(t)}(a|s))^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta}{1-\gamma} Q_\tau^{(t)}(s, a)\right)}{Z^{(t)}(s)}, \quad (81)$$

where we abuse the notation by letting  $Z^{(t)}(s) := \sum_{a'} \pi^{(t)}(a'|s)^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta}{1-\gamma} Q_\tau^{(t)}(s, a')\right)$ . It is worth emphasizing that  $\check{\pi}^{(t+1)}$  is produced on the basis of  $\pi^{(t)}$  as opposed to  $\check{\pi}^{(t)}$ ; it should be viewed as a one-step perfect update from a given policy  $\pi^{(t)}$ .

We first make note of the following fact: for any step size  $0 < \eta \leq (1-\gamma)/\tau$ , it follows from (66) — together with the construction (80) and (81) — that

$$\begin{aligned} & \left\| \log \pi^{(t+1)} - \log \check{\pi}^{(t+1)} \right\|_\infty \\ & \leq 2 \left\| \log \left( \pi^{(t)}(a|s)^{1-\eta\tau/(1-\gamma)} \exp\left(\frac{\eta}{1-\gamma} \widehat{Q}_\tau^{(t)}(s, a)\right) \right) - \log \left( \pi^{(t)}(a|s)^{1-\eta\tau/(1-\gamma)} \exp\left(\frac{\eta}{1-\gamma} Q_\tau^{(t)}(s, a)\right) \right) \right\|_\infty \\ & = \frac{2\eta}{1-\gamma} \left\| \widehat{Q}_\tau^{(t)} - Q_\tau^{(t)} \right\|_\infty. \end{aligned} \quad (82)$$

Next, let us recall the inequality (70) in the proof of Lemma 1 under exact policy evaluation  $\check{\pi}^{(t+1)}(\cdot|s)$ ; when applied to the current setting, it essentially indicates that

$$\begin{aligned} V_\tau^{(t)}(s_0) &= \mathbb{E}_{a_0 \sim \check{\pi}^{(t+1)}(\cdot|s_0)} \left[ \frac{1-\gamma}{\eta} \log Z^{(t)}(s_0) \right] - \frac{1-\gamma}{\eta} \text{KL} \left( \pi^{(t)}(\cdot|s_0) \parallel \check{\pi}^{(t+1)}(\cdot|s_0) \right) \\ &= \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ \frac{1-\gamma}{\eta} \log Z^{(t)}(s_0) \right] - \frac{1-\gamma}{\eta} \text{KL} \left( \pi^{(t)}(\cdot|s_0) \parallel \check{\pi}^{(t+1)}(\cdot|s_0) \right), \end{aligned} \quad (83)$$

where the last step follows since the quantity  $Z^{(t)}(s)$  does not depend on  $a$  at all. In order to control the first term of (83), we invoke the definition of  $\check{\pi}^{(t+1)}(\cdot|s)$  to show that

$$\begin{aligned} & \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ \frac{1-\gamma}{\eta} \log Z^{(t)}(s_0) \right] \\ & \stackrel{(i)}{=} \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ -\tau \log \check{\pi}^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0, a_0) + \left( \tau - \frac{1-\gamma}{\eta} \right) \left( \log \check{\pi}^{(t+1)}(a_0|s_0) - \log \pi^{(t)}(a_0|s_0) \right) \right] \\ & = \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ -\tau \log \pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0, a_0) \right] + \left( \tau - \frac{1-\gamma}{\eta} \right) \text{KL} \left( \pi^{(t+1)}(\cdot|s_0) \parallel \pi^{(t)}(\cdot|s_0) \right) \\ & \quad - \frac{1-\gamma}{\eta} \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ \log \check{\pi}^{(t+1)}(a_0|s_0) - \log \pi^{(t)}(a_0|s_0) \right] \\ & \leq \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ -\tau \log \pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0, a_0) \right] + \left( \tau - \frac{1-\gamma}{\eta} \right) \text{KL} \left( \pi^{(t+1)}(\cdot|s_0) \parallel \pi^{(t)}(\cdot|s_0) \right) \\ & \quad + 2 \left\| \widehat{Q}_\tau^{(t)} - Q_\tau^{(t)} \right\|_\infty, \end{aligned} \quad (84)$$

where the final step results from (82). Putting the above bound together with (83) guarantees that

$$V_\tau^{(t)}(s_0) \leq \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ -\tau \log \pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0, a_0) \right] - \frac{1-\gamma}{\eta} \text{KL} \left( \pi^{(t)}(\cdot|s_0) \parallel \check{\pi}^{(t+1)}(\cdot|s_0) \right)$$

$$\begin{aligned}
& - \left( \frac{1-\gamma}{\eta} - \tau \right) \text{KL} \left( \pi^{(t+1)}(\cdot|s_0) \parallel \pi^{(t)}(\cdot|s_0) \right) + 2 \|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\|_\infty \\
\leq & \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ -\tau \log \pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0, a_0) \right] + 2 \|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\|_\infty \\
\leq & \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ -\tau \log \pi^{(t+1)}(a_0|s_0) + r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot|s_0, a_0)} [V_\tau^{(t)}(s_1)] \right] + 2 \|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\|_\infty.
\end{aligned}$$

where the last identity makes use of the relation  $Q_\tau^{(t)}(s_0, a_0) = r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot|s_0, a_0)} [V_\tau^{(t)}(s_1)]$ . Invoking the above inequality recursively as in the expression (72) (see Lemma 1), we can expand it to establish

$$V_\tau^{(t)}(s_0) \leq V_\tau^{(t+1)}(s_0) + 2 \|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\|_\infty \sum_{i=0}^{\infty} \gamma^i = V_\tau^{(t+1)}(s_0) + \frac{2}{1-\gamma} \|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\|_\infty.$$

## C.5 Proof of Lemma 5

First of all, we follow the definition (8) of the entropy-regularized value function to deduce that

$$\begin{aligned}
V_\tau^*(\rho) - V_\tau^{(t)}(\rho) &= \mathbb{E}_{\substack{s_0 \sim \rho, a_i \sim \pi_\tau^*(\cdot|s_i), \\ s_{i+1} \sim P(\cdot|s_i, a_i), \forall i \geq 0}} \left[ \sum_{i=0}^{\infty} \gamma^i (r(s_i, a_i) - \tau \log \pi_\tau^*(a_i|s_i)) \right] - V_\tau^{(t)}(\rho) \\
&= \mathbb{E}_{\substack{s_0 \sim \rho, a_i \sim \pi_\tau^*(\cdot|s_i), \\ s_{i+1} \sim P(\cdot|s_i, a_i), \forall i \geq 0}} \left[ \sum_{i=0}^{\infty} \gamma^i \left( r(s_i, a_i) - \tau \log \pi_\tau^*(a_i|s_i) + V_\tau^{(t)}(s_i) - V_\tau^{(t)}(s_{i+1}) \right) \right] - V_\tau^{(t)}(\rho) \\
&= \mathbb{E}_{\substack{s_0 \sim \rho, a_i \sim \pi_\tau^*(\cdot|s_i), \\ s_{i+1} \sim P(\cdot|s_i, a_i), \forall i \geq 0}} \left[ V_\tau^{(t)}(s_0) + \sum_{i=0}^{\infty} \gamma^i \left( r(s_i, a_i) - \tau \log \pi_\tau^*(a_i|s_i) + \gamma V_\tau^{(t)}(s_{i+1}) - V_\tau^{(t)}(s_i) \right) \right] - V_\tau^{(t)}(\rho) \\
&\stackrel{(i)}{=} \mathbb{E}_{\substack{s_0 \sim \rho, a_i \sim \pi_\tau^*(\cdot|s_i), \\ s_{i+1} \sim P(\cdot|s_i, a_i), \forall i \geq 0}} \left[ \sum_{i=0}^{\infty} \gamma^i \left( r(s_i, a_i) - \tau \log \pi_\tau^*(a_i|s_i) + \gamma V_\tau^{(t)}(s_{i+1}) - V_\tau^{(t)}(s_i) \right) \right] \\
&\stackrel{(ii)}{=} \frac{1}{1-\gamma} \mathbb{E}_{\substack{s \sim d_\rho^{\pi_\tau^*} \\ s \sim d_\rho^{\pi_\tau^*}}} \left[ \sum_a \pi_\tau^*(a|s) \left( r(s, a) - \tau \log \pi_\tau^*(a|s) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} [V_\tau^{(t)}(s')] - V_\tau^{(t)}(s) \right) \right] \\
&\stackrel{(iii)}{=} \frac{1}{1-\gamma} \mathbb{E}_{\substack{s \sim d_\rho^{\pi_\tau^*} \\ s \sim d_\rho^{\pi_\tau^*}}} \left[ \sum_a \pi_\tau^*(a|s) \left( Q_\tau^{(t)}(s, a) - \tau \log \pi_\tau^*(a|s) \right) - V_\tau^{(t)}(s) \right]. \tag{85}
\end{aligned}$$

Here, (i) is due to the definition  $V_\tau^{(t)}(\rho) = \mathbb{E}_{s_0 \sim \rho} [V_\tau^{(t)}(s_0)]$ , (ii) follows by aggregating terms corresponding to the same state-action pair and the definition of  $d_\rho^{\pi_\tau^*}$  (cf. (5)), whereas (iii) results from the definition (11a) of the regularized Q-function.

To continue, we shall attempt to control each part of (85) separately. To begin with, observe that the first part of (85) can be bounded by Jensen's inequality, namely,

$$\begin{aligned}
\sum_a \pi_\tau^*(a|s) \left( Q_\tau^{(t)}(s, a) - \tau \log \pi_\tau^*(a|s) \right) &= \tau \sum_a \pi_\tau^*(a|s) \log \left( \frac{\exp(Q_\tau^{(t)}(s, a)/\tau)}{\pi_\tau^*(a|s)} \right) \\
&\leq \tau \log \left( \sum_a \pi_\tau^*(a|s) \frac{\exp(Q_\tau^{(t)}(s, a)/\tau)}{\pi_\tau^*(a|s)} \right) \\
&= \tau \log \left( \sum_a \exp(Q_\tau^{(t)}(s, a)/\tau) \right). \tag{86}
\end{aligned}$$

With regards to the second part of (85), it is seen from the definition of  $\pi^{(t+1)}$  (cf. (17)) that

$$Q_\tau^{(t)}(s, a) = \tau \log \pi_\tau^{(t+1)}(a|s) + \tau \log \left( \sum_a \exp(Q_\tau^{(t)}(s, a)/\tau) \right), \tag{87}$$

thus allowing one to derive

$$\begin{aligned}
V_\tau^{(t)}(s) &= \sum_a \pi_\tau^{(t)}(a|s) \left( Q_\tau^{(t)}(s, a) - \tau \log \pi_\tau^{(t)}(a|s) \right) \\
&\stackrel{(i)}{=} \tau \sum_a \pi_\tau^{(t)}(a|s) \left\{ \log \pi_\tau^{(t+1)}(a|s) + \log \left( \sum_a \exp \left( Q_\tau^{(t)}(s, a) / \tau \right) \right) - \log \pi_\tau^{(t)}(a|s) \right\} \\
&= \tau \log \left( \sum_a \exp \left( Q_\tau^{(t)}(s, a) / \tau \right) \right) + \tau \sum_a \pi_\tau^{(t)}(a|s) \left( \log \pi_\tau^{(t+1)}(a|s) - \log \pi_\tau^{(t)}(a|s) \right) \\
&= \tau \log \left( \sum_a \exp \left( Q_\tau^{(t)}(s, a) / \tau \right) \right) - \tau \text{KL} \left( \pi_\tau^{(t)}(a|s) \parallel \pi_\tau^{(t+1)}(\cdot|s) \right), \tag{88}
\end{aligned}$$

where (i) relies on the identity (87). Substituting the inequalities (86) and (88) into the expression (85), we can demonstrate with a little algebra that

$$V_\tau^*(\rho) - V_\tau^{(t)}(\rho) \leq \frac{1}{\eta} \mathbb{E}_{s \sim d_\rho^{\pi_\tau^*}} \left[ \text{KL} \left( \pi^{(t)}(\cdot|s) \parallel \pi^{(t+1)}(\cdot|s) \right) \right].$$

## C.6 Proof of Lemma 6

The results of this lemma, or some similar versions, have appeared in prior work (e.g. Mei et al. (2020, Lemma 10) and Agarwal et al. (2020b, Lemma 5.6)). We include the proof here primarily for the sake of self-completeness.

**Proof of Eqn. (62a).** The policy gradient of the unregularized value function  $V^{\pi_\theta}(s_0)$  is well-known as the policy gradient theorem (Sutton et al., 2000). Here, we deal with a slightly different variant – an entropy-regularized value function  $V_\tau^{\pi_\theta}(s_0)$  in the expression (2) with the softmax policy parameterization in (7). Invoking the Bellman equation and recognizing that  $V_\tau^{\pi_\theta}(s_0)$  can be viewed as an unregularized value function with instantaneous rewards  $r(s, a) - \tau \log \pi_\theta(a|s)$  for any  $(s, a)$ , we obtain

$$\begin{aligned}
\nabla_\theta V_\tau^{\pi_\theta}(s_0) &= \nabla_\theta \left[ \sum_{a_0} \pi_\theta(a_0|s_0) \left( r(s_0, a_0) - \tau \log \pi_\theta(a_0|s_0) + \gamma \mathbb{E}_{s' \sim P(\cdot|s_0, a_0)} [V_\tau^{\pi_\theta}(s')] \right) \right] \\
&\stackrel{(i)}{=} \nabla_\theta \left[ \sum_{a_0} \pi_\theta(a_0|s_0) \left( Q_\tau^{\pi_\theta}(s_0, a_0) - \tau \log \pi_\theta(a_0|s_0) \right) \right] \\
&= \sum_{a_0} (\nabla_\theta \pi_\theta(a_0|s_0)) \left( Q_\tau^{\pi_\theta}(s_0, a_0) - \tau \log \pi_\theta(a_0|s_0) \right) + \sum_{a_0} \pi_\theta(a_0|s_0) \nabla_\theta \left( Q_\tau^{\pi_\theta}(s_0, a_0) - \tau \log \pi_\theta(a_0|s_0) \right) \\
&\stackrel{(ii)}{=} \sum_{a_0} \left( \pi_\theta(a_0|s_0) \nabla_\theta \log \pi_\theta(a_0|s_0) \right) \left( Q_\tau^{\pi_\theta}(s_0, a_0) - \tau \log \pi_\theta(a_0|s_0) \right) \\
&\quad + \sum_{a_0} \pi_\theta(a_0|s_0) \nabla_\theta \left( r(s_0, a_0) + \gamma \sum_{s_1} P(s_1|s_0, a_0) V_\tau^{\pi_\theta}(s_1) - \tau \log \pi_\theta(a_0|s_0) \right),
\end{aligned}$$

where (i) relies on the definition (11a) of  $Q_\tau^{\pi_\theta}$ , and (ii) makes use of the identity

$$\nabla_\theta \pi_\theta(a_0|s_0) = \pi_\theta(a_0|s_0) \nabla_\theta \log \pi_\theta(a_0|s_0)$$

as well as the definition (11a) of  $Q_\tau^{\pi_\theta}$ . Given that

$$\sum_{a_0} \pi_\theta(a_0|s_0) \nabla_\theta \log \pi_\theta(a_0|s_0) = \sum_{a_0} \nabla_\theta \pi_\theta(a_0|s_0) = \nabla_\theta \left( \sum_{a_0} \pi_\theta(a_0|s_0) \right) = \nabla_\theta 1 = 0 \tag{89}$$

and that  $r(s, a)$  is independent of  $\theta$ , one can continue the above derivative to reach

$$\nabla_\theta V_\tau^{\pi_\theta}(s_0) = \sum_{a_0} \left( \pi_\theta(a_0|s_0) \nabla_\theta \log \pi_\theta(a_0|s_0) \right) \left( Q_\tau^{\pi_\theta}(s_0, a_0) - \tau \log \pi_\theta(a_0|s_0) \right)$$

$$\begin{aligned}
& + \gamma \sum_{a_0} \pi_\theta(a_0|s_0) \sum_{s_1} P(s_1|s_0, a_0) \nabla_\theta V_\tau^{\pi_\theta}(s_1) \\
= & \mathbb{E}_{\substack{a_i \sim \pi_\theta(\cdot|s_i), \\ s_{i+1} \sim P(\cdot|s_i, a_i), \forall i \geq 0}} \left[ (\nabla_\theta \log \pi_\theta(a_0|s_0)) \left( Q_\tau^{\pi_\theta}(s_0, a_0) - \tau \log \pi_\theta(a_0|s_0) \right) + \gamma \nabla_\theta V_\tau^{\pi_\theta}(s_1) \right].
\end{aligned}$$

Repeating the above calculations recursively, we arrive at

$$\begin{aligned}
\nabla_\theta V_\tau^{\pi_\theta}(s_0) & = \mathbb{E}_{\substack{a_i \sim \pi_\theta(\cdot|s_i), \\ s_{i+1} \sim P(\cdot|s_i, a_i), \forall i \geq 0}} \left[ \sum_{t=0}^{\infty} \gamma^t (\nabla_\theta \log \pi_\theta(a_t|s_t)) \left( Q_\tau^{\pi_\theta}(s_t, a_t) - \tau \log \pi_\theta(a_t|s_t) \right) \right] \\
& = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[ (\nabla_\theta \log \pi_\theta(a|s)) \left( Q_\tau^{\pi_\theta}(s, a) - \tau \log \pi_\theta(a|s) \right) \right] \\
& = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[ (\nabla_\theta \log \pi_\theta(a|s)) \left( A_\tau^{\pi_\theta}(s, a) + V_\tau^{\pi_\theta}(s) \right) \right] \\
& = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[ (\nabla_\theta \log \pi_\theta(a|s)) A_\tau^{\pi_\theta}(s, a) \right], \tag{90}
\end{aligned}$$

where the second line follows by aggregating the terms corresponding to the same state-action pair, and the third line invokes the definition (61) of  $A_\tau^{\pi_\theta}$ . To see why the last line holds, invoke (89) to reach

$$\begin{aligned}
\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[ V_\tau^{\pi_\theta}(s) \nabla_\theta \log \pi_\theta(a|s) \right] & = \sum_a V_\tau^{\pi_\theta}(s) \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) \\
& = V_\tau^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) = 0.
\end{aligned}$$

Further, it is easily seen that under the softmax parametrization in (7),

$$\frac{\partial \log \pi_\theta(a'|s')}{\partial \theta(s, a)} = \mathbf{1}[s' = s] (\mathbf{1}[a' = a] - \pi_\theta(a|s)) \tag{91}$$

for any  $(s, a), (s', a') \in \mathcal{S} \times \mathcal{A}$ . Combining with (90), it further implies that

$$\begin{aligned}
\frac{\partial V_\tau^{\pi_\theta}(s_0)}{\partial \theta(s, a)} & = \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a' \sim \pi_\theta(\cdot|s')} \left[ \frac{\partial \log \pi_\theta(a'|s')}{\partial \theta(s, a)} A_\tau^{\pi_\theta}(s', a') \right] \\
& = \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a' \sim \pi_\theta(\cdot|s')} \left[ \left( \mathbf{1}[s' = s] (\mathbf{1}[a' = a] - \pi_\theta(a|s)) \right) A_\tau^{\pi_\theta}(s', a') \right] \\
& \stackrel{(i)}{=} \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a' \sim \pi_\theta(\cdot|s')} \left[ \mathbf{1}[(s', a') = (s, a)] A_\tau^{\pi_\theta}(s', a') \right] \\
& = \frac{1}{1-\gamma} d_{s_0}^{\pi_\theta}(s) \pi_\theta(a|s) A_\tau^{\pi_\theta}(s, a).
\end{aligned}$$

where (i) follows from  $\mathbb{E}_{a' \sim \pi_\theta(\cdot|s')} A_\tau^{\pi_\theta}(s', a') = \sum_{a'} \pi_\theta(a'|s') A_\tau^{\pi_\theta}(s', a') = 0$  due to the definition (61). The proof regarding  $V_\tau^{\pi_\theta}(\rho)$  can be obtained by averaging the initial state  $s_0$  over the distribution  $\rho$ .

**Proof of Eqn. (62b).** In order to establish (62b), a crucial observation is that  $w_\theta := (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V_\tau^{\pi_\theta}(\rho)$  is exactly the solution to the following least-squares problem

$$\text{minimize}_{w \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}} \|\mathcal{F}_\rho^\theta w - \nabla_\theta V_\tau^{\pi_\theta}(\rho)\|_2^2. \tag{92}$$

From the definition (14) of the Fisher information matrix, we have

$$\mathcal{F}_\rho^\theta w = \mathbb{E}_{s \sim d_\rho^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[ (\nabla_\theta \log \pi_\theta(a|s)) (\nabla_\theta \log \pi_\theta(a|s))^\top w \right].$$



for any fixed vector  $w = [w_{s,a}]_{(s,a) \in \mathcal{S} \times \mathcal{A}}$ . As a result, for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$  one has

$$\begin{aligned}
(\mathcal{F}_\rho^\theta w)_{s,a} &= \mathbb{E}_{s' \sim d_\rho^{\pi_\theta}} \mathbb{E}_{a' \sim \pi_\theta(\cdot|s')} \left[ \frac{\partial \log \pi_\theta(a'|s')}{\partial \theta(s, a)} \left( \sum_{\tilde{s}, \tilde{a}} \frac{\partial \log \pi_\theta(a'|s')}{\partial \theta(\tilde{s}, \tilde{a})} w_{\tilde{s}, \tilde{a}} \right) \right] \\
&\stackrel{(i)}{=} \mathbb{E}_{s' \sim d_\rho^{\pi_\theta}} \mathbb{E}_{a' \sim \pi_\theta(\cdot|s')} \left[ \mathbf{1}[s' = s] \left( \mathbf{1}[a' = a] - \pi_\theta(a|s) \right) \left( \sum_{\tilde{s}, \tilde{a}} \mathbf{1}[\tilde{s} = s'] \left( \mathbf{1}[\tilde{a} = a'] - \pi_\theta(\tilde{a}|\tilde{s}) \right) w_{\tilde{s}, \tilde{a}} \right) \right] \\
&= \mathbb{E}_{s' \sim d_\rho^{\pi_\theta}} \mathbb{E}_{a' \sim \pi_\theta(\cdot|s')} \left[ \mathbf{1}[s' = s] \left( \mathbf{1}[a' = a] - \pi_\theta(a|s) \right) \left( w_{s', a'} - \sum_{\tilde{a}} \pi_\theta(\tilde{a}|s') w_{s', \tilde{a}} \right) \right] \\
&= d_\rho^{\pi_\theta}(s) \mathbb{E}_{a' \sim \pi_\theta(\cdot|s')} \left[ \left( \mathbf{1}[a' = a] - \pi_\theta(a|s) \right) (w_{s, a'} - c(s)) \right] \\
&= d_\rho^{\pi_\theta}(s) \mathbb{E}_{a' \sim \pi_\theta(\cdot|s')} \left[ \mathbf{1}[a' = a] w_{s, a'} - \pi_\theta(a|s) w_{s, a'} - \mathbf{1}[a' = a] c(s) + \pi_\theta(a|s) c(s) \right] \\
&= d_\rho^{\pi_\theta}(s) \left[ \pi_\theta(a|s) w_{s, a} - \pi_\theta(a|s) c(s) - \pi_\theta(a|s) c(s) + \pi_\theta(a|s) c(s) \right] \\
&= d_\rho^{\pi_\theta}(s) \pi_\theta(a|s) [w_{s, a} - c(s)],
\end{aligned}$$

where (i) makes use of the derivative calculation (91), and we define  $c(s) := \sum_a \pi_\theta(a|s) w_{s, a}$ . Consequently, the objective function of (92) can be written as

$$\begin{aligned}
\|\mathcal{F}_\rho^\theta w - \nabla_\theta V_\tau^{\pi_\theta}(\rho)\|_2^2 &= \sum_{s, a} \left( d_\rho^{\pi_\theta}(s) \pi_\theta(a|s) [w_{s, a} - c(s)] - \frac{1}{1-\gamma} d_\rho^{\pi_\theta}(s) \pi_\theta(a|s) A_\tau^{\pi_\theta}(s, a) \right)^2 \\
&= \sum_{s, a} \left( d_\rho^{\pi_\theta}(s) \pi_\theta(a|s) \left( w_{s, a} - c(s) - \frac{1}{1-\gamma} A_\tau^{\pi_\theta}(s, a) \right) \right)^2,
\end{aligned}$$

which is minimized by choosing  $w_{s, a} = \frac{1}{1-\gamma} A_\tau^{\pi_\theta}(s, a) + c(s)$  for all  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . This concludes the proof.

## D Convergence guarantees for CPI-style policy updates

Employing the SPI update as the improved policy, we arrive at the following CPI-style update

$$\pi^{(t+1)} = (1 - \beta) \pi^{(t)} + \beta \bar{\pi}^{(t+1)}. \tag{93a}$$

Here,  $\bar{\pi}^{(t+1)}$  corresponds to a one-step SPI update from  $\pi^{(t)}$ , namely,

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}, \quad \bar{\pi}^{(t+1)}(a|s) = \frac{1}{\bar{Z}^{(t)}(s)} \exp \left( Q_\tau^{(t)}(s, a) / \tau \right), \tag{93b}$$

where we denote

$$\bar{Z}^{(t)}(s) = \sum_{a \in \mathcal{A}} \exp(Q_\tau^{(t)}(s, a) / \tau) \quad \text{and} \quad Q_\tau^{(t)} = Q_\tau^{\pi^{(t)}}$$

as usual. Here,  $\beta \in (0, 1]$  is a parameter that controls the ‘‘conservatism’’ of the updates. We characterize the convergence rate of this update rule (93) in the following theorem.

**Theorem 4** (Linear convergence of CPI-style updates). *For any  $0 < \beta \leq 1$ , the update rule (93) satisfies*

$$V_\tau^*(\rho) - V_\tau^{(t)}(\rho) \leq \left\| \frac{\rho}{\mu_\tau^*} \right\|_\infty (1 - \beta(1 - \gamma))^t \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(0)}(\mu_\tau^*) \right), \quad \forall t \geq 0, \tag{94}$$

where  $\mu_\tau^*$  is the stationary distribution defined in (29).

According to Theorem 4, it takes the CPI-style policy update (93) at most

$$\frac{1}{\beta(1-\gamma)} \log \left( \left\| \frac{\rho}{\mu_\tau^*} \right\|_\infty \frac{V_\tau^*(\mu_\tau^*) - V_\tau^{(0)}(\mu_\tau^*)}{\epsilon} \right)$$

iterations to reach  $V_\tau^*(\rho) - V_\tau^{(t)}(\rho) \leq \epsilon$ . As it turns out, the CPI-style update rule can be analyzed using our framework through the following performance improvement lemma, which is an adaptation of Lemma 1. In what follows, we use  $\bar{Q}_\tau^{(t+1)}$  and  $\bar{V}_\tau^{(t+1)}$  to abbreviate  $Q_\tau^{\bar{\pi}^{(t+1)}}$  and  $V_\tau^{\bar{\pi}^{(t+1)}}$ , respectively.

**Lemma 7** (Performance improvement of CPI-style updates). *Consider the policy update rule (93a) with any  $\beta \in (0, 1]$ . For any distribution  $\rho$ , one has*

$$V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) \geq \frac{\beta\tau}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{(t+1)}} \left[ \text{KL} \left( \pi^{(t)}(\cdot|s) \parallel \bar{\pi}^{(t+1)}(\cdot|s) \right) \right].$$

*Proof.* See Appendix D.1. □

Combining the above result with Lemma 5 and following a similar approach to (60) give

$$\begin{aligned} V_\tau^*(\rho) - V_\tau^{(t+1)}(\rho) &= V_\tau^*(\rho) - V_\tau^{(t)}(\rho) + \left( V_\tau^{(t)}(\rho) - V_\tau^{(t+1)}(\rho) \right) \\ &\stackrel{(i)}{\leq} V_\tau^*(\rho) - V_\tau^{(t)}(\rho) - \frac{\beta\tau}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{(t+1)}} \left[ \text{KL} \left( \pi^{(t)}(\cdot|s) \parallel \bar{\pi}^{(t+1)}(\cdot|s) \right) \right] \\ &\stackrel{(ii)}{\leq} V_\tau^*(\rho) - V_\tau^{(t)}(\rho) - \frac{\beta\tau}{1-\gamma} \left\| \frac{d_\rho^{\bar{\pi}^*}}{d_\rho^{(t+1)}} \right\|_\infty^{-1} \mathbb{E}_{s \sim d_\rho^{\bar{\pi}^*}} \left[ \text{KL} \left( \pi^{(t)}(\cdot|s) \parallel \bar{\pi}^{(t+1)}(\cdot|s) \right) \right] \\ &\stackrel{(iii)}{\leq} V_\tau^*(\rho) - V_\tau^{(t)}(\rho) - \beta \left\| \frac{d_\rho^{\bar{\pi}^*}}{d_\rho^{(t+1)}} \right\|_\infty^{-1} \left( V_\tau^*(\rho) - V_\tau^{(t)}(\rho) \right) \\ &= \left( 1 - \beta \left\| \frac{d_\rho^{\bar{\pi}^*}}{d_\rho^{(t+1)}} \right\|_\infty^{-1} \right) \left( V_\tau^*(\rho) - V_\tau^{(t)}(\rho) \right). \end{aligned} \tag{95}$$

Here, (i) arises from Lemma 7, (ii) employs the pre-factor  $\left\| d_\rho^{\bar{\pi}^*} / d_\rho^{(t+1)} \right\|_\infty^{-1}$  to accommodate the change of distributions, whereas (iii) follows from Lemma 5 and the constraint that  $0 \leq \eta \leq \frac{1-\gamma}{\tau}$ . By taking  $\rho$  to be the stationary distribution  $\mu_\tau^*$  (cf. (29)), one has

$$\begin{aligned} V_\tau^*(\mu_\tau^*) - V_\tau^{(t+1)}(\mu_\tau^*) &\leq \left( 1 - \beta \left\| \frac{d_{\mu_\tau^*}^{\bar{\pi}^*}}{d_{\mu_\tau^*}^{(t+1)}} \right\|_\infty^{-1} \right) \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t)}(\mu_\tau^*) \right) \\ &\leq \left( 1 - \beta \left\| \frac{\mu_\tau^*}{(1-\gamma)\mu_\tau^*} \right\|_\infty^{-1} \right) \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t)}(\mu_\tau^*) \right) \\ &= (1 - \beta(1-\gamma)) \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t)}(\mu_\tau^*) \right), \end{aligned}$$

where we have used  $d_{\mu_\tau^*}^{\bar{\pi}^*} = \mu_\tau^*$  (cf. (29)) and  $d_{\mu_\tau^*}^{(t+1)} \geq (1-\gamma)\mu_\tau^*$  in the second step. This immediately concludes the proof.

## D.1 Proof of Lemma 7

First of all, we claim that

$$V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) = \frac{\tau}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{(t+1)}} \left[ \text{KL} \left( \pi^{(t)}(\cdot|s) \parallel \bar{\pi}^{(t+1)}(\cdot|s) \right) - \text{KL} \left( \pi^{(t+1)}(\cdot|s) \parallel \bar{\pi}^{(t+1)}(\cdot|s) \right) \right], \tag{96}$$

which we shall establish momentarily. Since the KL divergence  $\text{KL}(\pi(\cdot|s) \|\bar{\pi}^{(t+1)}(\cdot|s))$  is convex in  $\pi(\cdot|s)$  (Cover, 1999), the update rule (93a) together with Jensen's inequality necessarily implies that

$$\begin{aligned} \text{KL}(\pi^{(t+1)}(\cdot|s) \|\bar{\pi}^{(t+1)}(\cdot|s)) &\leq \beta \text{KL}(\bar{\pi}^{(t+1)}(\cdot|s) \|\bar{\pi}^{(t+1)}(\cdot|s)) + (1-\beta) \text{KL}(\pi^{(t)}(\cdot|s) \|\bar{\pi}^{(t+1)}(\cdot|s)) \\ &= (1-\beta) \text{KL}(\pi^{(t)}(\cdot|s) \|\bar{\pi}^{(t+1)}(\cdot|s)). \end{aligned}$$

Substituting the above inequality into (96) allows us to conclude that

$$V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) \geq \frac{\beta\tau}{1-\gamma} \mathbb{E}_{s \sim d_p^{(t+1)}} \left[ \text{KL}(\pi^{(t)}(\cdot|s) \|\bar{\pi}^{(t+1)}(\cdot|s)) \right].$$

The rest of this proof is then dedicated to establishing the claim (96), which is similar to the proof of Lemma 1. To begin with, we express  $V_\tau^{(t)}(s_0)$  as follows

$$\begin{aligned} V_\tau^{(t)}(s_0) &= \mathbb{E}_{a_0 \sim \pi^{(t)}(\cdot|s_0)} \left[ -\tau \log \pi^{(t)}(a_0|s_0) + Q_\tau^{(t)}(s_0, a_0) \right] \\ &= \mathbb{E}_{a_0 \sim \pi^{(t)}(\cdot|s_0)} \left[ -\tau \log \pi^{(t)}(a_0|s_0) + \tau \log \bar{\pi}^{(t+1)}(a_0|s_0) \right] + \tau \log \bar{Z}^{(t)}(s_0) \\ &= \tau \log \bar{Z}^{(t)}(s_0) - \tau \text{KL}(\pi^{(t)}(\cdot|s_0) \|\bar{\pi}^{(t+1)}(\cdot|s_0)) \\ &= \tau \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ \log \bar{Z}^{(t)}(s_0) \right] - \tau \text{KL}(\pi^{(t)}(\cdot|s_0) \|\bar{\pi}^{(t+1)}(\cdot|s_0)), \end{aligned}$$

where the first line makes use of the definitions (8) and (11a), the second line follows from (93), the third line uses the definition of the KL divergence, and the last line follows since  $\bar{Z}^{(t)}(s_0)$  does not depend on  $a$ . To continue, we subtract and add  $\tau \text{KL}(\pi^{(t+1)}(\cdot|s_0) \|\bar{\pi}^{(t+1)}(\cdot|s_0))$  to obtain

$$\begin{aligned} V_\tau^{(t)}(s_0) &= \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ \tau \log \bar{Z}^{(t)}(s_0) - \tau \log \pi^{(t+1)}(a_0|s_0) + \tau \log \bar{\pi}^{(t+1)}(a_0|s_0) \right] \\ &\quad + \tau \text{KL}(\pi^{(t+1)}(\cdot|s_0) \|\bar{\pi}^{(t+1)}(\cdot|s_0)) - \tau \text{KL}(\pi^{(t)}(\cdot|s_0) \|\bar{\pi}^{(t+1)}(\cdot|s_0)) \\ &= \mathbb{E}_{a_0 \sim \pi^{(t+1)}(\cdot|s_0)} \left[ -\tau \log \pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0, a_0) \right] + \tau \text{KL}(\pi^{(t+1)}(\cdot|s_0) \|\bar{\pi}^{(t+1)}(\cdot|s_0)) \\ &\quad - \tau \text{KL}(\pi^{(t)}(\cdot|s_0) \|\bar{\pi}^{(t+1)}(\cdot|s_0)) \\ &= V_\tau^{(t+1)}(s_0) + \frac{\tau}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{(t+1)}} \left[ \text{KL}(\pi^{(t+1)}(\cdot|s) \|\bar{\pi}^{(t+1)}(\cdot|s)) - \text{KL}(\pi^{(t)}(\cdot|s) \|\bar{\pi}^{(t+1)}(\cdot|s)) \right]. \end{aligned}$$

Here, the first step relies on the definition of KL divergence, the second step comes from (93), while the last step is obtained by using the relation  $Q_\tau^{(t)}(s_0, a_0) = r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot|s_0, a_0)} [V_\tau^{(t)}(s_1)]$  and then invoking the above equality recursively as in the expression (72) (see Lemma 1). Averaging the equality over the initial state distribution  $s_0 \sim \rho$  thus establishes the claim (96).

## E Proof for approximate entropy-regularized NPG (Theorem 2)

In this section, we complete the proofs of Theorem 2 in Section 4.3, which consists of (i) establishing the linear system in (58) and (ii) extracting the convergence rate from (58).

**Step 1: establishing the linear system (58).** In what follows, we shall justify the linear system relation by checking each row separately.

(1) **Bounding  $\|Q_\tau^* - \tau \log \hat{\xi}^{(t+1)}\|_\infty$ .** From the construction (57b) of  $\hat{\xi}^{(t+1)}$ , we have

$$Q_\tau^* - \tau \log \hat{\xi}^{(t+1)} = \alpha(Q_\tau^* - \tau \log \hat{\xi}^{(t)}) + (1-\alpha)(Q_\tau^* - Q_\tau^{(t)}) + (1-\alpha)(Q_\tau^{(t)} - \hat{Q}_\tau^{(t)}).$$

Taken together with the triangle inequality and the assumption  $\|Q_\tau^{(t)} - \widehat{Q}_\tau^{(t)}\|_\infty \leq \delta$ , this gives

$$\|Q_\tau^* - \tau \log \widehat{\xi}^{(t+1)}\|_\infty \leq \alpha \|Q_\tau^* - \tau \log \widehat{\xi}^{(t)}\|_\infty + (1 - \alpha) \|Q_\tau^* - Q_\tau^{(t)}\|_\infty + (1 - \alpha) \delta. \quad (97)$$

(2) **Bounding**  $-\min_{s,a} (Q_\tau^{(t+1)}(s, a) - \tau \log \widehat{\xi}^{(t+1)}(s, a))$ . Invoking the definition (57b) of  $\widehat{\xi}^{(t+1)}$  again implies that for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ ,

$$\begin{aligned} & - \left( Q_\tau^{(t+1)}(s, a) - \tau \log \widehat{\xi}^{(t+1)}(s, a) \right) \\ &= - \left( Q_\tau^{(t+1)}(s, a) - \tau \left( \alpha \log \widehat{\xi}^{(t)}(s, a) + (1 - \alpha) \widehat{Q}_\tau^{(t)}(s, a) / \tau \right) \right) \\ &= -\alpha \left( Q_\tau^{(t)}(s, a) - \tau \log \widehat{\xi}^{(t)}(s, a) \right) + (1 - \alpha) \left( \widehat{Q}_\tau^{(t)}(s, a) - Q_\tau^{(t)}(s, a) \right) + \left( Q_\tau^{(t)}(s, a) - Q_\tau^{(t+1)}(s, a) \right) \\ &\leq -\alpha \left( Q_\tau^{(t)}(s, a) - \tau \log \widehat{\xi}^{(t)}(s, a) \right) + (1 - \alpha) \delta + \frac{2\gamma\delta}{1 - \gamma}, \end{aligned}$$

where the last inequality follows from  $\|Q_\tau^{(t)} - \widehat{Q}_\tau^{(t)}\|_\infty \leq \delta$  and (56). Taking the maximum over  $(s, a) \in \mathcal{S} \times \mathcal{A}$  on both sides and using the definition  $\alpha = 1 - \frac{\eta\tau}{1 - \gamma}$  yield

$$-\min_{s,a} \left( Q_\tau^{(t+1)}(s, a) - \tau \log \widehat{\xi}^{(t+1)}(s, a) \right) \leq -\alpha \min_{s,a} \left( Q_\tau^{(t)}(s, a) - \tau \log \widehat{\xi}^{(t)}(s, a) \right) + (1 - \alpha) \delta \left( 1 + \frac{2\gamma}{\eta\tau} \right). \quad (98)$$

(3) **Bounding**  $\|Q_\tau^* - Q_\tau^{(t+1)}\|_\infty$ . Following the same arguments as for (78), we obtain

$$\begin{aligned} Q_\tau^*(s, a) - Q_\tau^{(t+1)}(s, a) &= \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[ \tau \log (\|\exp(Q_\tau^*(s', \cdot) / \tau)\|_1) - \tau \log (\|\widehat{\xi}^{(t+1)}(s', \cdot)\|_1) \right] \\ &\quad - \gamma \mathbb{E}_{\substack{s' \sim P(\cdot | s, a), \\ a' \sim \pi^{(t+1)}(\cdot | s')}} \left[ Q_\tau^{(t+1)}(s', a') - \tau \log \widehat{\xi}^{(t+1)}(s', a') \right] \\ &\leq \gamma \|Q_\tau^* - \tau \log \widehat{\xi}^{(t+1)}\|_\infty - \gamma \min_{s,a} \left( Q_\tau^{(t+1)}(s, a) - \tau \log \widehat{\xi}^{(t+1)}(s, a) \right), \end{aligned}$$

where the last line follows from (65). By plugging (97) and (98) into the above inequality, we arrive at the claimed bound regarding this term.

**Step 2: deducing convergence guarantees from the linear system (58).** We start by pinning down the eigenvalues and eigenvectors of the matrix  $B$ . Specifically, the three eigenvalues can be calculated as

$$\lambda_1 = \alpha + \gamma(1 - \alpha) = 1 - \eta\tau, \quad \lambda_2 = \alpha \quad \text{and} \quad \lambda_3 = 0, \quad (99)$$

whose corresponding eigenvectors are given respectively by

$$v_1 = \begin{bmatrix} \gamma \\ 1 \\ 0 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}, \quad \text{and} \quad v_3 = \begin{bmatrix} \alpha \\ \alpha - 1 \\ 0 \end{bmatrix}. \quad (100)$$

With some elementary computation, one can show that  $z_0$  and  $b$  introduced in (59) can be related to the eigenvectors of  $B$  in the following way:

$$z_0 \leq \begin{bmatrix} \|Q_\tau^* - Q_\tau^{(0)}\|_\infty \\ \|Q_\tau^* - \tau \log \widehat{\xi}^{(0)}\|_\infty \\ \|Q_\tau^{(0)} - \tau \log \widehat{\xi}^{(0)}\|_\infty \end{bmatrix}$$

$$\begin{aligned}
&= \frac{1}{1-\eta\tau} \left[ (1-\alpha) \|Q_\tau^* - Q_\tau^{(0)}\|_\infty + \alpha \left( \|Q_\tau^* - \tau \log \widehat{\xi}^{(0)}\|_\infty + \|Q_\tau^{(0)} - \tau \log \widehat{\xi}^{(0)}\|_\infty \right) \right] v_1 \\
&\quad + \|Q_\tau^{(0)} - \tau \log \widehat{\xi}^{(0)}\|_\infty v_2 + c_z v_3 \\
&\leq \frac{1}{1-\eta\tau} \left( \|Q_\tau^* - Q_\tau^{(0)}\|_\infty + 2\alpha\tau \|\log \pi_\tau^* - \log \pi^{(0)}\|_\infty \right) v_1 + \|Q_\tau^{(0)} - \tau \log \widehat{\xi}^{(0)}\|_\infty v_2 + c_z v_3, \tag{101}
\end{aligned}$$

where  $c_z$  is some scalar whose value is immaterial since the eigenvalue corresponding to  $v_3$  is  $\lambda_3 = 0$ , and the last line follows from the same reasoning for (52). Another useful identity is:

$$b = (1-\alpha)\delta \begin{bmatrix} \gamma \left(2 + \frac{2\gamma}{\eta\tau}\right) \\ 1 \\ 1 + \frac{2\gamma}{\eta\tau} \end{bmatrix} = (1-\alpha)\delta \left[ \left(2 + \frac{2\gamma}{\eta\tau}\right) v_1 + \left(1 + \frac{2\gamma}{\eta\tau}\right) v_2 \right]. \tag{102}$$

With these preparations in place, we can now invoke the recursion relationship (58) and the non-negativity of  $B$  to obtain

$$\begin{aligned}
z_{t+1} &\leq B^{t+1} z_0 + \sum_{s=0}^t B^{t-s} b \\
&\leq B^{t+1} \left[ \frac{1}{1-\eta\tau} \left( \|Q_\tau^* - Q_\tau^{(0)}\|_\infty + 2\alpha\tau \|\log \pi_\tau^* - \log \pi^{(0)}\|_\infty \right) v_1 + \|Q_\tau^{(0)} - \tau \log \widehat{\xi}^{(0)}\|_\infty v_2 + c_z v_3 \right] \\
&\quad + (1-\alpha)\delta \sum_{s=0}^t B^{t-s} \left[ \left(2 + \frac{2\gamma}{\eta\tau}\right) v_1 + \left(1 + \frac{2\gamma}{\eta\tau}\right) v_2 \right] \\
&= \left[ \lambda_1^t \left( \|Q_\tau^* - Q_\tau^{(0)}\|_\infty + 2\alpha\tau \|\log \pi_\tau^* - \log \pi^{(0)}\|_\infty \right) + (1-\alpha)\delta \left(2 + \frac{2\gamma}{\eta\tau}\right) \frac{1-\lambda_1^{t+1}}{1-\lambda_1} \right] v_1 \\
&\quad + \left[ \lambda_2^{t+1} \|Q_\tau^{(0)} - \tau \log \widehat{\xi}^{(0)}\|_\infty + (1-\alpha)\delta \left(1 + \frac{2\gamma}{\eta\tau}\right) \frac{1-\lambda_2^{t+1}}{1-\lambda_2} \right] v_2,
\end{aligned}$$

where the eigenvalues and eigenvectors of  $B$  are given in (99) and (100), respectively, and the second inequality relies on (101) and (102). Note that we are only interested in the first two entries of the vector  $z_t$ . Since the first two entries of the eigenvector  $v_2$  are non-positive, we can safely drop the term involving  $v_2$  in the above inequality to obtain

$$\begin{aligned}
&\begin{bmatrix} \|Q_\tau^* - Q_\tau^{(t+1)}\|_\infty \\ \|Q_\tau^* - \tau \log \widehat{\xi}^{(t+1)}\|_\infty \end{bmatrix} \\
&\leq \left\{ \lambda_1^t \left( \|Q_\tau^* - Q_\tau^{(0)}\|_\infty + 2\alpha\tau \|\log \pi_\tau^* - \log \pi^{(0)}\|_\infty \right) + (1-\alpha)\delta \left(2 + \frac{2\gamma}{\eta\tau}\right) \frac{1-\lambda_1^{t+1}}{1-\lambda_1} \right\} \begin{bmatrix} \gamma \\ 1 \end{bmatrix} \\
&\leq \left\{ (1-\eta\tau)^t \left( \|Q_\tau^* - Q_\tau^{(0)}\|_\infty + 2 \left(1 - \frac{\eta\tau}{1-\gamma}\right) \tau \|\log \pi_\tau^* - \log \pi^{(0)}\|_\infty \right) + \frac{2\delta}{1-\gamma} \left(1 + \frac{\gamma}{\eta\tau}\right) \right\} \begin{bmatrix} \gamma \\ 1 \end{bmatrix}. \tag{103}
\end{aligned}$$

When it comes to the log policies, we recall again the fact that  $\pi^{(t)}$  is related to  $\widehat{\xi}^{(t)}$  as

$$\forall s \in \mathcal{S}: \quad \pi^{(t)}(\cdot|s) = \frac{1}{\|\widehat{\xi}^{(t)}(s, \cdot)\|_1} \widehat{\xi}^{(t)}(s, \cdot). \tag{104}$$

Invoking the elementary property (66), we reach

$$\|\log \pi_\tau^* - \log \pi^{(t+1)}\|_\infty \leq 2 \|Q_\tau^*/\tau - \log \widehat{\xi}^{(t+1)}\|_\infty.$$

This together with the bound on  $\|Q_\tau^* - \tau \log \widehat{\xi}^{(t+1)}\|_\infty$  in (103) establishes our claim for  $\|\log \pi_\tau^* - \log \pi^{(t+1)}\|_\infty$ .

## F Proof for local quadratic convergence (Theorem 3)

Assuming that the policy  $\pi^{(t)}$  obeys Condition (30), we can control the difference of the corresponding discounted state visitation probabilities in terms of the sub-optimality gap w.r.t. the log policy. This is stated in the following lemma, whose proof is deferred to Section F.1.

**Lemma 8.** *Consider any policy  $\pi$  satisfying  $\|\log \pi - \log \pi_\tau^*\|_\infty \leq 1$ . It follows that*

$$\left\| 1 - \frac{d_\rho^{\pi_\tau^*}}{d_\rho^\pi} \right\|_\infty \leq 2 \left( \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi_\tau^*}}{\rho} \right\|_\infty - 1 \right) \|\log \pi - \log \pi_\tau^*\|_\infty.$$

In particular, by taking  $\rho = \mu_\tau^*$  one has

$$\left\| 1 - \frac{\mu_\tau^*}{d_{\mu_\tau^*}^{\pi_\tau^*}} \right\|_\infty = \left\| 1 - \frac{d_{\mu_\tau^*}^{\pi_\tau^*}}{d_{\mu_\tau^*}^{\mu_\tau^*}} \right\|_\infty \leq \frac{2\gamma}{1-\gamma} \|\log \pi - \log \pi_\tau^*\|_\infty.$$

First, by virtue of the SPI update rule (17) and the inequality (66), it is guaranteed that

$$\|\log \pi_\tau^* - \log \pi^{(t+1)}\|_\infty \leq \frac{2}{\tau} \|Q_\tau^* - Q_\tau^{(t)}\|_\infty \leq \frac{2\gamma}{\tau} \|V_\tau^* - V_\tau^{(t)}\|_\infty \leq \frac{2\gamma}{\tau} \left\| \frac{1}{\mu_\tau^*} \right\|_\infty \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t)}(\mu_\tau^*) \right), \quad (105)$$

where the last inequality comes from a change of distributions argument. Armed with Lemma 8 and the inequality (105), we arrive at

$$\left\| 1 - \frac{d_{\mu_\tau^*}^{\pi_\tau^*}}{d_{\mu_\tau^*}^{(t+1)}} \right\|_\infty \leq \frac{2\gamma}{1-\gamma} \|\log \pi^{(t+1)} - \log \pi_\tau^*\|_\infty \leq \frac{4\gamma^2}{(1-\gamma)\tau} \left\| \frac{1}{\mu_\tau^*} \right\|_\infty \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t)}(\mu_\tau^*) \right). \quad (106)$$

Substitution into (60) gives

$$\begin{aligned} V_\tau^*(\mu_\tau^*) - V_\tau^{(t+1)}(\mu_\tau^*) &\leq \left( 1 - \left\| \frac{d_{\mu_\tau^*}^{\pi_\tau^*}}{d_{\mu_\tau^*}^{(t+1)}} \right\|_\infty^{-1} \right) \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t)}(\mu_\tau^*) \right) \\ &= \left( 1 - \left\| 1 + \frac{d_{\mu_\tau^*}^{\pi_\tau^*} - d_{\mu_\tau^*}^{(t+1)}}{d_{\mu_\tau^*}^{(t+1)}} \right\|_\infty^{-1} \right) \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t)}(\mu_\tau^*) \right) \\ &\leq \left( 1 - \frac{1}{1 + \frac{4\gamma^2}{(1-\gamma)\tau} \left\| \frac{1}{\mu_\tau^*} \right\|_\infty \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t)}(\mu_\tau^*) \right)} \right) \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t)}(\mu_\tau^*) \right) \\ &= \frac{\frac{4\gamma^2}{(1-\gamma)\tau} \left\| \frac{1}{\mu_\tau^*} \right\|_\infty \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t)}(\mu_\tau^*) \right)^2}{1 + \frac{4\gamma^2}{(1-\gamma)\tau} \left\| \frac{1}{\mu_\tau^*} \right\|_\infty \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t)}(\mu_\tau^*) \right)} \leq \frac{4\gamma^2}{(1-\gamma)\tau} \left\| \frac{1}{\mu_\tau^*} \right\|_\infty \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t)}(\mu_\tau^*) \right)^2, \end{aligned}$$

where the second inequality makes use of the bound (106). This in turn reveals that

$$\frac{4\gamma^2}{(1-\gamma)\tau} \left\| \frac{1}{\mu_\tau^*} \right\|_\infty \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t+1)}(\mu_\tau^*) \right) \leq \left( \frac{4\gamma^2}{(1-\gamma)\tau} \left\| \frac{1}{\mu_\tau^*} \right\|_\infty \left( V_\tau^*(\mu_\tau^*) - V_\tau^{(t)}(\mu_\tau^*) \right) \right)^2,$$

which leads to our claimed result by a standard change of distributions.

### F.1 Proof of Lemma 8

For any policy  $\pi$ , denote by  $P_\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$  the state transition matrix induced by  $\pi$  as follows

$$\forall s, s' \in \mathcal{S} : \quad [P_\pi]_{s,s'} := \mathbb{E}_{a \sim \pi(\cdot|s)} [P(s'|s, a)]. \quad (107)$$

For any policy  $\pi$  satisfying  $\|\log \pi - \log \pi_\tau^*\|_\infty \leq 1$ , we develop an upper bound on  $\left| [P_\pi - P_{\pi_\tau^*}]_{s,s'} \right|$  as follows

$$\begin{aligned} \left| [P_\pi - P_{\pi_\tau^*}]_{s,s'} \right| &= \left| \sum_a P(s'|s, a) (\pi(a|s) - \pi_\tau^*(a|s)) \right| \leq \sum_a P(s'|s, a) \pi_\tau^*(a|s) \left| \frac{\pi(a|s)}{\pi_\tau^*(a|s)} - 1 \right| \\ &\stackrel{(i)}{\leq} (e-1) \sum_a P(s'|s, a) \pi_\tau^*(a|s) \left| \log \pi(a|s) - \log \pi_\tau^*(a|s) \right| \\ &\leq \|\log \pi - \log \pi_\tau^*\|_\infty (e-1) \sum_a P(s'|s, a) \pi_\tau^*(a|s) \\ &\leq 2 [P_{\pi_\tau^*}]_{s,s'} \|\log \pi - \log \pi_\tau^*\|_\infty, \end{aligned}$$

where (i) uses the assumption  $\|\log \pi^* - \log \pi\|_\infty \leq 1$  together with the elementary inequality  $|x| \leq (e-1) |\log(1+x)|$  when  $-1 < x \leq e-1$ . With the preceding bound in mind, we can demonstrate that

$$\left| (d_\rho^{\pi_\tau^*})^\top (P_\pi - P_{\pi_\tau^*}) \right| \leq 2 \|\log \pi - \log \pi_\tau^*\|_\infty (d_\rho^{\pi_\tau^*})^\top P_{\pi_\tau^*}. \quad (108)$$

Here and throughout, we overload the notation  $|z|$  for any vector  $z \in \mathbb{R}^{|\mathcal{S}|}$  to denote  $[\|z_i\|_{1 \leq i \leq |\mathcal{S}|}]$ .

In addition, the definitions of  $d_\rho^\pi$  and  $d_\rho^{\pi_\tau^*}$  admit the following matrix-vector representation:

$$(d_\rho^\pi)^\top = (1-\gamma)\rho^\top (I - \gamma P_\pi)^{-1}, \quad (109)$$

$$(d_\rho^{\pi_\tau^*})^\top = (1-\gamma)\rho^\top (I - \gamma P_{\pi_\tau^*})^{-1}, \quad (110)$$

thus allowing one to derive

$$\begin{aligned} (d_\rho^\pi - d_\rho^{\pi_\tau^*})^\top &= (1-\gamma)\rho^\top (I - \gamma P_{\pi_\tau^*})^{-1} [(I - \gamma P_{\pi_\tau^*}) - (I - \gamma P_\pi)] (I - \gamma P_\pi)^{-1} \\ &= \gamma (d_\rho^{\pi_\tau^*})^\top (P_\pi - P_{\pi_\tau^*}) (I - \gamma P_\pi)^{-1}. \end{aligned}$$

This together with the non-negativity of the matrix  $(I - \gamma P_\pi)^{-1}$  (Li et al., 2020b, Lemma 7) enables the following bound

$$\begin{aligned} \left| (d_\rho^\pi - d_\rho^{\pi_\tau^*})^\top \right| &\leq \gamma \left| (d_\rho^{\pi_\tau^*})^\top (P_\pi - P_{\pi_\tau^*}) \right| (I - \gamma P_\pi)^{-1} \\ &\leq 2 \|\log \pi - \log \pi_\tau^*\|_\infty \gamma (d_\rho^{\pi_\tau^*})^\top P_{\pi_\tau^*} (I - \gamma P_\pi)^{-1}, \end{aligned} \quad (111)$$

where the last inequality results from (108).

Furthermore, we make the observation that

$$\begin{aligned} \gamma (d_\rho^{\pi_\tau^*})^\top P_{\pi_\tau^*} &= (1-\gamma)\gamma\rho^\top (I - \gamma P_{\pi_\tau^*})^{-1} P_{\pi_\tau^*} = (1-\gamma)\gamma\rho^\top \left[ \sum_{i=0}^{\infty} (\gamma P_{\pi_\tau^*})^i \right] P_{\pi_\tau^*} \\ &= (1-\gamma)\rho^\top \left[ \sum_{i=1}^{\infty} (\gamma P_{\pi_\tau^*})^i \right] \\ &= (1-\gamma)\rho^\top [(I - \gamma P_{\pi_\tau^*})^{-1} - I] = (d_\rho^{\pi_\tau^*})^\top - (1-\gamma)\rho^\top \leq \left( \left\| \frac{d_\rho^{\pi_\tau^*}}{\rho} \right\|_\infty - (1-\gamma) \right) \rho^\top, \end{aligned}$$

where the last line comes from a change of distributions argument. Combining this bound with (111) gives

$$\begin{aligned} \left| (d_\rho^\pi - d_\rho^{\pi_\tau^*})^\top \right| &\leq 2 \|\log \pi - \log \pi_\tau^*\|_\infty \left( \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi_\tau^*}}{\rho} \right\|_\infty - 1 \right) (1-\gamma)\rho^\top (I - \gamma P_\pi)^{-1} \\ &= 2 \|\log \pi - \log \pi_\tau^*\|_\infty \left( \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi_\tau^*}}{\rho} \right\|_\infty - 1 \right) (d_\rho^\pi)^\top, \end{aligned}$$

where the last line arises from the expression (109). As a result, we establish the claimed bound

$$\left\| 1 - \frac{d_{\rho}^{\pi^*}}{d_{\rho}^{\pi}} \right\|_{\infty} \leq 2 \left\| \log \pi - \log \pi^* \right\|_{\infty} \left( \frac{1}{1 - \gamma} \left\| \frac{d_{\rho}^{\pi^*}}{\rho} \right\|_{\infty} - 1 \right).$$