Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, Yuejie Chi

To cite this article:
Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, Yuejie Chi (2021) Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization. Operations Research

Published online in Articles in Advance 02 Dec 2021

. https://doi.org/10.1287/opre.2021.2151

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–16
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization

Shicong Cen,[a] Chen Cheng,[b] Yuxin Chen,[c] Yuting Wei,[d] Yuejie Chi[a]

[a] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; [b] Department of Statistics, Stanford University, Stanford, California 94305; [c] Department of Electrical and Computer Engineering, Princeton University, Princeton, New Jersey 08544; [d] Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Contact: shicongc@andrew.cmu.edu (SC); chencheng@stanford.edu (CC); yuxin.chen@princeton.edu, https://orcid.org/0000-0001-9256-5815 (YC); ytwei@wharton.upenn.edu, https://orcid.org/0000-0002-3041-3434 (YW); yuejiechi@cmu.edu, https://orcid.org/0000-0002-6766-5459 (YC)

Received: August 5, 2020
Revised: December 21, 2020
Accepted: April 6, 2021
Published Online in Articles in Advance: December 2, 2021

OR/MS Subject Classifications: Analysis of algorithms: computational complexity; decision analysis: theory
Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2021.2151

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 *nonasymptotic* 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.

Funding: This work was supported by the National Science Foundation, Air Force Office of Scientific Research, Army Research Office, and Office of Naval Research. S. Cen and Y. Chi are supported in part by 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 National Science Foundation [Grants CCF-2106778, CCF-2007911, and DMS-2015447/2147546]. Y. Chen is supported in part by Grants AFOSR awards FA9550-19-1-0030 and FA9550-22-1-0198, ONR N00014-19-1-2120 and N00014-22-1-2354, ARO YIP award W911NF-20-1-0097, ARO W911NF-18-1-0303, NSF CCF-2106739/2221009, CCF-1907661, IIS-1900140 and IIS-2100158/2218773, the Google Research Scholar Award, and the Alfred P. Sloan Research Fellowship.

Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2021.2151.

Keywords: reinforcement learning • natural policy gradient methods • entropy regularization • global convergence

1. Introduction

Policy gradient (PG) methods and their variants (Williams 1992, Konda and Tsitsiklis 2000, Sutton et al. 2000, Kakade 2002, Peters and Schaal 2008), which aim to optimize (parameterized) policies via gradient-type methods, lie at the heart of recent advances in reinforcement learning (RL) (see, e.g., Mnih et al. (2015), Schulman et al. (2015), Silver et al. (2016), and 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*

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

(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 because of the intrinsic nonconcavity 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, Cen et al. 2021, Peters et al. 2010, Duan et al. 2016, Mnih et al. 2016, Haarnoja et al. 2017, Hazan et al. 2019, Xiao et al. 2019, Vieillard et al. 2020). 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 toward understanding PG methods (Bhandari and Russo 2019, Agarwal et al. 2020b, Mei et al. 2020), we aim to develop nonasymptotic 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 toward demystifying the effectiveness of entropy-regularized policy optimization in more complex settings.

1.2.1. Settings.

Consider a γ-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)

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

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, where $\tau > 0$ is the regularization parameter, $0 < \eta \le \frac{1-\gamma}{\tau}$ is the learning rate (or stepsize), $\pi^{(t)}$ indicates the t-th policy iterate, and Q_τ^π is the soft Q-function under policy π (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).

1.2.2. Our Contributions.

The results of this paper deliver fully nonasymptotic convergence rates of entropy-regularized NPG methods without any hidden constants, which are previewed as follows (in an orderwise manner). The definition of ϵ-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 ϵ-optimal policy for the regularized MDP (cf. Table 1), the algorithm (1) with a general learning rate $0 < \eta \le \frac{1-\gamma}{\tau}$ 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

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

Table 1. The Iteration Complexities of NPG Methods to Reach ϵ-Accuracy in Terms of Optimization Error, Where the Unregularized (Resp. Regularized) Version is Given by (13) (cf. (15)) with η the Learning Rate

Paper	Iteration complexity upper bound	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)	$\frac{1}{(1-\gamma)\min_{s\in\mathcal{S}}\rho(s)}\log(\frac{1}{\epsilon})$	Unregularized	exact line search
This work	$\frac{1}{1-\gamma}\log(\frac{1}{\epsilon})$	Regularized	constant: $\frac{1-\gamma}{\tau}$
This work	$\frac{1}{\eta\tau}\log(\frac{1}{\epsilon})$	Regularized	constant: $(0,\frac{1-\gamma}{\tau})$

Notes. We assume exact gradient evaluation and softmax parameterization and hide the dependencies that are logarithmic on problem parameters. Here, ϵ-accuracy or ϵ-optimality for the unregularized (resp. regularized) case means that $V^{\star}(s)-V^{\pi^{(t)}}(s)\le\epsilon$ (resp. $V_{\tau}^{\star}(s)-V_{\tau}^{\pi^{(t)}}(s)\le\epsilon$) holds simultaneously for all $s\in\mathcal{S}$; ρ denotes the initial state distribution, which clearly obeys $\frac{1}{\min_{s\in\mathcal{S}}\rho(s)}\ge|\mathcal{S}|$.

learning rate $0<\eta\le\frac{1-\gamma}{\tau}$ 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 entry-wise error of the soft Q-function estimates (see Theorem 2).

- **Quadratic convergence in the small-ϵ regime**. In the high-accuracy regime, where the target level ϵ is very small, the algorithm (1) with $\eta=\frac{1-\gamma}{\tau}$ converges superlinearly, in the sense that the iteration complexity to reach ϵ-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).

1.2.3. Comparisons with Prior Art. Agarwal et al. (2020b) proved that unregularized NPG methods with softmax parameterization attain an ϵ-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 step sizes 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 by 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 step sizes 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 $Q_{\tau}^{\star}-Q_{\tau}^{(t)}$ and $Q_{\tau}^{\star}-\tau\log\xi^{(t)}$ (to be defined later) and invoke the linear system theory to establish appealing contractions, which allow for the use of more aggressive constant step sizes 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, whereas 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 $O(1/\epsilon^2)$, whereas 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 that takes softmax PG methods (super)-exponential time (in terms of $|\mathcal{S}|$ and $\frac{1}{1-\gamma}$) to converge. In contrast, the bounds derived in the current paper are fully nonasymptotic, delineating clear dependencies on all salient problem parameters, which clearly demonstrate the

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

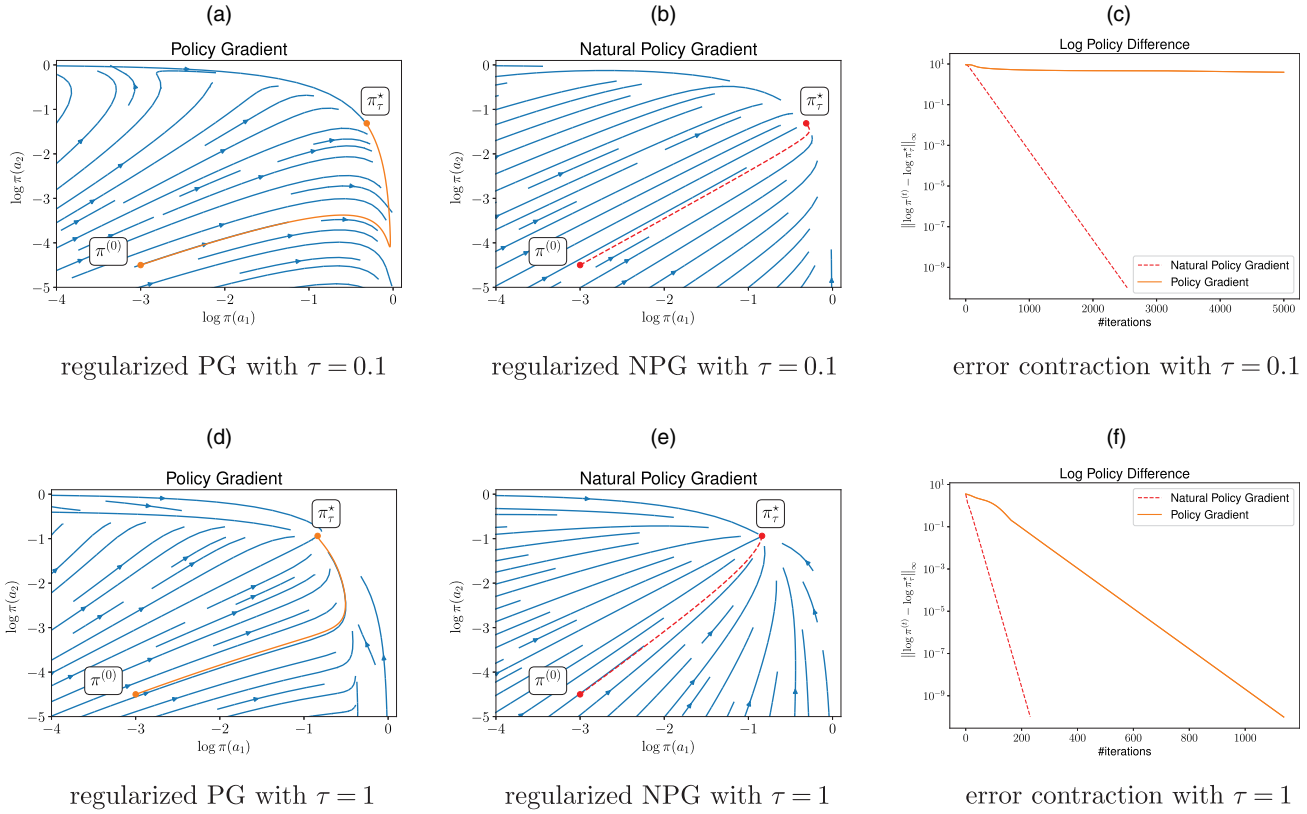
algorithmic advantages of NPG methods. Figure 1 depicts the policy paths of PG and NPG methods with entropy regularization for a simple bandit problem with three actions. It is evident from the plots that the NPG method follows a more direct path to the global optimum compared with the PG counterpart and hence, converges faster. In addition, both algorithms converge more rapidly as the regularization parameter τ 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), and 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), and 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) and 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

Figure 1. (Color online) Comparisons of PG and NPG Methods with Entropy Regularization for a Bandit Problem ($\gamma = 0$) with Three Actions, Whose Corresponding Rewards are 1.0, 0.9, and 0.1, Respectively



Notes. 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 π(a1), log π(a2)) 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 π(a1), log π(a2)) 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).

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

(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 suboptimality gap of soft policy iteration is small if the policy improvement is small in consecutive iterations.

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 $z = [z_i]_{1\le i\le n} \in \mathbb{R}^n$, the notation $|\cdot|$ denotes $|z| := [|z_i|]_{1\le i\le n}$; other functions are defined analogously. For any vectors $z = [z_i]_{1\le i\le n}$ and $w = [w_i]_{1\le i\le n}$, the notation $z \ge w$ (resp. $z \le w$) means $z_i \ge w_i$ (resp. $z_i \le w_i$) for all $1 \le i \le n$. The softmax function $\mathsf{softmax}: \mathbb{R}^n \mapsto \mathbb{R}^n$ is defined such that $[\mathsf{softmax}(\theta)]_i := \exp(\theta_i)/(\Sigma_i \exp(\theta_i))$ for a vector $\theta = [\theta_i]_{1\le i\le n} \in \mathbb{R}^n$. Given two probability distributions π_1 and π_2 over $\mathcal{A}$, the Kullback-Leibler (KL) divergence from π_2 to π_1 is defined by $\mathsf{KL}(\pi_1\|\pi_2) := \Sigma_{a\in\mathcal{A}}\pi_1(a)\log\frac{\pi_1(a)}{\pi_2(a)}$. Given two probability distributions p and q over $\mathcal{S}$, we introduce the notation $\|\frac{p}{q}\|_\infty := \max_{s\in\mathcal{S}}\frac{p(s)}{q(s)}$ and $\|\frac{1}{q}\|_\infty := \max_{s\in\mathcal{S}}\frac{1}{q(s)}$.

2. Model and Algorithms

2.1. Problem Settings

2.1.1. 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, $P: \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, and $r: \mathcal{S}\times\mathcal{A} \to [0,1]$ stands for the reward function.[1] 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 $\pi: \mathcal{S} \to \Delta(\mathcal{A})$ represents a (randomized) action selection rule; namely, $\pi(a|s)$ specifies the probability of executing action a in state s for each $(s,a) \in \mathcal{S}\times\mathcal{A}$.

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

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

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

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

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

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

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

2.1.3. 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 π given the initial state $s_0 \in \mathcal{S}$ is defined by

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

where the trajectory $(s_0, s_1, \cdots)$ is generated by the MDP $\mathcal{M}$ under policy π 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 ρ over $\mathcal{S}$, we define the distribution d_ρ^π as follows

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

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

2.1.4. 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 that naturally ensures that the policy lies in the probability simplex. Specifically, for any $\theta: \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ (called "logic values"), the corresponding softmax policy π_θ is generated through the softmax transform

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

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

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

2.1.5. 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

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

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

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

Equivalently, V_τ^π can be viewed as the value function of π by adjusting the instantaneous reward to be the policy-dependent regularized version as follows:

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

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_τ^π of a policy π, also known as the soft Q-function,[2] is related to V_τ^π as

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

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

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

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

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

2.2. Algorithm: NPG Methods With Entropy Regularization

2.2.1. Natural Policy Gradient Methods. Toward 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 θ until convergence, a first-order method commonly referred to as the *policy gradient* (PG) algorithm (see, e.g., Sutton et al. 2000). In comparison, the *natural policy gradient* (NPG) method (Kakade 2002) adopts a preconditioned gradient update rule

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

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

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

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., Section 7 in Schulman et al. 2015).

2.2.2. NPG Methods With Entropy Regularization. Equipped with entropy regularization, the NPG update rule can be written as

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

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, which, interestingly, is invariant to the choice of ρ. More precisely, if we let $\theta^{(t)}$ denote the tth iterate and $\pi^{(t)} = \mathsf{softmax}(\theta^{(t)})$ the associated policy, then the entropy-regularized NPG updates satisfy

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

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

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

Schulman et al. 2015 and Shani et al. (2019). As an important special case, the update rule (16) reduces to

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

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 Section 4.1 in Haarnoja et al. 2018).

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 η, initialization $\pi^{(0)}$.
2. For $t = 0, 1, 2, \cdots$ do
3. Compute the regularized Q-function $Q_\tau^{(t)}$. (defined in (11a)) of policy $\pi^{(t)}$.
4. Update the policy:

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

where

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

2.3. A Warm-Up Example: The Bandit Case

Inspired by Schulman et al. (2017a) and 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 $\pi_\theta : \mathcal{A} \to \Delta(\mathcal{A})$ that solves the following optimization problem

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

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

$$\pi_\tau^\star = \mathsf{softmax}(\mathrm{r}/\tau), \tag{20}$$

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

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

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

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

$$\left\|\log\pi^{(t)} - \log\pi_\tau^\star\right\|_\infty \le 2(1-\tau\eta)^t\left\|\log\pi^{(0)} - \log\pi_\tau^\star\right\|_\infty.$$

Although 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 that 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 $Q_\tau^\star$ and the associated log policy $\log\pi_\tau^\star$—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* $0 < \eta \le (1-\gamma)/\tau$, *the entropy-regularized NPG updates* (18) *satisfy*

$$\left\|Q_\tau^\star - Q_\tau^{(t+1)}\right\|_\infty \le C_1\gamma(1-\eta\tau)^t \tag{22a}$$

$$\left\|\log\pi_\tau^\star - \log\pi^{(t+1)}\right\|_\infty \le 2C_1\tau^{-1}(1-\eta\tau)^t \tag{22b}$$

for all $t \ge 0$, *where*

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

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

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

• **Linear convergence of soft Q-functions.** To reach $\|Q_\tau^\star - Q_\tau^{(t)}\|_\infty \le \epsilon$, the entropy-regularized NPG method needs at most $\frac{1}{\eta\tau}\log(\frac{C_1\gamma}{\epsilon})$ 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 η is fixed in the admissible range, the iteration complexity scales inverse proportionally with τ, suggesting that 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 $\frac{1}{\eta\tau}\log(\frac{2C_1}{\epsilon\tau})$ iterations to yield $\|\log\pi_\tau^\star - \log\pi^{(t+1)}\|_\infty \le \epsilon$.

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

$$\|V_\tau^\star - V_\tau^{(t+1)}\|_\infty \le 3C_1\gamma(1-\eta\tau)^t. \tag{24}$$

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

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \qquad V_\tau^\star(s) = -\tau\,\log\,\pi_\tau^\star(a|s) + Q_\tau^\star(s,a),$$

$$\Rightarrow \qquad V_\tau^\star(s) = \mathop{\mathbb{E}}_{a\sim\pi^{(t+1)}(\cdot|s)}[-\tau\log\pi_\tau^\star(a|s) + Q_\tau^\star(s,a)].$$

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

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

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 $\|Q_\tau^\star - Q_\tau^{(t)}\|_\infty \le \epsilon$ reduces to

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

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

3.1.1. Comparison With Entropy-Regularized Policy Gradient Methods.

Theorem 6 in Mei et al. (2020) proved that the entropy-regularized policy gradient method achieves[4]

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

and they further showed that $\inf_{k\ge0}\min_{s,a}\pi^{(k)}(a|s)$ is nonvanishing in t. It remains unclear, however, how $\inf_{t\ge0}\min_{s,a}\pi^{(t)}(a|s)$ 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.

3.1.2. Computing an ϵ-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 $V_\tau^\star$, as opposed to finding the optimal value function $V^\star$ of the original MDP. Fortunately, by choosing the regularization parameter τ to be sufficiently small (in accordance with the target accuracy level ϵ), we can guarantee that $V_\tau^\star \approx V^\star$ (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 τ:

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

and assume the error of the regularized value function satisfies $\|V_\tau^\star - V_\tau^{(t)}\|_\infty < \epsilon/2$. By virtue of Theorem 1, this optimization accuracy can be achieved via no more than $\frac{4\log|\mathcal{A}|}{(1-\gamma)\eta\epsilon}\log(\frac{2C_1\gamma}{\epsilon})$ iterations of entropy-regularized NPG updates with a general learning rate[5] or no more than $\frac{1}{1-\gamma}\log\left(\frac{\gamma\|Q_\tau^\star - Q_\tau^{(0)}\|_\infty}{\epsilon}\right)$ iterations with the

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

specific choice $\eta = \frac{1-\gamma}{\tau}$. It then follows that

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

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

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

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

3.1.3. Convergence Guarantee for Conservative Policy Iteration.

Our analysis framework also leads to a similar convergence guarantee for a type of policy update adopted in *conservative policy iteration* (CPI; see 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 the supplemental material 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 $t \ge 0$,

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

Here, δ 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 (see, 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.

Theorem 2 (Linear Convergence of Approximate Entropy-Regularized NPG). *When $0 < \eta \le (1-\gamma)/\tau$, the inexact entropy-regularized NPG updates (26) satisfy*

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

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

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

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

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

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

then the algorithm (26) achieves 2ϵ-accuracy (i.e., $\|Q_\tau^\star - Q_\tau^{(t)}\|_\infty \le 2\epsilon$) 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 δ 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 π, model-based policy evaluation achieves $\|\widehat{Q}_\tau^\pi - Q_\tau^\pi\|_\infty \le \delta$ 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 $\frac{\delta=(1-\gamma)^2\epsilon}{2\gamma}$ and invoke the union bound over $\tilde{O}(\frac{1}{1-\gamma})$ 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 ϵ-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.

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

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-ϵ Regime

Somewhat remarkably, the regularized NPG method with $\eta = \frac{1-\gamma}{\tau}$ achieves superlinear convergence in computing $V_\tau^\star$ 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 $\pi_\tau^\star$, denoted by $\mu_\tau^\star \in \Delta(\mathcal{S})$. It is straightforward to verify the following basic property

$$d_{\mu_\tau^\star}^{\pi_\tau^\star} = \mu_\tau^\star, \tag{29}$$

given that the state visitation distribution remains unchanged if the initial state is already in a steady state. Throughout this paper, we assume that $\min_s \mu_\tau^\star(s) > 0$. 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*

$$\left\| \log \pi^{(t)} - \log \pi_\tau^\star \right\|_\infty \le 1. \tag{30}$$

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

$$\begin{aligned} & V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) \\ & \le \left\| \frac{\rho}{\mu_\tau^\star} \right\|_\infty \frac{(1-\gamma)\tau}{4\gamma^2} \left\| \frac{1}{\mu_\tau^\star} \right\|_\infty^{-1} \left(\frac{4\gamma^2}{(1-\gamma)\tau} \left\| \frac{1}{\mu_\tau^\star} \right\|_\infty \right. \\ & \qquad \left. \times \left(V_\tau^\star(\mu_\tau^\star) - V_\tau^{(0)}(\mu_\tau^\star) \right) \right)^{2^t}. \end{aligned}$$

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

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

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

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 τ. 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 $V_\tau^\star(\rho)$) as opposed to the accuracy w.r.t. the original unregularized MDP. Thus, interpreting Theorem 3 in practice requires caution, since the approximation error $V_\tau^\star(\rho) - V^\star(\rho)$ 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.

4.1.1. 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 $0 < \eta \le (1-\gamma)/\tau$. For any distribution ρ, one has*

$$\begin{aligned} & V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) \\ & = \mathop{\mathbb{E}}_{s \sim d_\rho^{(t+1)}} \left[\left(\frac{1}{\eta} - \frac{\tau}{1-\gamma} \right) \mathsf{KL}(\pi^{(t+1)}(\cdot|s) \| (\pi^{(t)}(\cdot|s)) \right. \\ & \quad \left. + \frac{1}{\eta} \mathsf{KL}(\pi^{(t)}(\cdot|s) \| (\pi^{(t+1)}(\cdot|s)) \right]. \end{aligned} \tag{32}$$

Proof. See the supplemental material. □

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 ρ readily reveals a sort of pointwise monotonicity for the above range of learning rates in the sense that $V_\tau^{(t+1)}(s) \ge V_\tau^{(t)}(s)$ 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

$$\begin{aligned} Q_\tau^{(t+1)}(s,a) &= r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(|s,a)} [V_\tau^{(t+1)}(s')] \\ &\ge r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(|s,a)} [V_\tau^{(t)}(s')] = Q_\tau^{(t)}(s,a), \end{aligned} \tag{33}$$

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

4.1.2. A Key Contraction Operator: The Soft Bellman Optimality Operator.

An operator that plays a pivotal role in the theory of dynamic programming (Bellman

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

1952) is the renowned Bellman optimality operator $\mathcal{T}: \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, defined as follows

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

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

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

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

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

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 Section 3.1 in Dai et al. (2018).

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:*

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

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

$$\mathcal{T}_\tau(Q_\tau^\star) = Q_\tau^\star. \tag{37}$$

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

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

Proof. See the supplemental material. □

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 γ-contraction property (38) and the fixed-point property (37), the following soft Q-value iteration

$$Q_{\text{svi}}^{(t+1)} = \mathcal{T}_\tau\left(Q_{\text{svi}}^{(t)}\right), \qquad t \ge 0$$

is guaranteed to converge linearly to the optimal $Q_\tau^\star$ with a contraction rate γ, 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

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

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)), and (iv) uses the form of $\pi^{(t+1)}$ in (17), whereas (v) makes use of the expression (36). The inequality (39) further leads to $0 \le Q_\tau^\star - Q_\tau^{(t+1)} \le Q_\tau^\star - \mathcal{T}_\tau(Q_\tau^{(t+1)})$, and hence,

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

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

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

where the second relation comes from Equation (12) in Nachum et al. (2017). It then follows from an elementary property of the softmax function that

$$\begin{aligned}\|\log\pi^{(t+1)} - \log\pi_\tau^\star\|_\infty &\le \frac{2}{\tau}\|Q_\tau^{(t)} - Q_\tau^\star\|_\infty \\ &\le \frac{2}{\tau}\gamma^t\|Q_\tau^\star - Q_\tau^{(0)}\|_\infty,\end{aligned}$$

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

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

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

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

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

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

and consequently,

$$\begin{aligned} \|Q_\tau^\star - \tau\log\xi^{(t+1)}\|_\infty &\leq \alpha\|Q_\tau^\star - \tau\log\xi^{(t)}\|_\infty \\ &\quad + (1-\alpha)\|Q_\tau^\star - Q_\tau^{(t)}\|_\infty. \end{aligned} \tag{44}$$

Step 1: A Linear System that Describes the Error Recursions. In the case with general learning rates, the estimation error $\|Q_\tau^\star - Q_\tau^{(t)}\|_\infty$ 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 $\|Q_\tau^\star - \tau\log\xi^{(t)}\|_\infty$. In what follows, we leverage a simple yet powerful technique by describing the dynamics concerning $\|Q_\tau^\star - Q_\tau^{(t)}\|_\infty$ and $\|Q_\tau^\star - \tau\log\xi^{(t)}\|_\infty$ via a linear system, whose spectral properties dictate the convergence rate. Toward this, we start with the following key observation, whose proof is deferred to the supplemental material.

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

$$\begin{aligned} \|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty &\leq \gamma\|Q_\tau^\star - \tau\log\xi^{(t+1)}\|_\infty \\ &\quad + \gamma\alpha^{t+1}\|Q_\tau^{(0)} - \tau\log\xi^{(0)}\|_\infty, \end{aligned} \tag{45}$$

where α 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 quantity, it is convenient to combine (44) and (45) into the following linear system

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

where

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

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

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

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

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 $L^{(t)} := (1-\alpha)\|Q_\tau^\star - Q_\tau^{(t)}\|_\infty + \alpha\|Q_\tau^\star - \tau\log\xi^{(t)}\|_\infty$ 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 Equation (46) and the nonnegativity of (A, x_t, y), it is immediate to deduce that

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

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 nonzero singular value is larger than 1). In addition, the Woodbury matrix inversion formula together with the decomposition (48) yields

$$\gamma(\alpha^{-1}A - I)^{-1}y = \gamma\left\{\begin{bmatrix} 1 & \frac{\alpha}{1-\alpha} \\ \frac{1}{\gamma} & \frac{\alpha}{(1-\alpha)\gamma}\end{bmatrix} - I\right\}$$

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

which is a nonnegative vector. Consequently, this taken together with (49) gives

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

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

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

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

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

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

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

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

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

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

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 $0 < \eta \le (1-\gamma)/\tau$. For any state $s_0 \in \mathcal{S}$, one has*

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

Proof. See the supplemental material. □

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 $\left\| \widehat{Q}_\tau^{(t)} - Q_\tau^{(t)} \right\|_\infty \le \delta$, repeating the argument of (33) reveals that the soft Q-function estimates are not far from being monotone in t in the sense that

$$\begin{aligned} \forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad & Q_\tau^{(t)}(s,a) - Q_\tau^{(t+1)}(s,a) \\ & = \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} [V_\tau^{(t)}(s') - V_\tau^{(t+1)}(s')] \le \frac{2\gamma\delta}{1-\gamma}. \end{aligned} \tag{56}$$

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. Toward this end, we adopt a similar approach as in (42) by introducing the following auxiliary sequence $\widehat{\xi}^{(t)}$ defined recursively using $\widehat{Q}_\tau^{(t)}$:

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

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

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:

$$z_{t+1} \le B z_t + b, \tag{58}$$

where

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

Here, the system matrix B (in particular its eigenvalues) governs the contraction rate, whereas 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 the supplemental material.

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 Suboptimality 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 suboptimality gap for each iteration.

Lemma 5 (Suboptimality Gap). *Suppose that $\eta = (1-\gamma)/\tau$. For any distribution ρ, one has*

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

$$V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) \le \frac{1}{\eta}\mathop{\mathbb{E}}_{s\sim d_\rho^{\pi_\tau^\star}}\left[\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\middle\|(\pi^{(t+1)}(\cdot|s)\right)\right].$$

Proof. This result has appeared in Eqn. (486) of Mei et al. (2020). For completeness, we include a proof in the supplemental material. □

In words, Lemma 5 formalizes the connection between the suboptimality 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 $\eta = (1-\gamma)/\tau$:

$$\begin{aligned}
&V_\tau^\star(\rho) - V_\tau^{(t+1)}(\rho)\\
&= V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) + \left(V_\tau^{(t)}(\rho) - V_\tau^{(t+1)}(\rho)\right)\\
&\overset{\text{(i)}}{=} V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) - \frac{1}{\eta}\mathop{\mathbb{E}}_{s\sim d_\rho^{(t+1)}}\left[\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\middle\|(\pi^{(t+1)}(\cdot|s)\right)\right]\\
&\overset{\text{(ii)}}{\le} V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) - \frac{1}{\eta}\left\|\frac{d_\rho^{\pi_\tau^\star}}{d_\rho^{(t+1)}}\right\|_\infty^{-1}\mathop{\mathbb{E}}_{s\sim d_\rho^{\pi_\tau^\star}}\left[\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\middle\|(\pi^{(t+1)}(\cdot|s)\right)\right]\\
&\overset{\text{(iii)}}{\le} V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) - \left\|\frac{d_\rho^{\pi_\tau^\star}}{d_\rho^{(t+1)}}\right\|_\infty^{-1}\left(V_\tau^\star(\rho) - V_\tau^{(t)}(\rho)\right)\\
&= \left(1 - \left\|\frac{d_\rho^{\pi_\tau^\star}}{d_\rho^{(t+1)}}\right\|_\infty^{-1}\right)\left(V_\tau^\star(\rho) - V_\tau^{(t)}(\rho)\right). \qquad (60)
\end{aligned}$$

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

Step 3: Superlinear Convergence in the Small-ϵ Regime. The contraction property (60) implies that $V_\tau^{(t+1)}(\rho)$ converges superlinearly to $V_\tau^\star$ once $\pi^{(t)}$ gets sufficiently close to $\pi_\tau^\star$. In fact, once the ratio $d_\rho^{(t+1)}/d_\rho^{\pi_\tau^\star}$ becomes sufficiently close to 1, the contraction factor $1 - \left\|d_\rho^{\pi_\tau^\star}/d_\rho^{(t+1)}\right\|_\infty^{-1}$ in (60) is approaching 0, thereby accelerating convergence. This observation underlies Theorem 3, whose complete analysis is postponed until the supplemental material.

5. Discussions

This paper establishes nonasymptotic 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 toward 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, Shani et al. 2019, and Agarwal et al. 2020b) or bootstrapping (see, e.g., Haarnoja et al. 2018, Wu et al. 2020, and Xu 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 Agarwal et al. 2019, 2020b, and Sutton et al. 2000) 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.

Endnotes

[1] 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.

[2] In this paper, we use the terms "regularized" value (resp. Q) functions and "soft" value (resp. Q) functions interchangeably.

[3] To see this, invoke the optimality of $\pi_\tau^\star$ and the elementary entropy bound $0 \le \mathcal{H}(\rho,\pi) \le \frac{1}{1-\gamma}\log|\mathcal{A}|$ to obtain

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

[4] Here, we have assumed that the exact policy gradient is computed with respect to $V_\tau^{(t)}(\rho)$.

[5] 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 $\eta \ge 2(1-\gamma)\log|\mathcal{A}|\log\left(\frac{2C_1\gamma}{\epsilon}\right)$. Consequently, our finding hints at the potential advantage of entropy-regularized NPG methods over the unregularized counterpart even when solving the original MDP.

References

Agarwal A, Jiang N, Kakade SM (2019) Reinforcement Learn.: Theory and algorithms. Technical report, University of Washington Seattle, Seattle, WA.

Agarwal A, Kakade S, Yang LF (2020a) Model-based reinforcement Learn. with a generative model is minimax optimal. *Proc. 33rd Conf. Learn. Theory*, 67–83.

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

Agarwal A, Kakade SM, Lee JD, Mahajan G (2020b) Optimality and approximation with policy gradient methods in Markov decision processes. Proc. 33rd Conf. Learn. Theory, 64–66.

Ahmed Z, Le Roux N, Norouzi M, Schuurmans D (2019) Understanding the impact of entropy on policy optimization. *Proc. 36th Internat. Conf. Machine Learn.*, 151–160.

Amari SI (1998) Natural gradient works efficiently in Learn.. *Neural Comput.* 10(2):251–276.

Azar MG, Munos R, Kappen HJ (2013) Minimax PAC bounds on the sample complexity of reinforcement Learn. with a generative model. *Machine Learn.* 91(3):325–349.

Bellman R (1952) On the theory of dynamic programming. *Proc. Natl. Acad. Sci. USA* 38(8):716.

Bertsekas DP (2017) *Dynamic Programming and Optimal Control*, 4th ed. (Athena Scientific, Belmont, MA).

Bhandari J, Russo D (2019) Global optimality guarantees for policy gradient methods. Preprint, submitted June 5, https://arxiv.org/abs/1906.01786.

Bhandari J, Russo D (2020) A note on the linear convergence of policy gradient methods. Preprint, submitted July 21, https://arxiv.org/abs/2007.11120.

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

Cai Q, Yang Z, Jin C, Wang Z (2019) Provably efficient exploration in policy optimization. *Proc. 37th Conf. Machine Learn.*, PMLR 119:1283–1294.

Cen S, Wei Y, Chi Y (2021) Fast Policy Extragradient Methods for Competitive Games with Entropy Regularization. arXiv preprint arXiv:2105.15186.

Dai B, Shaw A, Li L, Xiao L, He N, Liu Z, Chen J, Song L (2018) SBEED: Convergent reinforcement Learn. with nonlinear function approximation. *Proc. 35th Internat. Conf. Machine Learn.*, PMLR, 80:1125–1134.

Duan Y, Chen X, Houthooft R, Schulman J, Abbeel P (2016) Benchmarking deep reinforcement Learn. for continuous control. *Proc. 33rd Internat. Conf. Machine Learn.*, PMLR, 48:1329–1338.

Even-Dar E, Kakade SM, Mansour Y (2009) Online Markov decision processes. *Math. Oper. Res.* 34(3):726–736.

Fazel M, Ge R, Kakade S, Mesbahi M (2018) Global convergence of policy gradient methods for the linear quadratic regulator. *Proc. 35th Internat. Conf. Machine Learn.*, PMLR, 80:1467–1476.

Geist M, Scherrer B, Pietquin O (2019) A theory of regularized Markov decision processes. Internat. Conf. Machine Learn., 2160–2169.

Grill JB, Darwiche Domingues O, Menard P, Munos R, Valko M (2019) Planning in entropy-regularized markov decision processes and games. *Advances in Neural Information Processing Systems* 32:12404–12413.

Haarnoja T, Tang H, Abbeel P, Levine S (2017) Reinforcement Learn. with deep energy-based policies. Proc. 34th Internat. Conf. Machine Learn., PMLR, 70:1352–1361.

Haarnoja T, Zhou A, Abbeel P, Levine S (2018) Soft actor-critic: off-policy maximum entropy deep reinforcement Learn. with a stochastic actor. *Proc. 35th Internat. Conf. Machine Learn.*, PMLR, 80:1861–1870.

Hazan E, Kakade S, Singh K, Van Soest A (2019) Provably efficient maximum entropy exploration. *Proc. 36th Internat. Conf. Machine Learn.*, PMLR, 97:2681–2691.

Jansch-Porto JP, Hu B, Dullerud G (2020) Convergence guarantees of policy optimization methods for Markovian jump linear systems. *arXiv preprint arXiv:2002.04090*.

Kakade SM (2002) A natural policy gradient. *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA), 1531–1538.

Kakade S, Langford J (2002) Approximately optimal approximate reinforcement Learn. *Proc. 19th Internat. Conf. Machine Learn.*, 267–274.

Karimi B, Miasojedow B, Moulines É, Wai HT (2019) Non-asymptotic analysis of biased stochastic approximation scheme. Proc. Thirty-Second Conf. Learn. Theory, 99:1944–1974.

Konda VR, Tsitsiklis JN (2000) Actor-critic algorithms. *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA) 1008–1014.

Li G, Wei Y, Chi Y, Gu Y, Chen Y (2020a) Breaking the sample size barrier in model-based reinforcement Learn. with a generative model. *34th Conf. Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, BC, Canada, 12861–12872.

Li G, Wei Y, Chi Y, Gu Y, Chen Y (2020b) Sample complexity of asynchronous Q-Learn.: Sharper analysis and variance reduction. Preprint, submitted Jun 4, https://arxiv.org/abs/2006.03041.

Li G, Wei Y, Chi Y, Gu Y, Chen Y (2021b) Softmax policy gradient methods can take exponential time to converge. Belkin M, Kpotufe S, eds. *Proc. 34th Conf. Learning Theory* (PMLR), 134:3107–3110.

Li G, Cai C, Chen Y, Gu Y, Wei Y, Chi Y (2021a) Is Q-Learn. minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*.

Liu B, Cai Q, Yang Z, Wang Z (2019). Neural proximal trust region policy optimization attains globally optimal policy. *33rd Conf. Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, BC, Canada.

Mei J, Xiao C, Szepesvari C, Schuurmans D, (2020) On the global convergence rates of softmaz policy gradient methods. *Proc. 37th Internat. Conf. Machine Learn.*, PMLR, 119:6820–6829.

Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K (2016). Asynchronous methods for deep reinforcement Learn. *Proc. 33rd Internat. Conf. Machine Learn.*, PMLR, 48:1928–1937.

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, et al (2015) Human-level control through deep reinforcement Learn. *Nature* 518(7540):529–533.

Mohammadi H, Zare A, Soltanolkotabi M, Jovanović MR (2019). Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem. Preprint, submitted December 26, https://arxiv.org/abs/1912.11899.

Nachum O, Norouzi M, Xu K, Schuurmans D (2017). Bridging the gap between value and policy based reinforcement Learn.. Preprint, submitted November 22, https://arxiv.org/abs/1702.08892.

Nemirovsky AS, Yudin DB (1983) Problem complexity and method efficiency in optimization. (J. Wiley & Sons).

Nesterov Y (2009) Primal-dual subgradient methods for convex problems. *Math. Programming* 120(1):221–259.

Neu G, Jonsson A, Gómez V (2017). A unified view of entropy-regularized Markov decision processes. Preprint, submitted May 22, https://arxiv.org/abs/1705.07798.

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

Peters J, Mulling K, Altun Y (2010) Relative entropy policy search. Proc. AAAI Conf. Artificial Intelligence, 24(1):1607–1612.

Puterman ML (2014) *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. (John Wiley & Sons, Hoboken, NJ).

Schulman J, Chen X, Abbeel P (2017a). Equivalence between policy gradients and soft Q-Learn. Preprint, submitted April 21, https://arxiv.org/abs/1704.06440.

Schulman J, Levine S, Abbeel P, Jordan M, Moritz P (2015) Trust region policy optimization. *Proc. 32nd Conf. Machine Learn.*, PMLR, 37:1889–1897.

Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017b) Proximal policy optimization algorithms. Preprint, submitted July 20, https://arxiv.org/abs/1707.06347.

Published in Operations Research on December 02, 2021 as DOI: 10.1287/opre.2021.2151.
This article has not been copyedited or formatted. The final version may differ from this version.

Shani L, Efroni Y, Mannor S (2019) Adaptive trust region policy optimization: global convergence and faster rates for regularized MDPs. *Proc. AAAI Conf. Artificial Intelligence* 34(4):5668–5675.

Silver D, Huang A, Maddison CJ, 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 RS, McAllester DA, Singh SP, Mansour Y (2000) Policy gradient methods for reinforcement Learn. with function approximation. NIPS'99, 1057–1063.

Tu S, Recht B (2019) The gap between model-based and model-free methods on the linear quadratic regulator: an asymptotic viewpoint. *Proc. Thirty-Second Conf. Learn. Theory*, PMLR, 99: 3036–3083.

Vieillard N, Kozuno T, Scherrer B, Pietquin O, Munos R, Geist M (2020) Leverage the average: an analysis of KL regularization in RL. Preprint, submitted March 31, https://arxiv.org/abs/2003.14089.

Wang L, Cai Q, Yang Z, Wang Z (2019) Neural policy gradient methods: Global optimality and rates of convergence. Preprint, submitted August 29, https://arxiv.org/abs/1909.01150.

Williams RJ (1992) Simple statistical gradient-following algorithms for connectionist reinforcement Learn. *Machine Learn.* 8(3-4): 229–256.

Williams RJ, Peng J (1991) Function optimization using connectionist reinforcement Learn. algorithms. *Connect. Sci.* 3(3):241–268.

Wu Y, Zhang W, Xu P, Gu Q (2020) A finite time analysis of two time-scale actor critic methods. Preprint, submitted May 4, https://arxiv.org/abs/2005.01350.

Xiao C, Huang R, Mei J, Schuurmans D, Müller M (2019) Maximum entropy Monte-Carlo planning. *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY), 9520–9528.

Xu T, Wang Z, Liang Y (2020) Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. Preprint, submitted May 7, https://arxiv.org/abs/2005.03557.

Zhang K, Hu B, Basar T (2019a) Policy optimization for $\mathscr{H}_2$ linear control with $\mathscr{H}_\infty$ robustness guarantee: implicit regularization and global convergence. *Proc. 2nd Conf. Learn. Dynam. Control*, PMLR, 120:179–190.

Zhang K, Koppel A, Zhu H, Başar T (2019b). Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM J. Control Optim.* 58(6):3586–3612.

Shicong Cen is a second-year PhD student in the Department of Electrical and Computer Engineering of Carnegie Mellon University, advised by Professor Yuejie Chi. He received his bachelor's degree from School of Mathematical Sciences, Peking University. His research interests lie in the theoretical foundations of optimization methods in machine learning and reinforcement learning.

Chen Cheng is a second-year PhD student in the Department of Statistics at Stanford University, jointly advised by Professor John Duchi and Professor Andrea Montanari. He received his bachelor's degree from the School of Mathematical Sciences, Peking University. His research interests lie in statistical theory and algorithms for high-dimensional data, random matrix theory, and reinforcement learning.

Yuxin Chen is currently an assistant professor of electrical and computer engineering at Princeton University. His research interests include high-dimensional statistics, mathematical optimization, and reinforcement learning. He has received the AFOSR and ARO Young Investigator Awards, the Princeton graduate mentoring award, and the 2020 ICCM best paper award (gold medal), and was selected as a finalist for the Best Paper Prize for Young Researchers in Continuous Optimization, 2019.

Yuting Wei is currently an assistant professor in the Statistics and Data Science Department at the Wharton School, University of Pennsylvania. Prior to this, she spent two years at Carnegie Mellon University as an assistant professor, and one year at Stanford University as a Stein Fellow. She obtained her PhD in statistics at University of California at Berkeley, receiving the 2018 Erich L. Lehmann Citation for her PhD dissertation. Her research interests include high-dimensional statistics, machine learning, and reinforcement learning.

Yuejie Chi is a professor in the Department of Electrical and Computer Engineering at Carnegie Mellon University. Her research interests lie in the theoretical and algorithmic foundations of data science, signal processing, machine learning and inverse problems. Among others, Dr. Chi received the Presidential Early Career Award for Scientists and Engineers (PECASE), and the inaugural IEEE Signal Processing Society Early Career Technical Achievement Award for contributions to high-dimensional structured signal processing.