Theoretical Guarantees of Fictitious Discount Algorithms for Episodic Reinforcement Learning and Global Convergence of Policy Gradient Methods

Xin Guo,[1,2] **Anran Hu,** [1] **Junzi Zhang** [2*]

[1] University of California, Berkeley
[2] Amazon.com
xinguo@berkeley.edu, anran_hu@berkeley.edu, junziz@amazon.com

Abstract

When designing algorithms for finite-time-horizon episodic reinforcement learning problems, a common approach is to introduce a fictitious discount factor and use stationary policies for approximations. Empirically, it has been shown that the fictitious discount factor helps reduce variance, and stationary policies serve to save the per-iteration computational cost. Theoretically, however, there is no existing work on convergence analysis for algorithms with this fictitious discount recipe. This paper takes the first step towards analyzing these algorithms. It focuses on two vanilla policy gradient (VPG) variants: the first being a widely used variant with discounted advantage estimations (DAE), the second with an additional fictitious discount factor in the score functions of the policy gradient estimators. Non-asymptotic convergence guarantees are established for both algorithms, and the additional discount factor is shown to reduce the bias introduced in DAE and thus improve the algorithm convergence asymptotically. A key ingredient of our analysis is to connect three settings of Markov decision processes (MDPs): the finite-time-horizon, the average reward and the discounted settings. To our best knowledge, this is the first theoretical guarantee on fictitious discount algorithms for the episodic reinforcement learning of finite-time-horizon MDPs, which also leads to the (first) global convergence of policy gradient methods for finite-time-horizon episodic reinforcement learning.

1 Introduction

This paper studies episodic reinforcement learning with each episode consisting of a finite-time-horizon Markov decision process (MDP). For such finite-time-horizon episodic reinforcement learning problems, a popular heuristic approach is to introduce a fictitious discount factor and use stationary policies when designing algorithms; see for instance, the renowned DQN (Mnih et al. 2015), DDPG (Lillicrap et al. 2015), and recent works of (François-Lavet, Fonteneau, and Ernst 2015; Xu, van Hasselt, and Silver 2018; Burda et al. 2018; Hessel et al. 2018; Fedus et al. 2019; Tessler and Mannor 2020).

Empirically, it has been shown that discount factors serve to reduce variance (Thomas 2014; Haarnoja et al. 2017), and stationary policies help save per-iteration computational costs. Theoretically, fictitious discount algorithms designed for *average reward* MDPs have been analyzed (Marbach 1998; Marbach and Tsitsiklis 2001) and the *asymptotic* local convergence[1] has been established (Marbach and Tsitsiklis 2003).

It remains open, however, to establish the non-asymptotic global convergence for this fictitious-discount-factor approach in the *finite-time-horizon* framework. The major challenges are to characterize the *bias* introduced by the discount factor, and to close the gap between the *non-stationary* optimal policies for finite-time-horizon MDPs and the stationary algorithm policies.

This paper takes the first steps towards rigorously analyzing the global and non-asymptotic convergence of fictitious discount algorithms for finite-time-horizon episodic reinforcement learning. It focuses on the convergence analysis of two concrete algorithms in the context of policy gradient methods. The first one is a widely used variant of the vanilla policy gradient (VPG) method with discounted advantage estimations (DAE). This variant was originally proposed for average reward problems (Marbach 1998; Baxter and Bartlett 1999, 2001; Marbach and Tsitsiklis 2001), later extended to episodic deep reinforcement learning setting (Schulman et al. 2015b) and implemented in popular solvers such as Spinning Up (Achiam 2018). The second one is a new doubly discounted variant of VPG, with the introduction of an additional fictitious discount factor in the score functions of the policy gradient estimators. This additional discount factor is shown to help reduce the bias in DAE and thus improve asymptotically the algorithm convergence.

Our approach. There are three main ingredients in our analysis. The first is establishing quantitative connections among three settings of MDPs: the finite-time-horizon, the average award, and the discounted settings (*cf.* §2). These relations enable us to connect the finite-time-horizon sub-optimality gap with the average reward (*cf.* Theorem 14) and the discounted (*cf.* Theorem 18) ones. The second is utilizing the convergence property of value iteration algorithms to analyze the gap between the stationary policies of the average reward MDPs and the non-stationary optimal policies of

[*] Work done prior to joining or outside of Amazon.

[1] In this paper, "local convergence" indicates convergence to stationary points of value functions, and "global convergence" means convergence in terms of the value function sub-optimality gaps.

the finite-time-horizon MDPs (*cf.* Lemma 6). The third one is deriving the gradient domination (*cf.* Lemma 8) and Lipschitz gradient (*cf.* Lemma 10) properties for average reward MDPs, which is critical to obtain the sub-optimality of algorithm policies for the average reward problem (*cf.* Theorem 13).

Contributions. The contributions of this paper are two-fold:

- It establishes the first (and non-asymptotic) connections between (a) the sub-optimality gap in finite-time-horizon MDPs and (b) the sub-optimality gaps in the average reward and the discounted reformulations (*cf.* Theorems 14 and 18).
- It obtains, for the first time, theoretical guarantees on fictitious discount algorithms for the episodic reinforcement learning of finite-time-horizon MDPs (*cf.* Theorems 15 and 19). The convergence is global, and not asymptotic. Moreover, it demonstrates explicit dependencies on both the time horizon and the fictitious discount factor. The analysis in this paper leads to the first global convergence of policy gradient methods for finite-time-horizon episodic reinforcement learning.

Related work. Since the seminal work of D. Blackwell (Blackwell 1962), earlier works on the relationship among different settings of MDPs have been focusing on the discounted and average reward settings (Hordijk and Yushkevich 2002; Lasserre 1988; Kakade 2001a; Lewis and Puterman 2002; Mahadevan 1996; Schneckenreither 2020). In contrast, our focus is on the remaining two relations, namely (i) the connection between the finite-time-horizon and the discounted problems and (ii) the connection between the finite-time-horizon and the average reward problems.

Theoretical study on policy gradient methods started with the asymptotic local convergence (Sutton et al. 2000; Konda and Tsitsiklis 2003; Marbach and Tsitsiklis 2001). Later, non-asymptotic rate of such local convergence has been established in a series of works (Papini et al. 2018; Xu, Gao, and Gu 2019). Recently, more attention has been shifted to the global convergence of policy gradient methods. However, the majority of these results have been on the discounted settings (Zhang et al. 2019; Bhandari and Russo 2019; Agarwal et al. 2019; Wang et al. 2019; Shani, Efroni, and Mannor 2019; Mei et al. 2020; Cen et al. 2020; Zhang et al. 2020b). Recent progress has been made on a particular class of finite-time-horizon MDPs, i.e., linear quadratic finite-time-horizon MDPs and their variants (Hambly, Xu, and Yang 2020) (Zhang et al. 2021), (Hambly, Xu, and Yang 2021). This paper, instead, studies global convergence of policy gradient methods for finite-time-horizon, finite-state-action MDPs with *general dynamics and rewards*.

Outline. §2 introduces three settings of MDPs and their mutual connections. §3 introduces DAE REINFORCE and establishes its global sub-optimality guarantee. A doubly discounted variant is then proposed in §4 with its global convergence analysis, showing the benefits of the additional discount factor. §5 concludes.

2 Problem setup and preliminaries

2.1 Problem Setup

Consider a Markov decision process $\mathcal{M}$ with a finite state space $\mathcal{S} = \{1, \ldots, S\}$, a finite action space $\mathcal{A} = \{1, \ldots, A\}$, a transition probability $p(s'|s,a)$ for the probability of transitioning from state s to state s' when taking action a, and a reward function $r(s,a)$ denoting the (deterministic) instantaneous reward for taking action a in state s. Here, the initial state is assumed to follow a distribution $\rho \in \mathcal{P}(\mathcal{S})$, where $\mathcal{P}(\mathcal{S}) \subseteq \mathbf{R}^{|\mathcal{S}|}$ denotes the set of probability measures on over the set $\mathcal{S}$. Denote $R_{\max}$ the maximum reward such that $R_{\max} = \max_{s\in\mathcal{S}, a\in\mathcal{A}} |r(s,a)|$.

The focus of this paper is the finite-time-horizon MDP. Given a finite time horizon $H \geq 1$, decisions are made in the duration of timestamps from $h = 0$ to $h = H-1$. This duration is also referred to as an "episode". Such a horizon can either be naturally defined by the expiration time (*e.g.*, the length of a video game) or manually specified by the decision maker (*e.g.*, the length of affordable decision period). A (randomized) policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ is a mapping from the state space to a distribution over the action space. For notational simplicity, we use $\pi(a|s)$ to denote the a-th entry of $\pi(s)$, *i.e.*, the probability of taking action a at state s under a policy π. Then for any (randomized) policy sequence $\boldsymbol{\pi}^H = \{\pi_h\}_{h=0}^{H-1}$, the performance metric $V^H(\boldsymbol{\pi}^H)$ is the mean reward collected over the finite horizon episode of length H, *i.e.*,

$$V^H(\boldsymbol{\pi}^H) = \frac{1}{H}\mathbf{E}\sum\nolimits_{h=0}^{H-1} r(s_h, a_h), \tag{1}$$

where $s_0 \sim \rho$, $a_h \sim \pi_h(s_h)$ and $s_{h+1} \sim p(\cdot|s_h, a_h)$ for $h = 0, \ldots, H-2$. The finite-time-horizon problem is the following optimization problem:

$$\text{maximize}_{\boldsymbol{\pi}^H = \{\pi_0, \ldots, \pi_{H-1}\}}\, V^H(\boldsymbol{\pi}^H). \tag{2}$$

Note that the optimal policy sequence $\boldsymbol{\pi}^{H,\star} = \{\pi_h^{H,\star}\}_{h=0}^{H-1}$ of problem (2) may be nonstationary, and we write $V^{H,\star} = V^H(\boldsymbol{\pi}^{H,\star})$. When the policy sequence $\boldsymbol{\pi}^H = \{\pi\}_{h=0}^{H-1}$ is stationary, we will write it as π for notational simplicity. Here and below we use $P_\pi \in \mathbf{R}^{S\times S}$ to denote the transition probability of the Markov chain induced by policy π, *i.e.*, $P_\pi(s,s') = \sum_{a\in\mathcal{A}} p(s'|s,a)\pi(a|s)$.

Throughout this paper, we make the following assumption as in (Ortner 2020). Note that this assumption naturally holds when the transition probability p is component-wisely positive.

Assumption 1. *For any deterministic stationary policy π, the induced Markov chain with transition matrix P_π is irreducible and aperiodic.*

With Assumption 1, we have the following proposition.

Proposition 1. *Given Assumption 1, then there exist constants $C_{p,S,A} > 1$ and $\alpha_{p,S,A} \in [0,1)$ that depend only on the transition probability model p, number of states S and number of actions A of the MDP $\mathcal{M}$, such that for any policy π and $h \geq 0$,*

$$d_{\mathrm{TV}}(\rho P_\pi^h, \mu_\pi) \leq C_{p,S,A}\alpha_{p,S,A}^h, \tag{3}$$

where μ_π is the (unique) stationary distribution of the transition matrix P_π.

The analysis of the above finite-time-horizon MDP will rely on two related MDPs: the average reward problem and the discounted one, both of which have stationary optimal policies under Assumption 1.

Discounted problem. It is to consider an infinite horizon and solve for

$$\text{maximize}_{\boldsymbol{\pi}=\{\pi_h\}_{h=0}^\infty} V^\gamma(\boldsymbol{\pi})$$

with

$$V^\gamma(\boldsymbol{\pi}) = (1-\gamma)\mathbf{E}\sum\nolimits_{h=0}^\infty \gamma^h r(s_h, a_h),$$

where $s_0 \sim \rho$, $a_h \sim \pi_h(s_h)$ and $s_{h+1} \sim p(\cdot|s_h, a_h)$ for $h \geq 0$. Here $\gamma \in [0,1)$ is the discount factor, penalizing future rewards. It is well-known that for this discounted problem, there exists a stationary optimal policy sequence $\boldsymbol{\pi}^{\gamma,\star} = \{\pi_h^{\gamma,\star}\}_{h=0}^\infty$, where all $\pi_h^{\gamma,\star} = \pi^{\gamma,\star}$ ($h \geq 0$) are equal (Puterman 2014). Similarly, we denote $V^{\gamma,\star} = V^\gamma(\boldsymbol{\pi}^{\gamma,\star})$. Again, when the policy sequence $\boldsymbol{\pi} = \{\pi\}_{h=0}^\infty$ is stationary, we will write it as π for notational simplicity.

Average reward problem. The infinite horizon average reward of a (stationary) policy π is defined as

$$\begin{aligned}\eta(\pi) &= \lim_{H\to\infty} V^H(\pi) = \lim_{H\to\infty} \frac{1}{H}\mathbf{E}\sum_{h=0}^{H-1} r(s_h, a_h) \\ &= \sum_{s\in\mathcal{S},\, a\in\mathcal{A}} \mu_\pi(s)\pi(a|s)r(s,a),\end{aligned} \tag{4}$$

where μ_π is defined in Proposition 1. The goal is to find π that maximizes $\eta(\cdot)$. Note $\eta(\pi)$ is well-defined as the limit in (4) is guaranteed to exist and be finite, and independent of the initial state distribution ρ under Assumption 1 (Puterman 2014). Since $|\eta(\pi)| \leq R_{\max}$ and the set of all (stationary) policies (viewed as a subset $\mathbf{R}^{SA}$) is compact, the optimal (stationary) policy $\pi^\star$ (that maximizes $\eta(\cdot)$) exists and we denote the corresponding value function as $\eta^\star = \eta(\pi^\star)$.

2.2 Connections of finite-time-horizon with discounted and average reward problems

Now we introduce our first set of main results, which characterize the connections within these three different MDP problems.

The first result bounds the error between $V^\gamma(\pi)$ (for the discounted problem) and $V^H(\pi)$ (for the finite-time-horizon problem) under an arbitrary stationary policy π.

Lemma 2. *Given Assumption 1, then for any stationary policy π,*

$$\begin{aligned}|V^\gamma(\pi) - V^H(\pi)| \leq 2R_{\max}C_{p,S,A}\Bigg(&\frac{\gamma}{H(1-\gamma)}\alpha_{p,S,A}^H \\ &+\frac{\alpha_{p,S,A} + |H(1-\gamma)-1|}{(1-\alpha_{p,S,A})H}\Bigg),\end{aligned} \tag{5}$$

where $C_{p,S,A} > 1$ and $\alpha_{p,S,A} \in [0,1)$ are the constants in Proposition 1, and depend only on the transition probability model p, the number of states S and the number of actions A of $\mathcal{M}$, the underlying MDP.

The next lemma establishes a bound between $V^\gamma(\pi)$ (for the discounted problem) and $\eta(\pi)$ (for the average reward problem) under any stationary policy π.

Lemma 3. *Given Assumption 1, then*

$$|V^\gamma(\pi) - \eta(\pi)| \leq \frac{2(1-\gamma)R_{\max}C_{p,S,A}}{1-\alpha_{p,S,A}}, \tag{6}$$

where the constants $C_{p,S,A} > 1$ and $\alpha_{p,S,A} \in [0,1)$ are the same as in Lemma 2.

Maximizing over π, then immediately from Lemma 3, we have

Corollary 4. *Given Assumption 1, then*

$$|V^{\gamma,\star} - \eta^\star| \leq \frac{2(1-\gamma)R_{\max}C_{p,S,A}}{1-\alpha_{p,S,A}}, \tag{7}$$

where the constants $C_{p,S,A} > 1$ and $\alpha_{p,S,A} \in [0,1)$ are the same as in Lemma 2.

The following statement controls the gap between $V^H(\pi)$ (for the finite-time-horizon problem) and $\eta(\pi)$ (for the average reward problem) under any stationary policy π.

Lemma 5. *Given Assumption 1, then*

$$|V^H(\pi) - \eta(\pi)| \leq \frac{2R_{\max}C_{p,S,A}}{H(1-\alpha_{p,S,A})}, \tag{8}$$

where the constants $C_{p,S,A} > 1$ and $\alpha_{p,S,A} \in [0,1)$ are the same as in Lemma 2.

And finally, the bound of the gap between the optimal value functions $V^{H,\star}$ (for the finite-time-horizon problem) and $\eta^\star$ (for the average reward problem) is as follows.

Lemma 6. *Given Assumption 1, then*

$$|V^{H,\star} - \eta^\star| \leq \frac{2R_{\max}D_{p,S,A}}{H}, \tag{9}$$

where $D_{p,S,A} > 1$ is a constant that depends only on the transition probability model p, the number of states S and the number of actions A of the underlying MDP $\mathcal{M}$.

Remark 1. *Lemma 6 cannot be directly implied by Lemma 5. The key issue is that the optimal policy for the average reward value function $\eta(\cdot)$ is stationary, while the optimal policy for the finite-horizon value function $V^H(\cdot)$ may be non-stationary. To bridge this gap between stationary and non-stationary policies, we need the convergence property of value iteration algorithms (cf. Appendix A.2).*

These properties show that the three different settings are closely related for a large horizon H, and are critical for the subsequent analyses.

2.3 Gradient properties

In this section, we review the basics of policy gradient methods and state some useful properties of policy gradients in the average reward and the discounted settings.

Policy gradient methods. Policy gradient methods start by parametrizing the policy with parameter $\theta \in \Theta$, which we denote as π_θ. Here Θ is the parameter space and the parametrization maps θ to a randomized policy $\pi_\theta : \mathcal{S} \to \mathcal{P}(\mathcal{A})$. The (vanilla) policy gradient (VPG) methods then proceed by performing stochastic gradient ascent on a (regularized) value function in the parameter space, namely, for each iteration k, θ^k is updated to θ^{k+1} with

$$\theta^{k+1} = \theta^k + \alpha^k g_k. \tag{10}$$

Here θ^0 is the initial parameter, α^k is the step-size, and g_k is a (possibly biased) stochastic gradient estimator of a regularized value function.

Throughout this paper, we will focus on the following regularized value function of the average reward problem:

$$\bar{L}(\theta) = \eta(\pi_\theta) + \Omega(\theta),$$

and the regularized value function of the discounted problem:

$$L^\gamma(\theta) = \frac{1}{1-\gamma} V^\gamma(\pi_\theta) + \Omega(\theta).$$

Here $\Omega : \Theta \to \mathbf{R}$ is a regularization term that serves to improve the convergence (Zhao et al. 2016; Mnih et al. 2016; Henkel 2018).

Below we specify additional assumptions about the problem setting. Note that the same set of assumptions have been made in (Agarwal et al. 2019; Zhang et al. 2020a).

Assumption 2. *(Setting)*

- *The policy is a soft-max policy parameterization, i.e.,* $\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}$, *with* $\Theta = \mathbf{R}^{SA}$.
- *The regularization term is (with* $\lambda > 0$*)*

$$\Omega(\theta) = \frac{\lambda}{SA} \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \log \pi_\theta(a|s).$$

- *The initial distribution is component-wisely positive, i.e.,* $\rho(s) > 0$ *for any* $s \in \mathcal{S}$.
- *The reward function* $r(s,a) \in [0,1]$, $\forall\, s \in \mathcal{S}$, $a \in \mathcal{A}$.

Some remarks on Assumption 2:

- The soft-max policy parametrization is simple yet forms the basis of the widely-used (neural network) energy based policies (Haarnoja et al. 2017).
- The regularization term is a simplified version of the popular (relative) entropy regularization terms (Peters, Mülling, and Altun 2010; Schulman, Chen, and Abbeel 2017), and has been demonstrated to be necessary to avoid exponential lower bounds when working with the soft-max policy parametrization in (Li et al. 2021).
- The positivity assumption on the initial distribution is standard in the global convergence literature of policy gradient methods (Agarwal et al. 2019; Bhandari and Russo 2019; Mei et al. 2020).
- The last assumption on the range of r is merely for the simplicity of the subsequent discussions and can be easily relaxed to the general constant bound $r(s,a) \in [-R_{\max}, R_{\max}]$, $\forall\, s \in \mathcal{S}$, $a \in \mathcal{A}$.

Properties of policy gradients. We are now ready to provide some useful properties regarding the gradients of the discounted and the average reward problems.

We first slightly tighten the gradient domination property established in (Agarwal et al. 2019, Theorem 5.2) for the discounted problems by utilizing the uniform ergodic property in Assumption 1.

Proposition 7. (Gradient domination for discounted problems) *Given Assumptions 1 and 2, suppose that* $\|\nabla_\theta L^\gamma(\theta)\|_2 \le \lambda/(2SA)$. *Then*

$$V^{\gamma,\star} - V^\gamma(\pi_\theta) \le 2\lambda \min\left\{ \left\| \frac{d_\rho^{\pi^{\gamma,\star}}}{\rho} \right\|_\infty, \frac{S \|d_\rho^{\pi^{\gamma,\star}}\|_\infty}{(1-\gamma)(1-\alpha_{p,S,A})} \right\}.$$

Here for any (randomized) policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$,

$$d_\rho^\pi(s) = (1-\gamma) \sum_{t=0}^\infty \gamma^t \mathbf{Prob}_\rho^\pi(s_t = s)$$

is the discounted state visitation distribution, where $\mathbf{Prob}_\rho^\pi(s_t = s)$ is the probability of arriving at s in step t starting from $s_0 \sim \rho$ following policy π in $\mathcal{M}$. In addition, the division in $d_\rho^{\pi^\star}/\rho$ is component-wise.

We next establish analogously the gradient domination property for the average reward problem.

Lemma 8. (Gradient domination for average reward problems) *Given Assumptions 1 and 2, suppose that* $\|\nabla_\theta \bar{L}(\theta)\|_2 \le \lambda/(2SA)$. *Then*

$$\eta^\star - \eta(\pi_\theta) \le \lambda \frac{S \|\mu_{\pi^\star}\|_\infty}{1 - \alpha_{p,S,A}},$$

where $\mu_{\pi^\star}$ *and* $\alpha_{p,S,A}$ *are defined as in Proposition 1.*

The two statements above on gradient domination capture the sub-optimality results for policies satisfying certain gradient conditions.

Now recall the strongly smoothness property of the objectives for discounted problems (Agarwal et al. 2019).

Proposition 9. (Strongly smoothness for discounted problems (Agarwal et al. 2019, Lemma D.4)) *Given Assumptions 1 and 2,* L^γ *is strongly smooth with parameter* $\beta_\lambda = \frac{8}{(1-\gamma)^3} + \frac{2\lambda}{S}$, *i.e.,*

$$\|\nabla_\theta L^\gamma(\theta_1) - \nabla_\theta L^\gamma(\theta_2)\|_2 \le \beta_\lambda \|\theta_1 - \theta_2\|_2$$

for any $\theta_1, \theta_2 \in \Theta$.

We can establish analogously the strongly smoothness property for the average reward problem.

Lemma 10. (Strongly smoothness for average reward problems) *Under Assumptions 1 and 2,* $\bar{L}$ *is strongly smooth with parameter* $\bar{\beta}_\lambda = 22\sqrt{S}\left(\frac{2C_{p,S,A}}{1-\alpha_{p,S,A}} + 1\right)^3 + 2\lambda/S$, *i.e.,*

$$\|\nabla_\theta \bar{L}(\theta_1) - \nabla_\theta \bar{L}(\theta_2)\|_2 \le \bar{\beta}_\lambda \|\theta_1 - \theta_2\|_2,$$

for any $\theta_1, \theta_2 \in \Theta$. *Here the constants* $C_{p,S,A} > 1$ *and* $\alpha_{p,S,A} \in [0,1)$ *are defined as in Proposition 1.*

These two statements are critical for the subsequent analyses of the algorithms.

3 DAE REINFORCE algorithm

In this section, we first introduce a widely used vanilla policy gradient implementation (Achiam 2018), which we call the DAE REINFORCE algorithm (following its usage of DAE in (Schulman et al. 2015b)). In DAE REINFORCE, a stationary parametrized policy $\pi_\theta(a|s)$ is considered, and the parameter is updated by

$$\theta^{k+1} = \theta^k + \alpha^k \hat{g}_k, \tag{11}$$

where

$$\hat{g}_k = \frac{1}{NH}\sum_{i=1}^{N}\sum_{h=0}^{H-1} \nabla_\theta \log \pi_{\theta^k}(a_h^i|s_h^i) \times \overbrace{\left(\sum_{h'=h}^{H-1}\gamma^{h'-h} r_{h'}^i - b(s_h^i)\right)}^{\text{advantage function}} + \nabla_\theta \Omega(\theta^k). \tag{12}$$

Here $\gamma \in (0,1)$ is a fictitious discount factor, N is the mini-batch size of the updates, $r_h^i = r(s_h^i, a_h^i)$, $\tau_i = (s_0^i, a_0^i, r_0^i, \ldots, s_{H-1}^i, a_{H-1}^i, r_{H-1}^i)$ $(i = 1, \ldots, N,\ h = 0, \ldots, H-1)$ are i.i.d. trajectories sampled under policy π_{θ^k}, and b is a baseline function that is independent of the trajectories. Throughout the paper, we assume that the baseline b is a.s. uniformly bounded, *i.e.*, $\max_{s\in\mathcal{S}} |b(s)| \le B$ a.s. for some constant $B > 0$.

In the rest of the section, we establish the convergence of (a slightly modified version of) DAE REINFORCE, which we call Truncated DAE REINFORCE and summarize in Algorithm 1. Note that the estimator $\hat{g}_k$ is truncated in (13) (and for notational simplicity under the same symbol) with a truncation parameter $\beta \in (0,1)$. The same truncation has been adopted for studying the standard REINFORCE algorithm (without DAE) in (Zhang et al. 2020a), where β is introduced to ensure that the advantage function estimation is sufficiently accurate.[2]

Algorithm 1: Truncated DAE REINFORCE

1: **Input:** Initialization θ^0, step-sizes α^k for $k \ge 0$.
2: **for** $k = 0, 1, \ldots$ **do**
3: Sample N i.i.d. trajectories $\{\tau_i\}_{i=1}^N$ under policy π_{θ^k}.
4: Compute gradient estimator $\hat{g}_k$ as

$$\hat{g}_k = \frac{1}{N\lfloor \beta H\rfloor}\sum_{i=1}^{N}\sum_{h=0}^{\lfloor\beta H\rfloor-1} \nabla_\theta \log \pi_{\theta^k}(a_h^i|s_h^i) \times \left(\sum_{h'=h}^{H-1}\gamma^{h'-h} r_{h'}^i - b(s_h^i)\right) + \nabla_\theta \Omega(\theta^k). \tag{13}$$

5: Update $\theta^{k+1} = \theta^k + \alpha^k \hat{g}_k$.
6: **end for**

The main idea behind our convergence analysis is to use the average reward as a bridge to connect the original finite-time-horizon MDP and the DAE REINFORCE algorithm. The proof consists of two parts. The first part is to establish the sub-optimality of θ_k, evaluated for the average reward problem. The second part is to establish the convergence of the algorithm for the finite-horizon problem by utilizing the connection between the average reward setting and the finite-horizon setting.

We begin the analysis by estimating the (upper) bound on the difference between the exact gradient and the sample gradient. Hereafter, we use $\mathbf{E}_k$ to denote the conditional expectation given the k-th iteration θ^k.

Lemma 11. *Given Assumptions 1 and 2, then*

$$\begin{aligned}\left\|\mathbf{E}_k[\hat{g}_k] - \nabla \bar{L}(\theta^k)\right\|_2 &\le \frac{16 C_{p,S,A}}{\lfloor \beta H\rfloor (1-\alpha_{p,S,A})}\left(1 + \frac{C_{p,S,A}}{1-\alpha_{p,S,A}}\right) \\ &\quad + 8 C_{p,S,A}\frac{1-\gamma}{(1-\alpha_{p,S,A})^2} \\ &\quad + 4\gamma^{(1-\beta)H}\left(1 + \frac{C_{p,S,A}}{1-\alpha_{p,S,A}}\right).\end{aligned} \tag{14}$$

Here the constants $C_{p,S,A} > 1$ and $\alpha_{p,S,A} \in [0,1)$ are defined in Proposition 1.

This lemma leads to the following bounds on the stochastic gradients, which are key to establishing the convergence of Algorithm 1.

Lemma 12. *Given Assumptions 1 and 2, then*

$$\begin{aligned}\|\hat{g}_k\|_2 &\le G^\gamma + 2\lambda \quad a.s., \\ \mathbf{E}_k \hat{g}_k^T \nabla_\theta \bar{L}(\theta^k) &\ge \|\nabla_\theta L^\gamma(\theta^k)\|_2^2 - (\bar{G} + 2\lambda)\bar{\Delta}, \\ \mathbf{E}_k \|\hat{g}_k\|_2^2 &\le 2\|\nabla_\theta \bar{L}(\theta^k)\|_2^2 + \bar{M}.\end{aligned}$$

Here $G^\gamma = \frac{2(1+(1-\gamma)B)}{1-\gamma}$, $\bar{G} = 4\left(1 + \frac{C_{p,S,A}}{1-\alpha_{p,S,A}}\right)$, $\bar{M} = 2\bar{\Delta}^2 + (G^\gamma + 2\lambda)^2/N$, $\bar{\Delta}$ is the right-hand side of (14)*, the constants $C_{p,S,A} > 1$ and $\alpha_{p,S,A} \in [0,1)$ are defined in Proposition 1.*

Remark 2. *The second bound in Lemma 12 shows that $\hat{g}_k$ is nearly unbiased, while the third bound shows that $\hat{g}_k$ satisfies a bounded second-order moment growth condition. These conditions slightly generalize the standard ones used in analyzing stochastic gradient methods (Bottou, Curtis, and Nocedal 2018).*

Now, we obtain first the sub-optimality behavior of θ^k from the Truncated DAE REINFORCE algorithm (*cf.* Algorithm 1) in the average reward setting.

Theorem 13. *Given Assumptions 1 and 2, let $\bar{\beta}_\lambda = 22\sqrt{S}\left(\frac{2C_{p,S,A}}{1-\alpha_{p,S,A}} + 1\right)^3 + 2\lambda/S$. For a fixed $\beta \in (0,1)$ and any $\epsilon > 0$, $\delta \in (0,1)$, set $\alpha^k = \frac{1}{2\bar{\beta}_\lambda}\frac{1}{\sqrt{k+3}\log_2(k+3)}$ and λ is the positive (larger) root of the following quadratic equation:*

$$2(\bar{G} + 2\lambda)\bar{\Delta} = (\lambda - \epsilon)^2/(4S^2A^2),$$

where $\bar{G}$ and $\bar{\Delta}$ are defined as in Lemma 12. Then

$$\begin{aligned}\min_{k=0,\ldots,K} \eta^\star - \eta(\pi_{\theta^k}) \le &\frac{\|\mu_{\pi^\star}\|_\infty}{1-\alpha_{p,S,A}}\left(S\epsilon + 4S^3A^2\bar{\Delta}\right. \\ &\left. + 4S^2A\sqrt{\bar{\Delta}\epsilon + 4S^2A^2\bar{\Delta}^2 + \bar{G}\bar{\Delta}}\right)\end{aligned} \tag{15}$$

[2] In §4, we show that β can be dropped if an additional discount factor is introduced in the gradient estimator.

with probability at least $1-\delta$, for any K such that

$$K \geq O\left(\frac{S^4A^4\bar{\beta}_\lambda^2(\bar{D}+\sqrt{2\bar{C}_\gamma\log(2/\delta)})^2}{\epsilon^4} \times \log^2\left(\frac{SA\bar{\beta}_\lambda(\bar{D}+\sqrt{2\bar{C}_\gamma\log(2/\delta)})}{\epsilon}\right)\right). \tag{16}$$

Here the constants $C_{p,S,A} > 1$ and $\alpha_{p,S,A} \in [0,1)$ are defined in Proposition 1, and the constants $\bar{D}$ and $\bar{C}_\gamma$ are bounded by

$$\bar{D} = O(\bar{M}+\lambda+1),$$
$$\bar{C}_\gamma = O\left(\frac{(G^\gamma+2\lambda)^2}{S}\left(\frac{C_{p,S,A}^2}{(1-\alpha_{p,S,A})^2}+\lambda^2+(G^\gamma+2\lambda)^2\right)\right), \tag{17}$$

where the constants hidden in the big-O notation may depend on θ^0.

Next, by Lemma 5 and Lemma 6, we have the following theorem.

Theorem 14. *Given Assumption 1, for any $H \geq 1$, if there exists a policy $\hat{\pi}$ such that $|\eta^\star - \eta(\hat{\pi})| \leq \epsilon$ for some $\epsilon > 0$, then*

$$V^{H,\star} - V^H(\hat{\pi}) \leq \frac{2R_{\max}D_{p,S,A}}{H} + \epsilon + \frac{2R_{\max}C_{p,S,A}}{H(1-\alpha_{p,S,A})}. \tag{18}$$

Here the constants $C_{p,S,A} > 1$, $D_{p,S,A} > 1$ and $\alpha_{p,S,A} \in [0,1)$ are the constants in Proposition 1 and Lemma 6, which depend only on the transition probability model p, the number of states S and the number of actions A of the underlying MDP $\mathcal{M}$.

Combining Theorems 13 and 14 we can derive the convergence for Truncated DAE REINFORCE algorithm.

Theorem 15. *Given Assumptions 1 and 2, let $\gamma = 1-H^{-\sigma}$ for some $\sigma \in (0,1)$. For a fixed $\beta \in (0,1)$ and any $\epsilon > 0$, $\delta \in (0,1)$, set λ, $\bar{\beta}_\lambda$ and α^k to be the same as in Theorem 13. Then for any K such that* (16) *is satisfied,* [3] *with probability at least $1-\delta$,*

$$\min_{k=0,\ldots,K} V^{H,\star} - V^H(\pi_{\theta^k}) \leq O\left(\frac{S}{1-\alpha_{p,S,A}}\epsilon\right) + \mathbf{bias}_H^{\mathrm{DAE}}, \tag{19}$$

where

$$\mathbf{bias}_H^{\mathrm{DAE}} = O\left(\frac{S^2AC_{p,S,A}^3}{(1-\alpha_{p,S,A})^4}H^{-\sigma/2} + \frac{S^3A^2C_{p,S,A}^2}{(1-\alpha_{p,S,A})^3}H^{-\sigma} + \left(D_{p,S,A}+\frac{C_{p,S,A}}{1-\alpha_{p,S,A}}\right)H^{-1}\right). \tag{20}$$

Here $C_{p,S,A} > 1$, $D_{p,S,A} > 1$ and $\alpha_{p,S,A} \in [0,1)$ are the constants in Proposition 1 and Lemma 6.

The choice of γ is for ease of presentation. See also (Liu and Su 2020; Dong, Van Roy, and Zhou 2021).

[3] See Appendix B.4 for more explicit bounds on the constants involved in (16).

4 Doubly Discounted REINFORCE algorithm

In Algorithm 1, a fictitious discount factor is introduced when computing advantage function estimates, while for the rest part it remains undiscounted. This introduces a bias term $\mathbf{bias}_H^{\mathrm{DAE}}$ as shown in Theorem 15, which remains nonzero for a fixed planning horizon H even when the number of iterations K goes to infinity and ϵ goes to 0. In this section, we propose the Doubly Discounted REINFORCE algorithm (*cf.* Algorithm 2) to reduce the bias introduced by DAE.

Algorithm 2: Doubly Discounted REINFORCE

1: **Input:** Initialization θ^0, step-sizes α^k for $k \geq 0$.
2: **for** $k = 0, 1, \ldots$ **do**
3: Sample N i.i.d. trajectories $\{\tau_i\}_{i=1}^N$ under policy π_{θ^k}.
4: Compute gradient estimator $\tilde{g}_k$ as

$$\tilde{g}_k = \frac{1}{N}\sum_{i=1}^N\sum_{h=0}^{H-1}\gamma^h\nabla_\theta\log\pi_{\theta^k}(a_h^i|s_h^i) \times\left(\sum_{h'=h}^{H-1}\gamma^{h'-h}r_{h'}^i - b(s_h^i)\right) + \nabla_\theta\Omega(\theta^k). \tag{21}$$

5: Update $\theta^{k+1} = \theta^k + \alpha^k\tilde{g}_k$.
6: **end for**

Compared with Algorithm 1, Algorithm 2 introduces an additional discount factor when computing the score functions and gets rid of the artificial parameter $\beta \in (0,1)$ needed in Truncated DAE REINFORCE. As a result, the estimator (21) coincides with the vanilla policy gradient estimator for solving discounted problems (Zhang et al. 2020a) with a fixed-length trajectory truncation (Liu et al. 2020). Note that a similar observation has been made for natural actor-critic methods in (Thomas 2014).

Similar to the idea of §3, we first establish the suboptimality of the Doubly Discounted REINFORCE algorithm, evaluated for the discounted problem. Parallel to Lemma 12, we have the following stochastic gradient bounds.

Lemma 16. *Given Assumptions 1 and 2, then*

$$\|\tilde{g}_k\|_2 \leq G+2\lambda \quad a.s.,$$
$$\mathbf{E}_k\tilde{g}_k^T\nabla_\theta L^\gamma(\theta^k) \geq \|\nabla_\theta L^\gamma(\theta^k)\|_2^2 - (G+2\lambda)\Delta,$$
$$\mathbf{E}_k\|\tilde{g}_k\|_2^2 \leq 2\|\nabla_\theta L^\gamma(\theta^k)\|_2^2 + M.$$

Here $G = \frac{2(1+B(1-\gamma))}{(1-\gamma)^2}$, and the constants Δ and M are defined by

$$\Delta = 2\frac{\gamma^H}{1-\gamma}\left(H+\frac{1}{1-\gamma}\right), \quad M = 2\Delta^2 + (G+2\lambda)^2/N.$$

Based on the above conditions, we now establish the suboptimality of θ^k from the Doubly Discounted REINFORCE algorithm for the discounted problem.

Theorem 17. *Given Assumptions 1 and 2, let $\beta_\lambda = 8/(1-\gamma)^3 + 2\lambda/S$. For any $\epsilon > 0$ and $\delta \in (0,1)$, set $\alpha^k =$*

$\frac{1}{2\beta_\lambda}\frac{1}{\sqrt{k+3}\log_2(k+3)}$ *and* λ *to be the positive (larger) root of the following quadratic equation:*

$$2(G+2\lambda)\Delta = (\lambda-\epsilon)^2/(4S^2A^2).$$

Then

$$\begin{aligned}&\min_{k=0,\ldots,K} V^{\gamma,\star} - V^{\gamma}(\pi_{\theta^k})\\ &\leq \min\left\{\left\|\frac{d_\rho^{\pi^{\gamma,\star}}}{\rho}\right\|_\infty, \frac{S\|d_\rho^{\pi^{\gamma,\star}}\|_\infty}{(1-\gamma)(1-\alpha_{p,S,A})}\right\} \\ &\times (2\epsilon + 8S^2A^2\Delta + 8SA\sqrt{\Delta\epsilon + 4S^2A^2\Delta^2 + G\Delta})\end{aligned} \tag{22}$$

with probability at least $1-\delta$*, for any* K *such that*

$$\begin{aligned}K \geq O\Bigg(&\frac{S^4A^4\beta_\lambda^2(D+\sqrt{2C\log(2/\delta)})^2}{\epsilon^4}\\ &\times \log^2\left(\frac{SA\beta_\lambda(D+\sqrt{2C\log(2/\delta)})}{\epsilon}\right)\Bigg).\end{aligned} \tag{23}$$

Here the constant $\alpha_{p,S,A} \in (0,1)$ *is defined in Proposition 1, and the constants* D *and* C *are bounded by*

$$\begin{aligned}&D = O(M + 1/(1-\gamma) + \lambda),\\ &C = O((G+2\lambda)^2(1/(1-\gamma)^4 + \lambda^2 + (G+2\lambda)^2)),\end{aligned} \tag{24}$$

where the constants hidden in the big-O notation may depend on θ^0*.*

The next result is parallel to Theorem 14, and is based on Lemma 2, Corollary 4, and Lemma 6.

Theorem 18. *Given Assumption 1, if there exists a policy* $\hat{\pi}$ *such that* $V^{\gamma,\star} - V^{\gamma}(\hat{\pi}) \leq \epsilon$ *for some* $\epsilon > 0$*, then for any* $H \geq 1$*,*

$$\begin{aligned}V^{H,\star} - V^H(\hat{\pi}) \leq\ & 2R_{\max}C_{p,S,A}\frac{\gamma}{H(1-\gamma)}\alpha_{p,S,A}^H + \epsilon\\ &+ \frac{2R_{\max}}{H}\left(\frac{C_{p,S,A}(H(1-\gamma)+\alpha_{p,S,A})}{1-\alpha_{p,S,A}} + D_{p,S,A}\right),\end{aligned} \tag{25}$$

where $C_{p,S,A} > 1$*,* $D_{p,S,A} > 1$ *and* $\alpha_{p,S,A} \in [0,1)$ *are the constants in Proposition 1 and Lemma 6, which depend only on the transition probability model* p*, the number of states* S *and the number of actions* A *of the underlying MDP* $\mathcal{M}$*.*

Combining Theorems 17 and 18, we obtain the final convergence result for the Doubly Discounted REINFORCE algorithm (in parallel to Theorem 15).

Theorem 19. *Given Assumptions 1 and 2, let* $\gamma = 1 - H^{-\sigma}$ *for some* $\sigma \in (0,1)$*. For any* $\epsilon > 0$*,* $\delta \in (0,1)$*, set* λ*,* β_λ *and* α^k *to be the same as in Theorem 17. Then for any* K *such that* (23) *is satisfied,* [4] *with probability at least* $1-\delta$*,*

$$\begin{aligned}&\min_{k=0,\ldots,K} V^{H,\star} - V^H(\pi_{\theta^k})\\ &\leq O\left(\epsilon\min\left\{\left\|\frac{1}{\rho}\right\|_\infty, \frac{SH^\sigma}{1-\alpha_{p,S,A}}\right\}\right) + \textbf{bias}_H^{\mathrm{DD}},\end{aligned} \tag{26}$$

[4] See Appendix C.2 for more explicit bounds on the constants involved in (23).

where

$$\begin{aligned}\textbf{bias}_H^{\mathrm{DD}} = O\Bigg(&\frac{C_{p,S,A}}{1-\alpha_{p,S,A}}H^{-\sigma} + D_{p,S,A}H^{-1}\\ &+\frac{S^3A^2}{1-\alpha_{p,S,A}}H^{\frac{1+5\sigma}{2}}e^{-H^{1-\sigma}/2}\\ &+C_{p,S,A}\alpha_{p,S,A}^H H^{-(1-\sigma)}\Bigg).\end{aligned} \tag{27}$$

Here $C_{p,S,A} > 1$*,* $D_{p,S,A} > 1$ *and* $\alpha_{p,S,A} \in [0,1)$ *are the constants in Proposition 1 and Lemma 6.*

Comparison with DAE REINFORCE. Here we compare the convergence of (truncated) DAE REINFORCE (*cf.* Algorithm 1) and Doubly Discounted REINFORCE (*cf.* Algorithm 2). Note that in both (19) and (26), the global suboptimality bounds consist of two parts: a vanishing ϵ term that goes to zero as the number of iterations K goes to infinity and a remaining bias term ($\textbf{bias}_H^{\mathrm{DAE}}$ and $\textbf{bias}_H^{\mathrm{DD}}$, respectively) resulting from the fictitious discount factor. Below we focus on comparing the bias terms with the same fictitious discount factor $\gamma = 1 - H^{-\sigma}$, with $\sigma \in (0,1)$. Recall that

$$\textbf{bias}_H^{\mathrm{DAE}} = O\left(\frac{S^2AC_{p,S,A}^3}{(1-\alpha_{p,S,A})^4}H^{-\frac{\sigma}{2}}\right) + \textit{lower order terms},$$

$$\textbf{bias}_H^{\mathrm{DD}} = O\left(\frac{C_{p,S,A}}{1-\alpha_{p,S,A}}H^{-\sigma}\right) + \textit{lower order terms}.$$

Comparing the above two bounds, we see the power of the additional discounting. Indeed, with further discounting, Doubly Discounted REINFORCE improves over DAE REINFORCE, especially in terms of H (from $H^{-\sigma/2}$ to $H^{-\sigma}$) as it grows. More precisely, the constant before the $H^{-\sigma}$ term is improved from $O(S^3A^2C_{p,S,A}^2/(1-\alpha_{p,S,A})^3)$ to $O(C_{p,S,A}/(1-\alpha_{p,S,A}))$, the constant before the H^{-1} term is improved from $O(D_{p,S,A} + C_{p,S,A}/(1-\alpha_{p,S,A})$ to $O(D_{p,S,A})$, while the $H^{-\sigma/2}$ term is improved to be exponentially decaying as H grows.

5 Conclusion and extensions

This paper focuses on two concrete fictitious discount algorithms in the context of policy gradient methods, namely DAE REINFORCE and Doubly Discounted REINFORCE. Rigorous convergence analyses are established for the two algorithms, which, for the first time, shed light on the non-asymptotic global convergence of fictitious discount algorithms.

Given recent development in (global) convergence analysis of algorithms in the discounted setting (Agarwal et al. 2019; Wang et al. 2019; Shani, Efroni, and Mannor 2019) and in the average reward framework (Neu, Jonsson, and Gómez 2017; Abbasi-Yadkori et al. 2019), it is natural to extend our study for natural policy gradient (Kakade 2001b), natural actor-critic (Peters and Schaal 2008), TRPO (Schulman et al. 2015a), PPO (Schulman et al. 2017), as well as deep learning based algorithms such as DQN (Mnih et al. 2015) and DDPG (Lillicrap et al. 2015).

Meanwhile, it remains to see if one can generalize our work to more general state and action spaces, and to remove the need for an exploratory initial distribution (*i.e.*, $\rho > 0$ component-wisely).

References

Abbasi-Yadkori, Y.; Bartlett, P.; Bhatia, K.; Lazic, N.; Szepesvari, C.; and Weisz, G. 2019. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, 3692–3702.

Achiam, J. 2018. OpenAI Spinning Up: Vanilla Policy Gradient.

Agarwal, A.; Jiang, N.; and Kakade, S. 2019. Reinforcement Learning: Theory and Algorithms. Technical report, Department of Computer Science, University of Washington.

Agarwal, A.; Kakade, S. M.; Lee, J. D.; and Mahajan, G. 2019. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv preprint arXiv:1908.00261*.

Baxter, J.; and Bartlett, P. 2001. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15: 319–350.

Baxter, J.; and Bartlett, P. L. 1999. Direct gradient-based reinforcement learning: I. gradient estimation algorithms. Technical report, Citeseer.

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

Blackwell, D. 1962. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 719–726.

Bottou, L.; Curtis, F.; and Nocedal, J. 2018. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2): 223–311.

Burda, Y.; Edwards, H.; Storkey, A.; and Klimov, O. 2018. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*.

Cen, S.; Cheng, C.; Chen, Y.; Wei, Y.; and Chi, Y. 2020. Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization. *arXiv preprint arXiv:2007.06558*.

Dong, S.; Van Roy, B.; and Zhou, Z. 2021. Simple Agent, Complex Environment: Efficient Reinforcement Learning with Agent State. *arXiv preprint arXiv:2102.05261*.

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

Fedus, W.; Gelada, C.; Bengio, Y.; Bellemare, M. G.; and Larochelle, H. 2019. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*.

Feinberg, E. A.; and Shwartz, A. 1996. Constrained discounted dynamic programming. *Mathematics of Operations Research*, 21(4): 922–945.

François-Lavet, V.; Fonteneau, R.; and Ernst, D. 2015. How to discount deep reinforcement learning: Towards new dynamic strategies. *arXiv preprint arXiv:1512.02011*.

Gao, B.; and Pavel, L. 2017. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*.

Gergely Neu, A. G.; Szepesvári, C.; and Antos, A. 2010. Online Markov decision processes under bandit feedback. In *Proceedings of the Twenty-Fourth Annual Conference on Neural Information Processing Systems*.

Haarnoja, T.; Tang, H.; Abbeel, P.; and Levine, S. 2017. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, 1352–1361. PMLR.

Hambly, B. M.; Xu, R.; and Yang, H. 2020. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *Available at SSRN*.

Hambly, B. M.; Xu, R.; and Yang, H. 2021. Policy Gradient Methods Find the Nash Equilibrium in N-player General-sum Linear-quadratic Games. *Available at SSRN 3894471*.

Haviv, M.; and Van der Heyden, L. 1984. Perturbation bounds for the stationary probabilities of a finite Markov chain. *Advances in Applied Probability*, 804–818.

Henkel, F. 2018. *A Regularization Study for Policy Gradient Methods/submitted by Florian Henkel*. Ph.D. thesis, Universität Linz.

Hessel, M.; Modayil, J.; Van Hasselt, H.; Schaul, T.; Ostrovski, G.; Dabney, W.; Horgan, D.; Piot, B.; Azar, M.; and Silver, D. 2018. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*.

Hordijk, A.; and Yushkevich, A. A. 2002. Blackwell optimality. In *Handbook of Markov decision processes*, 231–267. Springer.

Kakade, S. 2001a. Optimizing average reward using discounted rewards. In *International Conference on Computational Learning Theory*, 605–615. Springer.

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

Konda, V.; and Tsitsiklis, J. 2003. On actor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4): 1143–1166.

Lasserre, J. 1988. Conditions for existence of average and Blackwell optimal stationary policies in denumerable Markov decision processes. *Journal of mathematical analysis and applications*, 136(2): 479–489.

Lewis, M. E.; and Puterman, M. L. 2002. Bias optimality. In *Handbook of Markov decision processes*, 89–111. Springer.

Li, G.; Wei, Y.; Chi, Y.; Gu, Y.; and Chen, Y. 2021. Softmax policy gradient methods can take exponential time to converge. *arXiv preprint arXiv:2102.11270*.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Liu, S.; and Su, H. 2020. γ-Regret for Non-Episodic Reinforcement Learning. *arXiv e-prints*, arXiv–2002.

Liu, Y.; Zhang, K.; Basar, T.; and Yin, W. 2020. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33.

Mahadevan, S. 1996. Sensitive discount optimality: Unifying discounted and average reward reinforcement learning. In *ICML*, 328–336. Citeseer.

Marbach, P. 1998. *Simulation-based optimization of Markov decision processes*. Ph.D. thesis, Massachusetts Institute of Technology.

Marbach, P.; and Tsitsiklis, J. 2001. Simulation-based optimization of Markov reward processes. *IEEE Transactions on Automatic Control*, 46(2): 191–209.

Marbach, P.; and Tsitsiklis, J. N. 2003. Approximate gradient methods in policy-space optimization of Markov reward processes. *Discrete Event Dynamic Systems*, 13(1): 111–148.

Mei, J.; Xiao, C.; Szepesvari, C.; and Schuurmans, D. 2020. On the Global Convergence Rates of Softmax Policy Gradient Methods. *arXiv preprint arXiv:2005.06392*.

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

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533.

Neu, G.; Jonsson, A.; and Gómez, V. 2017. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*.

Ortner, R. 2020. Regret bounds for reinforcement learning via Markov chain concentration. *Journal of Artificial Intelligence Research*, 67: 115–128.

Papini, M.; Binaghi, D.; Canonaco, G.; Pirotta, M.; and Restelli, M. 2018. Stochastic variance-reduced policy gradient. *arXiv preprint arXiv:1806.05618*.

Peters, J.; Mülling, K.; and Altun, Y. 2010. Relative entropy policy search. In *AAAI*, volume 10, 1607–1612. Atlanta.

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

Puterman, M. L. 2014. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Rosenthal, J. S. 1995. Convergence rates for Markov chains. *Siam Review*, 37(3): 387–405.

Ryu, E. K.; and Boyd, S. 2016. Primer on monotone operator methods. *Appl. Comput. Math*, 15(1): 3–43.

Schneckenreither, M. 2020. Average Reward Adjusted Discounted Reinforcement Learning: Near-Blackwell-Optimal Policies for Real-World Applications. *arXiv preprint arXiv:2004.00857*.

Schulman, J.; Chen, X.; and Abbeel, P. 2017. Equivalence between policy gradients and soft Q-learning. *arXiv preprint arXiv:1704.06440*.

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

Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; and Abbeel, P. 2015b. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

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

Shani, L.; Efroni, Y.; and Mannor, S. 2019. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. *arXiv preprint arXiv:1909.02769*.

Stewart, G. 1990. Matrix perturbation theory.

Sutton, R.; McAllester, D.; Singh, S.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, 1057–1063.

Tessler, C.; and Mannor, S. 2020. Reward Tweaking: Maximizing the Total Reward While Planning for Short Horizons. *arXiv preprint arXiv:2002.03327*.

Thomas, P. 2014. Bias in natural actor-critic algorithms. In *International conference on machine learning*, 441–448. PMLR.

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

Wang, N.-Y.; Wu, L.; et al. 2014. Convergence rate and concentration inequalities for Gibbs sampling in high dimension. *Bernoulli*, 20(4): 1698–1716.

Xu, P.; Gao, F.; and Gu, Q. 2019. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*.

Xu, Z.; van Hasselt, H.; and Silver, D. 2018. Meta-gradient reinforcement learning. *arXiv preprint arXiv:1805.09801*.

Zhang, J.; Kim, J.; O'Donoghue, B.; and Boyd, S. 2020a. Sample Efficient Reinforcement Learning with REINFORCE. *arXiv preprint arXiv:2010.11364*.

Zhang, J.; Koppel, A.; Bedi, A. S.; Szepesvari, C.; and Wang, M. 2020b. Variational policy gradient method for reinforcement learning with general utilities. *arXiv preprint arXiv:2007.02151*.

Zhang, K.; Koppel, A.; Zhu, H.; and Başar, T. 2019. Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383*.

Zhang, K.; Zhang, X.; Hu, B.; and Başar, T. 2021. Derivative-Free Policy Optimization for Risk-Sensitive and Robust Control Design: Implicit Regularization and Sample Complexity. *arXiv preprint arXiv:2101.01041*.

Zhao, T.; Niu, G.; Xie, N.; Yang, J.; and Sugiyama, M. 2016. Regularized policy gradients: direct variance reduction in policy gradient estimation. In *Asian Conference on Machine Learning*, 333–348. PMLR.