Linear Convergence of Entropy-Regularized Natural Policy Gradient with Linear Function Approximation

Semih Cayci [1] **Niao He** [2] **R. Srikant** [3 1 4]

Abstract

Natural policy gradient (NPG) methods with function approximation achieve impressive empirical success in reinforcement learning problems with large state-action spaces. However, theoretical understanding of their convergence behaviors remains limited in the function approximation setting. In this paper, we perform a finite-time analysis of NPG with linear function approximation and softmax parameterization, and prove for the first time that widely used entropy regularization method, which encourages exploration, leads to linear convergence rate. We adopt a Lyapunov drift analysis to prove the convergence results and explain the effectiveness of entropy regularization in improving the convergence rates.

1. Introduction

The goal of reinforcement learning (RL) is to sequentially maximize the expected total reward in a Markov decision process (MDP) (Sutton & Barto, 2018; Szepesvári, 2010; Bertsekas & Tsitsiklis, 1996). Policy gradient (PG) methods, which aim to find the optimal policy in the parameter space by using gradient ascent (Williams, 1992; Sutton et al., 1999; Konda & Tsitsiklis, 2000), have demonstrated significant empirical success in a broad class of challenging RL problems (Mnih et al., 2016; Silver et al., 2016; Nachum et al., 2017; Duan et al., 2016).

Among the variants of policy gradient methods, natural policy gradient (NPG), which uses Fisher information matrix for pre-conditioning the gradient steps as a quasi-Newton method (Amari, 1998; Kakade, 2001; Peters & Schaal, 2008; Bhatnagar et al., 2007), has been particularly popular as a consequence of their impressive empirical performance and flexibility with function approximation (Schulman et al., 2015; Shani et al., 2020; Even-Dar et al., 2009). In practical applications, policy gradient methods are often combined with an entropy regularizer to encourage exploration, which yields considerably improved empirical performance (Haarnoja et al., 2018; Nachum et al., 2017; Ahmed et al., 2019).

Despite the impressive empirical success of the policy gradient methods in practical applications, particularly when used in conjunction with function approximation and entropy regularization, their theoretical convergence properties remain elusive. In the tabular setting, recent work show intriguing linear convergence of NPG with entropy-regularization under softmax parameterization (Mei et al., 2020; Cen et al., 2020; Lan, 2021). This motivates us to address the following open question in this paper: *Do entropy-regularized NPG methods with linear function approximation converge linearly to the optimal policy?* We will answer the question affirmatively and shed light on the role of the practically-used entropy regularization in the convergence of NPG methods with function approximation assuming true gradient.

1.1. Main Contributions

In this paper, we establish sharp non-asymptotic bounds on the global convergence of entropy-regularized natural policy gradient under softmax parameterization with linear function approximation. To the best of our knowledge, this is the first work that shows linear convergence of NPG in the function approximation regime. Our main contributions include the following:

- *Linear convergence of entropy-regularized NPG:* We prove for the first time that entropy-regularized NPG under softmax parameterization and linear function approximation achieves $\exp(-\Omega(T))$ convergence rate, up to a function approximation error, under a mild regularity condition on the basis vectors. We show that entropy regularization encourages exploration so that all actions are explored with some probability bounded away from zero (Lemma 2), which leads to improved convergence rates. This is an important step in explain-

[1]Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA [2]Department of Computer Science, ETH Zurich, Zurich, Switzerland [3]c3.ai DTI [4]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. Correspondence to: Semih Cayci <scayci@illinois.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

ing the empirical success of entropy regularization in RL applications.

- *Lyapunov drift analysis:* Our results rely on an intuitive Lyapunov drift analysis, which enables proving sharp convergence bounds and also studying the impact of regularization.

1.2. Related Work

NPG with function approximation: In (Agarwal et al., 2020), unregularized NPG with softmax parameterization and linear function approximation was studied, and $O(1/\sqrt{T})$ convergence rate is proved. Our analysis is inspired by (Agarwal et al., 2020) to analyze entropy-regularized NPG. We prove that $e^{-\Omega(T)}$ rate of convergence is achieved under a mild regularity condition on the basis vectors. In (Wang et al., 2019), $O(1/\sqrt{T})$ rate of convergence for NPG with neural network approximation is proved in the so-called neural tangent kernel (NTK) regime.

NPG in the tabular setting: The convergence properties of NPG in the tabular setting is relatively better understood compared to the function approximation setting (Agarwal et al., 2020; Bhandari & Russo, 2019; Cen et al., 2020; Mei et al., 2020). In (Cen et al., 2020), linear convergence of tabular-NPG is proved by exploiting a relation to the policy iteration in the tabular setting. In another recent work, (Mei et al., 2020) proves linear convergence of entropy-regularized tabular-NPG by establishing a Polyak-Łojasiewicz inequality. Similar results are obtained in (Lan, 2021) for more general regularizers. These results rely on the properties of the tabular setting. In this article, we adopt a different Lyapunov-drift approach, which makes use of potential functions for the analysis in the function approximation setting.

2. System Model and Algorithms

In this section, we will introduce the reinforcement learning setting and natural policy gradient algorithm.

2.1. Markov Decision Processes

In this work, we consider a γ-discounted Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$ where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces; $\mathcal{P}$ is a transition model $r : \mathcal{S} \times \mathcal{A} \to [r_{min}, r_{max}]$, $0 < r_{min} < r_{max} < \infty$ is the reward function. In this work, we consider a finite but arbitrarily large state space $\mathcal{S}$ and a finite action space $\mathcal{A}$.

A randomized *policy* $\pi : \mathcal{S} \to \mathcal{A}$ corresponds to a decision-making rule by specifying the probability of taking an action $a \in \mathcal{A}$ at a given state $s \in \mathcal{S}$. A policy π introduces a trajectory by specifying $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$. The corresponding value function of a policy π is as follows:

$$V^{\pi}(s) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s\Big], \tag{1}$$

where $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$. For an initial state distribution $\mu \in \Delta(\mathcal{S})$, we define

$$V^{\pi}(\mu) = \sum_{s \in \mathcal{S}} \mu(s) V^{\pi}(s). \tag{2}$$

Policy parameterization: In this work, we consider softmax parameterization with linear function approximation. Namely, we consider the log-linear policy class $\Pi = \{\pi_\theta : \theta \in \mathbb{R}^d\}$, where:

$$\pi_\theta(a|s) = \frac{\exp(\theta^\top \phi_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi_{s,a'})}, \tag{3}$$

for a set of d-dimensional basis vectors $\{\phi_{s,a} \in \mathbb{R}^d : s \in \mathcal{S}, a \in \mathcal{A}\}$ with $\sup_{s \in \mathcal{S}, a \in \mathcal{A}} \|\phi_{s,a}\|_2 \leq 1$, and policy parameter $\theta \in \mathbb{R}^d$. Π is a restricted policy class, which is a strict subset of all stochastic policies (Agarwal et al., 2020). We will aim to find the best policy within Π throughout this paper.

Entropy regularization: The value function $V^{\pi_\theta}(\mu)$ is a non-concave function of $\theta \in \mathbb{R}^d$ (Bhandari & Russo, 2019; Agarwal et al., 2020), and there exist suboptimal near-deterministic policies. In order to encourage exploration and evade suboptimal policies, entropy regularization is commonly used in practice (Silver et al., 2016; Haarnoja et al., 2018; Ahmed et al., 2019). For a policy $\pi \in \Pi$, let $H^{\pi}(\mu) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \mathcal{H}\big(\pi(\cdot|s_t)\big) \Big| s_0 \sim \mu\Big]$, where $\mathcal{H}(\pi(\cdot|s)) = -\sum_{a \in \mathcal{A}} \pi(a|s) \log\big(\pi(a|s)\big)$ is the entropy functional. Then, for $\lambda > 0$, the entropy-regularized value function is defined as follows:

$$V_\lambda^{\pi}(\mu) = V^{\pi}(\mu) + \lambda H^{\pi}(\mu). \tag{4}$$

Note that $\pi_{\mathbf{0}}(\cdot|\cdot) = 1/|\mathcal{A}|$ maximizes the regularizer $H^{\pi}(\mu)$. Hence, the additional $\lambda H^{\pi}(\mu)$ term in (4) encourages exploration increasingly with $\lambda > 0$.

Objective: The objective in this paper is to maximize the entropy-regularized value function in (4) for a given $\lambda > 0$ and initial state distribution $\mu \in \Delta(\mathcal{S})$:

$$\theta^* = \arg\max_{\theta \in \mathbb{R}^d} V_\lambda^{\pi_\theta}(\mu), \tag{5}$$

We denote the optimal policy as $\pi^* = \pi_{\theta^*}$ throughout the paper, and assume that $\|\theta^*\|_2 < \infty$, or equivalently $\inf_{s \in \mathcal{S}, a \in \mathcal{A}} \pi^*(a|s) > 0$, which automatically holds for sufficiently large $\lambda > 0$.

Q-function and advantage function: We define the (entropy-regularized) Q-function under a policy π as follows:

$$Q_\lambda^\pi(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) V_\lambda^\pi(s'). \quad (6)$$

The advantage function under a policy π is defined as follows:

$$A_\lambda^\pi(s,a) = Q_\lambda^\pi(s,a) - V_\lambda^\pi(s) - \lambda \log \pi(a|s). \quad (7)$$

We have the following characterization of $V_\lambda^\pi(\mu)$.

Proposition 1. *For any $\pi \in \Pi$, $\lambda > 0$ and $\mu \in \Delta(\mathcal{S})$, we have:*

$$V_\lambda^\pi(\mu) = \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mu(s)\pi(a|s)\big(Q_\lambda^\pi(s,a) - \lambda \log \pi(a|s)\big).$$

We can bound the entropy-regularized value function by using Prop. 1 as follows:

$$\frac{r_{min}}{1-\gamma} \leq V_\lambda^{\pi_\theta}(\mu) \leq \frac{r_{max} + \lambda \log |\mathcal{A}|}{1-\gamma}, \quad (8)$$

for any $\lambda > 0, \theta \in \mathbb{R}^d, \mu \in \Delta(\mathcal{S})$ since $r \in [r_{min}, r_{max}]$ and $\mathcal{H}(P) \leq \log |\mathcal{A}|$ for any $P \in \Delta(\mathcal{A})$.

2.2. Policy Gradient Theorem and Compatible Function Approximation

For any initial state distribution $\mu \in \Delta(\mathcal{S})$, let

$$d_\mu^\pi(s) = (1-\gamma) \sum_{s_0 \in \mathcal{S}} \mu(s_0) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\pi(s_t = s | s_0).$$

In the following, the gradient of the entropy-regularized value function with respect to π is characterized by extending (Sutton & Barto, 2018).

Proposition 2 (Policy gradient)**.** *For any $\theta \in \mathbb{R}^d$, $\lambda > 0$ and $\mu \in \Delta(\mathcal{S})$, we have:*

$$\nabla_\theta V_\lambda^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s,a \sim d_\mu^{\pi_\theta} \cdot \pi_\theta(\cdot|s)} \Big[\nabla_\theta \log \pi_\theta(a|s) \times \Big(Q_\lambda^{\pi_\theta}(s,a) - \lambda \log \pi_\theta(a|s) \Big) \Big],$$

where $\nabla_\theta \log \pi_\theta(a|s) = \phi_{s,a} - \sum_{a' \in \mathcal{A}} \pi_\theta(a'|s) \phi_{s,a'}$.

By using Proposition 2, the gradient update of natural policy gradient can be computed by the following lemma.

Lemma 1 (Compatible function approximation)**.** *Let*

$$L(w,\theta) = \mathbb{E}_{s \sim d_\mu^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \Big[\Big(\nabla_\theta^\top \log \pi_\theta(a|s) w - \big(Q_\lambda^{\pi_\theta}(s,a) - \lambda \log \pi_\theta(a|s) \big) \Big)^2 \Big], \quad (9)$$

be the approximation error, and

$$G_\lambda^{\pi_\theta}(\mu) = \mathbb{E}_{s \sim d_\mu^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \Big[\nabla_\theta \log \pi_\theta(a|s) \nabla_\theta^\top \log \pi_\theta(a|s) \Big],$$

be the Fisher information matrix under policy π_θ. Then, we have:

$$G_\lambda^{\pi_\theta}(\mu) w_\lambda^{\pi_\theta} = (1-\gamma) \nabla_\theta V_\lambda^{\pi_\theta}(\mu), \quad (10)$$

where

$$w_\lambda^{\pi_\theta} \in \arg\min_{w \in \mathbb{R}^d} L(w,\theta), \quad (11)$$

for any $\theta \in \mathbb{R}^d$.

By using these results, we define the natural policy gradient algorithm in the next subsection.

Algorithm 1 Entropy-regularized NPG with step-size $\eta > 0$

Input: Step-size $\eta > 0$;
Initialize: $\theta_0 = \mathbf{0}$, i.e., $\pi_0(a|s) = \frac{1}{|\mathcal{A}|}$ for all s, a
for $i = 1$ **to** $m-1$ **do**
  Compute $w_t = w_\lambda^{\pi_t}$ by using (11);
  $\theta_{t+1} = \theta_t + \eta w_t$;
end for

2.3. Natural Policy Gradient Algorithm with Entropy Regularization

For simplicity, we assume access to exact policy evaluation throughout the paper. Entropy-regularized NPG is defined in the following.

For a constant step-size $\eta > 0$, the natural policy gradient algorithm updates the parameter as follows:

$$\theta \leftarrow \theta + \eta w_\lambda^{\pi_\theta}, \quad (12)$$

where $w_\lambda^{\pi_\theta}$ is obtained by (11). The pseudocode for NPG with a constant step-size $\eta > 0$ is given in Algorithm 1. For any $t \geq 0$, we denote $\pi_t = \pi_{\theta_t}$ throughout the paper.

3. Main Results

In this section, we will prove that the entropy-regularized NPG achieves linear convergence rate. We make the following assumptions.

Assumption 1 (Concentrability coefficient)**.** *Let the concentrability coefficient be defined as*

$$C_t = \mathbb{E}_{s \sim d_\mu^{\pi_t}, a \sim \pi_t(\cdot|s)} \Big[\Big(\frac{d_\mu^{\pi^*}(s) \pi^*(a|s)}{d_\mu^{\pi_t}(s) \pi_t(a|s)} \Big)^2 \Big].$$

We assume that there exists a constant $C^\star < \infty$ such that $C_t \leq C^\star$ for all t.

Assumption 1 is a standard regularity assumption in the theoretical analysis of RL algorithms (Wang et al., 2019; Antos et al., 2008; Scherrer et al., 2015).

Assumption 2 (Approximation error)**.** *We assume that* $\min_{w\in\mathbb{R}^d} L(w,\theta_t) \le \epsilon_a$, *for all* $t \ge 1$.

This assumption is standard in the RL literature (Agarwal et al., 2020; Wang et al., 2019; Liu et al., 2019). From $\nabla_\theta \log \pi_\theta(a|s)$ for linear function approximation, we observe that ϵ_a is a measure of expressiveness of the linear function approximation with the basis vectors $\{\phi_{s,a} : (s,a) \in \mathcal{S}\times\mathcal{A}\}$.

Remark: Note that, for $\lambda > 0$ and $d < |\mathcal{S}\times\mathcal{A}|$, the approximation error by the compatible function approximation is positive, i.e., $\min_w L(w,\theta_t) > 0$, even if $Q_\lambda^{\pi_\theta}$ can be perfectly approximated by the basis vectors $\{\phi_{s,a} : (s,a) \in \mathcal{S}\times\mathcal{A}\}$. This is because there is an additional term due to entropy regularization $-\lambda \log \pi_\theta(a|s)$ in (9), which is approximated by $\nabla_\theta^\top \log \pi_\theta(a|s) w$, and d should be large for this approximation error to be small.

Assumption 3 (Regularity of the parametric model)**.** *Let*

$$F(\mu) = \mathbb{E}_{s\sim\mu, a\sim Unif(\mathcal{A})}\left[\varphi_{s,a}\varphi_{s,a}^\top\right],$$

where $\varphi_{s,a} = \phi_{s,a} - \mathbb{E}_{a'\sim Unif(\mathcal{A})}[\phi_{s,a'}]$ *We assume that* $F(\mu)$ *is non-singular.*

Assumption 3 is a regularity condition on $\{\phi_{s,a} : (s,a) \in \mathcal{S}\times\mathcal{A}\}$. It basically implies the initial Fisher information matrix, $G_\lambda^{\pi_0}(\mu)$, is non-singular since $\sigma_1(G_\lambda^{\pi_0}(\mu)) \ge (1-\gamma)\sigma_1(F(\mu))$. Similar regularity conditions, such as boundedness of the relative condition number, are assumed in the RL literature (Agarwal et al., 2020).

Remark: The minimum eigenvalue of $F(\mu)$ is bounded away from 0 for $d < |\mathcal{S}\times\mathcal{A}|$, which is called *the function approximation regime*, and typically decreases to 0 as d increases, i.e., the tabular case.

The following is a key result in the convergence proof. We denote the smallest eigenvalue of a matrix A as $\sigma_1(A)$.

Lemma 2 (Non-singularity lemma)**.** *Under Assumptions 1-3, there exist constants* $\sigma > 0$ *and* $p > 0$ *such that the following holds under Algorithm 1:*

$$\inf_{t\ge 0} \sigma_1\left(G_\lambda^{\pi_t}(\mu)\right) \ge \sigma,$$
$$\inf_{t\ge 0} \min_{\substack{a\in\mathcal{A}\\ s\in supp(\mu)}} \pi_t(a|s) \ge p,$$

the step-size is $\eta \le \min\left\{\frac{(1-\gamma)^2\sigma^2 r_{min}}{(r_{max}+\lambda\log|\mathcal{A}|)^2}, \frac{1}{2\lambda}\right\}$.

Lemma 2 implies that all actions are explored with a probability bounded away from zero because of entropy regularization.

Definition 1 (Potential function)**.** *For any* $\pi \in \Pi$, *we define the potential function* $\Phi : \Pi \to \mathbb{R}^+$ *as follows:*

$$\Phi(\pi) = \sum_{s\in\mathcal{S}} d_\mu^{\pi^*}(s) \sum_{a\in\mathcal{A}} \pi^*(a|s) \log \frac{\pi^*(a|s)}{\pi(a|s)},$$
$$= \sum_{s\in\mathcal{S}} d_\mu^{\pi^*}(s) D_{KL}(\pi^*(\cdot|s)\|\pi(\cdot|s)).$$

where D_{KL} *is the Kullback-Leibler divergence.*

Note that $\Phi(\pi) \ge 0$ for all π, and $\Phi(\pi) = 0$ if an only if $\pi(\cdot|s) = \pi^*(\cdot|s)$ for all $s \in supp(d_\mu^{\pi^*})$.

3.1. Main Convergence Result

The main result of this work is the following.

Theorem 1 (Convergence of entropy-regularized NPG)**.** *Under Assumptions 1-3, the entropy-regularized NPG with the constant step-size specified in Lemma 2 satisfies:*

$$\Phi(\pi_T) \le (1-\eta\lambda)^T \log|\mathcal{A}| + \frac{\sqrt{C^\star\cdot\epsilon_a}}{\lambda},$$
$$V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_T}(\mu) \le \left(1-\eta\lambda\right)^T \frac{\log|\mathcal{A}|}{\eta(1-\gamma)} + \frac{\sqrt{C^\star\cdot\epsilon_a}}{\lambda\eta(1-\gamma)},$$

for any $T \ge 1$, *where* $C^\star$ *is the concentrability coefficient.*

Note that the existing best-known convergence result for unregularized NPG with function approximation in (Agarwal et al., 2020) gives the bound $O\left(\frac{1}{1-\gamma}\sqrt{\frac{\log|\mathcal{A}|}{T}} + \sqrt{\frac{C^\star\epsilon_a}{(1-\gamma)^3}}\right)$. Our result shows that using entropy regularization boosts the convergence speed of NPG significantly: $T = O\left(\frac{1}{\lambda\eta}\log\left(\frac{\log|\mathcal{A}|}{\eta(1-\gamma)\epsilon}\right)\right)$ iterations are required to achieve an ϵ-optimal policy of the regularized problem up to the compatible function approximation error ϵ_a.

3.2. Proof Sketch

By using the smoothness of $\nabla_\theta \log \pi_\theta(a|s)$ and performance difference lemma (Kakade & Langford, 2002; Agarwal et al., 2020), we have the following drift bound:

$$\Phi(\pi_{t+1}) - \Phi(\pi_t) \le -\eta\lambda\Phi(\pi_t) + \eta\sqrt{C^\star\epsilon_a}$$
$$- \eta(1-\gamma)\left(V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_t}(\mu)\right)$$
$$- \eta\sum_{s,a} d_\mu^{\pi^*}(s)\pi^*(a|s)V_\lambda^{\pi_t}(s) + \eta^2\|w_t\|_2^2.$$

By Lemma 2,

$$\eta^2\|w_t\|_2^2 \le \frac{2\eta^2}{\sigma^2(1-\gamma)^2}\left(\sum_s d_\mu^{\pi_t}(s)V_\lambda^{\pi_t}(s)\right)^2.$$

The step-size specified in Lemma 2 implies that

$$-\eta\sum_{s,a} d_\mu^{\pi^*}(s)\pi^*(a|s)V_\lambda^{\pi_t}(s) + \eta^2\|w_t\|_2^2 \le 0.$$

Then, by induction, we prove both non-singularity of $G_\lambda^{\pi_t}(\mu)$ and the drift bound. The convergence results in Theorem 1 follows directly by using the drift bound for $\Phi(\pi_t)$ and then to bound the optimality gap. The detailed analysis can be found in Appendix.

4. Conclusion and Future Work

In this work, we analyzed the convergence of natural policy gradient under softmax parameterization with linear function approximation, and established sharp finite-time convergence bounds. In particular, we proved that entropy-regularized NPG with linear function approximation achieves linear convergence rate, which is significantly faster than the sublinear rates previously obtained in the function approximation setting.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.

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

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

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

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

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Bhatnagar, S., Ghavamzadeh, M., Lee, M., and Sutton, R. S. Incremental natural actor-critic algorithms. *Advances in neural information processing systems*, 20:105–112, 2007.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pp. 1329–1338. PMLR, 2016.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870. PMLR, 2018.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

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

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

Lan, G. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135*, 2021.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

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

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

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. Trust-pcl: An off-policy trust region method for continuous control. *arXiv preprint arXiv:1707.01891*, 2017.

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

Scherrer, B., Ghavamzadeh, M., Gabillon, V., Lesner, B., and Geist, M. Approximate modified policy iteration and its application to the game of tetris. *J. Mach. Learn. Res.*, 16:1629–1676, 2015.

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

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

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. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

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

Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y., et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pp. 1057–1063. Citeseer, 1999.

Szepesvári, C. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

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

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

A. Convergence Analysis of Entropy-Regularized NPG

In this section, we will prove Theorem 1. The following lemmas will be useful in the proof.

Lemma 3 (Performance difference lemma)**.** *For any $\theta, \theta' \in \mathbb{R}^d$ and $\mu \in \Delta(\mathcal{S})$, we have:*

$$V_\lambda^{\pi_\theta}(\mu) - V_\lambda^{\pi_{\theta'}}(\mu) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)}\Big[A_\lambda^{\pi_{\theta'}}(s,a) + \lambda \log\frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)}\Big], \tag{13}$$

where $A_\lambda^{\pi_\theta}$ is the advantage function defined in (7).

Lemma 3 is an extension of the performance difference lemma in (Kakade, 2001).

The following lemma directly follows from $\log \pi_\theta(a|s) = \phi_{s,a} - \mathbb{E}_{a'\sim\pi_\theta(\cdot|s)}\phi_{s,a'}$, and it was first used in (Agarwal et al., 2020).

Lemma 4 (Smoothness of $\log \pi_\theta(a|s)$)**.** *For any $(s,a) \in \mathcal{S}\times\mathcal{A}$, $\log\pi_\theta(a|s)$ is smooth:*

$$\|\nabla_\theta \log \pi_\theta(a|s) - \nabla_\theta \log \pi_{\theta'}(a|s)\|_2 \le \|\theta - \theta'\|_2, \tag{14}$$

for any $\theta, \theta' \in \mathbb{R}^d$.

The proof of Theorem 1 relies on a Lyapunov analysis with the Lyapunov function Φ in Definition 1.

Proof of Theorem 1. For any $t \ge 0$, we have the following Lyapunov drift inequality:

$$\Phi(\pi_{t+1}) - \Phi(\pi_t) = \sum_{s,a} d_\mu^{\pi^*}(s)\pi^*(a|s)\log\frac{\pi_t(a|s)}{\pi_{t+1}(a|s)}, \tag{15}$$

$$\le -\eta\sum_{s,a} d_\mu^{\pi^*}(s)\pi^*(a|s)\nabla_\theta^\top \log\pi_t(a|s)w_t + \frac{\eta^2\|w_t\|_2^2}{2}, \tag{16}$$

which follows from the smoothness of $\log\pi_\theta(a|s)$ shown in Lemma 4 (Agarwal et al., 2020). Then, by adding and subtracting the advantage function $A_\lambda^{\pi_t}(s,a)$ into the sum on the RHS of the above inequality, and using the definition of $A_\lambda^{\pi_t}(s,a)$, we have:

$$\begin{aligned}\Phi(\pi_{t+1}) - \Phi(\pi_t) \le{}& -\eta\sum_{s,a} d_\mu^{\pi^*}(s)\pi^*(a|s)\Big(\nabla_\theta^\top\log\pi_t(a|s)w_t - \big(Q_\lambda^{\pi_t}(s,a) - \lambda\log\pi_t(a|s)\big)\Big)\\ &- \eta\sum_{s,a} d_\mu^{\pi^*}(s)\pi^*(a|s)V_\lambda^{\pi_t}(s) - \eta\lambda\Phi(\pi_t) - \eta(1-\gamma)\Big(V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_t}(\mu)\Big) + \frac{1}{2}\eta^2\|w_t\|_2^2,\end{aligned} \tag{17}$$

where we used Lemma 3 for $\sum_{s,a} d_\mu^{\pi^*}(s)\pi^*(a|s)\Big(A_\lambda^{\pi_t}(s,a) + \lambda\log\frac{\pi_t(a|s)}{\pi^*(a|s)}\Big)$. In the following, we bound the terms in (17).

$$\begin{aligned}&-\eta\sum_{s,a} d_\mu^{\pi^*}(s)\pi^*(a|s)\Big(\nabla_\theta^\top\log\pi_t(a|s)w_t - \big(Q_\lambda^{\pi_t}(s,a) - \lambda\log\pi_t(a|s)\big)\Big)\\ &\quad\le \eta\sqrt{\sum_{s,a} d_\mu^{\pi^*}(s)\pi^*(a|s)\Big(\nabla_\theta^\top\log\pi_t(a|s)w_t - \big(Q_\lambda^{\pi_t}(s,a) - \lambda\log\pi_t(a|s)\big)\Big)^2},\\ &\quad\le \eta\sqrt{C_t L(w_t,\theta_t)}\\ &\quad\le \eta\sqrt{C^\star\epsilon_a},\end{aligned} \tag{18}$$

where the second inequality holds by Cauchy-Schwarz inequality and Assumption 1, which implies $C_t \le C^\star < \infty$, and the last inequality follows from Assumption 2.

By the entropy-regularized NPG update, we have

$$w_t = \frac{1}{1-\gamma}\Big[G_\lambda^{\pi_t}(\mu)\Big]^{-1}\nabla_\theta V_\lambda^{\pi_t}(\mu).$$

We have $\|\nabla_\theta \log \pi_t(a|s)\|_2 \leq 2$ by Cauchy-Schwarz inequality. From Proposition 2, we conclude that:

$$\|\nabla_\theta V_\lambda^{\pi_t}(\mu)\|_2 \leq \frac{2}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_t}, a\sim\pi_t(\cdot|s)}[Q_\lambda^{\pi_t}(s,a) - \lambda \log \pi_t(a|s)], \tag{19}$$

since $Q_\lambda^{\pi_t}(s,a) - \lambda \log \pi_t(a|s) \geq 0$. By Proposition 1, this implies the following:

$$\|\nabla_\theta V_\lambda^{\pi_t}(\mu)\|_2 \leq \frac{2}{1-\gamma}\sum_s d_\mu^{\pi_t}(s) V_\lambda^{\pi_t}(s). \tag{20}$$

By Lemma 2, we have $\|[G_\lambda^{\pi_t}(\mu)]^{-1}\|_2 \leq 1/\sigma$. Using these two results with Cauchy-Schwarz inequality, we obtain:

$$\|w_t\|_2 \leq \frac{2/\sigma}{1-\gamma}\sum_s d_\mu^{\pi_t}(s) V_\lambda^{\pi_t}(s). \tag{21}$$

By substituting (18) and (21) into (17), we have the following inequality:

$$\begin{aligned}\Phi(\pi_{t+1}) - \Phi(\pi_t) \leq -\eta\lambda\Phi(\pi_t) - \eta(1-\gamma)\Big(V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_t}(\mu)\Big) + \eta\sqrt{C^\star \epsilon_a} \\ - \eta\sum_{s,a} d_\mu^{\pi^*}(s)\pi^*(a|s)V_\lambda^{\pi_t}(s) + \frac{2\eta^2}{\sigma^2(1-\gamma)^2}\Big(\sum_s d_\mu^{\pi_t}(s)V_\lambda^{\pi_t}(s)\Big)^2,\end{aligned} \tag{22}$$

where the step-size η is chosen to make the summation of the last two terms on the RHS negative by using the bounds on $V_\lambda^\pi(s)$ provided in (8). Therefore, we have the following Lyapunov drift inequality:

$$\Phi(\pi_{t+1}) - \Phi(\pi_t) \leq -\eta\lambda\Phi(\pi_t) - \eta(1-\gamma)\Big(V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_t}(\mu)\Big) + \eta\sqrt{C^\star\epsilon_a}. \tag{23}$$

We will use (23) in two ways to obtain the results in Theorem 1. First, note that

$$\Phi(\pi_{t+1}) \leq (1-\eta\lambda)\Phi(\pi_t) + \eta\sqrt{C^\star\epsilon_a}. \tag{24}$$

By induction and noting that $\Phi(\pi_0) \leq \log|\mathcal{A}|$, we obtain:

$$\Phi(\pi_T) \leq (1-\eta\lambda)^T \log|\mathcal{A}| + \frac{\sqrt{\epsilon_a C^\star}}{\lambda}, \tag{25}$$

for any $T \geq 1$.

Using the bound on $\Phi(\pi_T)$ and rearranging the terms in the Lyapunov drift inequality (23), we bound the optimality gap:

$$V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_T}(\mu) \leq \big(1-\lambda\eta\big)^T \frac{\log|\mathcal{A}|}{\eta(1-\gamma)} + \frac{\sqrt{\epsilon_a C^\star}}{\lambda\eta(1-\gamma)},$$

which concludes the proof. □

In the following, we provide a proof sketch for Lemma 2.

Proof sketch for Lemma 2. The proof consists of three steps.

Step 1: For any policy $\pi \in \Pi$ with $\min_{a\in\mathcal{A}, s\in supp(\mu)} \pi(a|s) \geq p$ for $p > 0$, we can show that $\sigma_1(G_\lambda^\pi(\mu)) \geq \sigma(p) > 0$ for some $\sigma(p) > 0$ under Assumption 3. The proof follows from noting that

$$u^\top F(\mu) u = \sum_s \mu(s)\text{Var}_{a\sim\text{Unif}(\mathcal{A})}(\phi_{s,a}^\top u),$$

for any $u \in \mathbb{R}^d$. Therefore, for any $u \in \mathbb{R}^d$ and $s \in \mathcal{S}$ such that $\mu(s)\text{Var}_{a\sim\text{Unif}(\mathcal{A})}(\phi_{s,a}^\top u) > 0$, we have $\mu(s)\text{Var}_{a\sim\pi(\cdot|s)}(\phi_{s,a}^\top u) > 0$ since $\min_{a\in\mathcal{A}} \pi(a|s) > 0$.

Step 2: In the second step, we show that $\Phi(\pi) \leq \varepsilon$ for $\varepsilon > 0$ implies the following:

$$\pi(a|s) \geq \exp\Big(-\frac{\varepsilon}{\delta_\mu^*(s,a)} - \frac{\mathcal{H}(\pi^*(\cdot|s))}{\pi^*(a|s)}\Big) = p^*(s,a,\varepsilon) > 0,$$

for any $s \in supp(d_\mu^{\pi^*})$ where δ_μ^* is the state-action visitation distribution under π^*. This bound directly follows from the definition of the potential function Φ (see Definition 1).

Step 3: Let $p = \min_{s\in\mu(s),a\in\mathcal{A}} p^*(s,a,\varepsilon)$ for

$$\varepsilon = \log|\mathcal{A}| + \frac{\sqrt{C^\star \cdot \epsilon_a}}{\lambda}.$$

For any policy $\pi \in \Pi$ with

$$\min_{s\in supp(\mu),a\in\mathcal{A}} \pi(a|s) \geq p,$$

we have shown in Step 1 that $\sigma_1(G_\lambda^\pi(\mu)) \geq \sigma(p) = \sigma$. Let the step-size be $\eta \leq \min\{\sigma^2\eta_0, \frac{1}{2\lambda}\}$ where

$$\eta_0 = \frac{(1-\gamma)^2 r_{min}}{(r_{max} + \lambda|\mathcal{A}|)^2},$$

and let

$$\tau = \inf\Big\{t \geq 1 : \eta > \min\Big\{\Big(\sigma_1(G_\lambda^{\pi_t}(\mu)\Big)^2\eta_0, \frac{1}{2\lambda}\Big\}\Big\}.$$

Then, the inequality (25) holds for any $t < \tau$. Lemma 2 holds if and only if $\tau = \infty$. Suppose to the contrary that $\tau < \infty$. Hence,

$$\begin{aligned}\Phi(\pi_\tau) &\leq (1-\lambda\eta)^\tau \log|\mathcal{A}| + \frac{\sqrt{C^\star \cdot \epsilon_a}}{\lambda},\\ &\leq \log|\mathcal{A}| + \frac{\sqrt{C^\star \cdot \epsilon_a}}{\lambda},\end{aligned}$$

which implies

$$\min_{\substack{s\in supp(\mu)\\ a\in\mathcal{A}}} \pi_\tau(a|s) \geq p,$$

and therefore $\sigma_1(G_\lambda^{\pi_\tau}(\mu)) \geq \sigma(p)$ by Step 1, which contradicts with the definition of τ. This implies that $\tau = \infty$. Hence, the inequality in (25) holds and we have $\min_{s\in supp(\mu),a\in\mathcal{A}} \pi_t(a|s) \geq p > 0$ and $\sigma_1(G_\lambda^{\pi_t}(\mu)) \geq \sigma(p) = \sigma > 0$ for any $t \geq 1$, which concludes the proof. □