Understanding Policy Gradient Algorithms: A Sensitivity-Based Approach

Shuang Wu [1] **Ling Shi** [2] **Jun Wang** [3] **Guangjian Tian** [1]

Abstract

The REINFORCE algorithm from Williams is popular in policy gradient (PG) for solving reinforcement learning (RL) problems. Meanwhile, the theoretical form of PG is from Sutton et al. Although both formulae prescribe PG, their precise connections are not yet illustrated. Recently, Nota and Thomas (2020) have found that the ambiguity causes implementation errors. Motivated by the ambiguity and implementation incorrectness, we study PG from a perturbation perspective. In particular, we derive PG in a unified framework, precisely clarify the relation between PG implementation and theory, and echo back the findings by Nota and Thomas. Diving into factors contributing to empirical successes of the existing erroneous implementations, we find that small approximation error and the experience replay mechanism play critical roles.

1. Introduction

The policy gradient (PG) method refers to methods for reinforcement learning (RL) that directly search parameterized decision-making policy using the gradient of a performance metric with respect to the policy parameters. PG-based algorithms (e.g., TRPO (Schulman et al., 2015), DDPG (Lillicrap et al., 2016), PPO (Schulman et al., 2017), TD3 (Fujimoto et al., 2018), SAC (Haarnoja et al., 2018)) have prevailed in solving reinforcement learning (RL) problems for their applicability for both continuous and discrete action space and compatibility with value function approximations.

Despite its popularity, our understanding of PG is limited. Recently, several researchers have found that while PG-based algorithms succeeded in practice, PG implementations are hard to comprehend. Engstrom et al. (2019) showed that code-level details significantly affect performances of TRPO and PPO. Ilyas et al. (2020) showed through experiments that some claims regarding PG are not consistent with those observed in experiments.

Our motivation is two-fold. The first is the ambiguity between empirically implemented PG rooted in (Williams, 1988; 1992) and the theoretical formula derived in (Sutton et al., 1999). Williams (1988; 1992) developed REINFORCE to approximate PG based on empirical trajectory. A few years later, Sutton et al. (1999) derived the theoretical form of PG. While the result in (Sutton et al., 1999) lays down the theoretical foundation of PG, most PG implementations (such as those mentioned at the beginning) follow the recipe prescribed by Williams (1992). Although Sutton et al. (1999) claimed that Williams' REINFORCE algorithm implies the theoretical formula, the precise derivation has not been revealed. The gap between theory and implementation even leads to concerns regarding the correctness of current implementations. In particular, Nota & Thomas (2020) found that most of the modern PG implementations[1] (such as those mentioned at the beginning) followed Williams' episodic PG recipe incorrectly. While they adopted a discounted action-value function estimator, they dropped the discounting factor imposed on the empirical sample path. The asymmetry is inconsistent with the theoretical PG formula prescribed by Sutton et al. (1999).

The other motivation is based on the lack of a unified treatment of various PG formulae. Sutton et al. (1999) derived a unified formula for different setups, but the derivation is performed separately for different setups. Moreover, the related derivations are only performed for stochastic policies and are not directly extendable for the deterministic PG in (Silver et al., 2014). Since all these formulae are essentially gradient of a policy, it is natural to expect that there is a unified framework to derive and treat these equations.

Motivated by the ambiguities between theory and practice and lack of unification, we revisit PG using the perturbation approach originated from (Cao & Chen, 1997). Although

[1]Huawei Noah's Ark Lab [2]Hong Kong University of Science and Technology [3]University College London. Correspondence to: Shuang Wu <wushuang.noah@huawei.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]This issue is so widely spread that even influential platforms, such as `OpenAI Spinning Up` https://spinningup.openai.com/en/latest/index.html and `MATLAB toolbox` https://www.mathworks.com/help/reinforcement-learning/agents.html, inherit this issue.

the approach has not attracted much attention in the modern RL community, we find it helpful in deriving and elucidating PG formulas in a unified and conceptually straightforward way. In particular, we show that the theoretical PG formula corresponds to a general state-based weighting scheme rather than the previous assertion that PG is an expectation with respect to a probability distribution over the state space. Furthermore, we prove that Williams' PG becomes the unbiased PG estimator of the true PG in (Sutton et al., 1999) in various setups by incorporating the weighting scheme. Finally, we analyze why the existing erroneous implementations achieve good empirical performance. Our contributions are three-fold.

1. We introduce a unified perspective for PG in various setups. Using the perturbation approach, we show that one can derive various PG formulae in a unified framework based on the definition of a gradient.

2. We bridge the theoretical PG formula from (Sutton et al., 1999) and the practical implementations (Williams, 1988; 1992). While Williams' episodic REINFORCE algorithm corresponds to the exact PG in standard RL tasks under suitable assumptions, it needs an additional discounting adjustment in a discounted reward setup, which echos back the findings in (Nota & Thomas, 2020).

3. We analyze why existing PG-based algorithms work well in practice. We show that the approximation error decreases to zero when the discounting factor approaches one. We also demonstrate that the experience replay mechanism automatically fixes the implementation errors by breaking transition correlations.

Our refinements on PG may not seem critical from an algorithmic advancement perspective since the existing PG implementations often achieve high scores on RL benchmarks. However, besides algorithmic advancement that achieves better performance, it is equally important to understand algorithms in a theoretically solid position. Our results elucidate PG in a principled yet easy-to-follow perspective, clarify the ambiguity between theory and implementation, and shed light on why existing algorithms work well.

2. Problem Setup and Preliminaries

We clarify different setups of RL and show basic results of PG in this section.

2.1. Markov Decision Process

Markov decision process A Markov decision process (MDP) consists of a quadruple $(\mathcal{S}, \mathcal{A}, P, R)$: state space S of agent state s, action space A of accessible actions a, state transition probability $P(s'|s,a)$ under current state s and selected action a, and the associated per-stage reward function $R(s,a)$.

Policy and Markov chain A stochastic policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes the space of probability distributions over $\mathcal{A}$, is a stochastic assignment of actions given a state. A deterministic policy $\mu : \mathcal{S} \to \mathcal{A}$ reduces the stochastic mapping to a Dirac measure. We can associate a Markov chain and a per-stage reward with every fixed policy, which are $P^\pi(s'|s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[P(s'|s,a)]$ and $R^\pi(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[R(s,a)]$.

Entropy regularization We also consider entropy-regularized for stochastic policies. Formally, the entropy of a policy conditioned on a state s is defined as $\mathcal{H}[\pi(\cdot|s)] := -\mathbb{E}_{a\sim\pi(\cdot|s)}[\log \pi(a|s)]$. We regularize the per-stage reward under a policy by $\tilde{R}^\pi(s) := R^\pi(s) + \tau\mathcal{H}[\pi(\cdot|s)]$, where $\tau \geq 0$ is a constant reflecting preference to randomness. Since $\tilde{R}^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[R(s, a - \tau \log \pi(a|s))]$, we define the regularized reward function as $\tilde{R}(s,a) := R(s,a) - \tau \log \pi(a|s)$. All results in this paper apply to $\tau = 0$, which corresponds to the standard MDP setup.

2.2. Objectives and Tasks

Discounted reward The most widely studied RL setup is to maximize the discounted total reward with $\gamma \in (0,1)$,

$$J_\gamma(\pi;\rho_0) := \mathop{\mathbb{E}}_{\pi,\rho_0}\Big[\sum_{k=0}^{\infty}\gamma^k \tilde{R}(s_k,a_k)\Big], \tag{1}$$

where the expectation $\mathbb{E}_{\pi,\rho_0}$ is evaluated with respect to an initial state distribution $s_0 \sim \rho_0(\cdot)$, a stochastic policy $a_k \sim \pi(\cdot|s_k)$, and an underlying Markov transition probability $s_{k+1} \sim P(\cdot|s_k,a_k)$.

The discount factor γ requires a manual setting as it is not usually available in real-world problems. Mahadevan (1996) illustrated that the optimal policy varies for different γ. We also consider two other objectives.

Total reward for episodic task [2] One common task in RL is to achieve some goal within a finite number of steps. This corresponds to an episodic task. The corresponding objective is then the eventually accumulated total reward

$$J_{\text{tot}}(\pi;\rho_0) := \mathop{\mathbb{E}}_{\pi,\rho_0}\Big[\sum_{k=0}^{\infty}\tilde{R}(s_k,a_k)\Big]. \tag{2}$$

To ensure that $J_{\text{tot}}(\pi)$ is bounded, we assume that under all policies, the state finally transits to a zero-reward terminal

[2] This setup is distinct from the finite horizon setup in the MDP literature. The optimal policy is a time-varying mapping from the state to the action (distributions) in the finite horizon case, which is not attainable by a time-invariant policy.

Table 1. Formulae for state visitation counts $d_\bullet^{\pi,\rho_0}(s)$, state-value function $V_\bullet^\pi(s)$, action-value function $Q_\bullet^\pi(s,a)$, and Bellman equations.

	discounted (total w/ $\gamma=1$) reward	average reward
$V_\bullet^\pi$	$V_\gamma^\pi(s) = \mathop{\mathbb{E}}_{s_{k+1}\sim P^\pi(\cdot\|s_k)}\left[\sum_{k=0}^\infty \gamma^k \tilde{R}^\pi(s_k)\Big\| s_0=s\right]$	$V_{\rm av}^\pi(s) = \mathop{\mathbb{E}}_{s_{k+1}\sim P^\pi(\cdot\|s_k)}\left[\sum_{k=0}^\infty \tilde{R}^\pi(s_k) - J_{\rm av}(\pi)\Big\| s_0=s\right]$
$Q_\bullet^\pi$	$Q_\gamma^\pi(s,a) = R(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot\|s,a)}\left[V_\gamma^\pi(s')\right]$	$Q_{\rm av}^\pi(s,a) = R(s,a) - J_{\rm av}(\pi) + \mathop{\mathbb{E}}_{s'\sim P(\cdot\|s,a)}\left[V_{\rm av}^\pi(s')\right]$
Bellman	$V_\gamma^\pi(s) = \tilde{R}^\pi(s) + \gamma \mathop{\mathbb{E}}_{s'\sim P^\pi(\cdot\|s)}\left[V_\gamma^\pi(s')\right]$	$V_{\rm av}^\pi(s,a) = \tilde{R}^\pi(s) - J_{\rm av}(\pi) + \mathop{\mathbb{E}}_{s'\sim P(\cdot\|s,a)}\left[V_{\rm av}^\pi(s')\right]$
$d_\bullet^{\pi,\rho_0}$	$d_\gamma^{\pi,\rho_0}(s) = \sum_{k=0}^\infty \gamma^k \mathop{\mathbb{E}}_{s_0\sim\rho_0(\cdot)}\left[P^\pi(s_k=s\|s_0)\right]$	$d_{\rm av}^{\pi,\rho_0}(s) = \lim_{k\to\infty} \mathop{\mathbb{E}}_{s_0\sim\rho_0(\cdot)}\left[P^\pi(s_k=s\|s_0)\right]$

state (corresponding to task completion) and stays in the terminal state forever.

Assumption 2.1 (Terminal state)**.** There exists a terminal state z that is accessible from all states. Moreover, for all actions a, $P(z|z,a)=1$ and $R(z,a)=\tau\log\pi(a|z)$ (i.e., $\tilde{R}(z,a)=0$).

Average reward for continuing task Another common task in decision-making is to achieve good average performance while applying actions without stopping. Accordingly, the natural performance criterion is the time-averaged reward over the infinite horizon given by

$$J_{\rm av}(\pi;\rho_0) := \lim_{T\to\infty} \mathop{\mathbb{E}}_{\pi,\rho_0}\left[\frac{1}{T+1}\sum_{k=0}^T \tilde{R}(s_k,a_k)\right]. \quad (3)$$

To ensure that $J_{\rm av}(\pi)$ is meaningful, we make the following assumptions for continuing tasks.

Assumption 2.2 (Ergodicity)**.** The MDP admits ergodic chains under all policies.

Because of ergodicity, $J_{\rm av}(\pi;\rho_0)$ is independent of ρ_0. In the sequel, we write $J_{\rm av}(\pi;\rho_0)$ as $J_{\rm av}(\pi)$ for convenience.

2.3. PG Theory and Implementation

Theory Given a parameterized policy π_θ and an objective $J_\bullet(\pi_\theta)$, Sutton et al. (1999) showed that the PG (without the entropy regularization, i.e., $\tau=0$) is

$$\nabla_\theta J_\bullet(\pi_\theta) = \sum_s d_\bullet^{\pi,\rho_0}(s)\sum_a \nabla_\theta\pi(a|s) Q_\bullet^\pi(s,a). \quad (4)$$

where $d_\bullet^{\pi,\rho_0}$ and $Q_\bullet^{\pi,\rho_0}$ are from Table 1[3]. The formulae in the discounted reward setup reduce to those in the total reward setup by setting $\gamma=1$.

[3]The quantity $d_{\rm tot}^{\pi,\rho_0}(z)=\infty$ since the Markov chain stays in the terminal state z after a finite number of steps. Nevertheless, as $Q_{\rm tot}^\pi(z,\cdot)\equiv 0$, we can either omit z or set $d_{\rm tot}^{\pi,\rho_0}(z)$ as any finite number while computing $\nabla_\theta J_{\rm tot}(\pi_\theta)$.

Implementation The key aspect of (4) is that PG does not involve $\nabla_\theta d_\bullet^{\pi,\rho_0}$, which enables PG approximation by sampling. However, $d_\bullet^{\pi,\rho_0}$ is not directly accessible. Instead, one can only access the conditional distribution $s'\sim P(\cdot|s,a)$ and thus collect sampled trajectories $\{s_0,a_0,s_1,a_1,\dots\}$[4]. Williams (1988; 1992) proposed the episodic REINFORCE algorithm to approximate PG

$$\hat{\nabla}_\theta J_{\rm tot}(\pi_\theta) = \sum_{k=0}^T \nabla_\theta\log\pi(a_k|s_k)\hat{Q}_{\rm tot}^\pi(s_k,a_k), \quad (5)$$

where $\hat{Q}_{\rm tot}^\pi(s_k,a_k) := \sum_{t=k}^T R(s_t,a_t)$.

3. Perturbation Approach for PG

We use the perturbation approach from (Cao & Chen, 1997) to study PG. We show PG formulae can be derived in a unified manner and clarify how to produce unbiased estimates of PG using empirical trajectory. For conciseness, we will drop the notation θ in policies parameterized by θ when its meaning is clear from the context.

3.1. Unified Derivation of PG

The optimization problems in (1), (2), and (3) with a parameterized policy π_θ become parametric optimizations. We apply the fundamental definition of a derivative to derive the PG. Let θ and $\theta'=\theta+\delta\theta$ stand for parameters of two policies. The corresponding policies are π and $\pi'=\pi+\delta\pi$, respectively. We will use the definition of a gradient to derive PG

$$\nabla_\theta J(\theta) = \lim_{\delta\theta\to 0}\frac{J(\theta+\delta\theta)-J(\theta)}{\delta\theta}. \quad (6)$$

Eqn. (6) formulates PG in a unified perspective. However, it requires one to derive $J(\theta+\delta\theta)-J(\theta)$ as a function of $\delta\theta$ or $\delta\pi$ in a closed-form expression. We will show a unified formula to express the performance difference $J(\theta+\delta\theta)-J(\theta)$ for the three objectives, which allows us to derive PG for both stochastic and deterministic policies.

[4]Assuming that $R(s,a)$ is known in advance.

Performance difference To simplify notations in a general setup, we define

$$(\pi' \circ Q^\pi_\bullet)(s) := \mathop{\mathbb{E}}_{a\sim\pi'(\cdot|s)}\Big[Q^\pi_\bullet(s,a) - \tau\log\pi'(a|s)\Big]$$

to represent the contribution to long-term returns at state s for applying π' instead of π. From Table 1, we can derive $(\pi \circ Q^\pi_\bullet)(s) = V^\pi_\bullet(s)$. The performance difference can be expressed as an accumulative difference between applying π' and π.

Proposition 3.1 (Performance difference)**.** *The difference of performances between two policies is*

$$\begin{aligned}&J_\bullet(\pi';\rho_0) - J_\bullet(\pi;\rho_0)\\ &\quad= \sum_s d^{\pi',\rho_0}_\bullet(s)\Big[(\pi'\circ Q^\pi_\bullet)(s) - (\pi\circ Q^\pi_\bullet)(s)\Big].\end{aligned}$$

The proof relies on manipulating the Bellman equations of the two policies. We present the proof for the discounted reward case as an example. The proofs for the other two objectives are in the appendix.

Proof. We let ρ_0 and $d^{\pi,\rho_0}_\bullet$ be row vector representations for $\rho_0(s)$ and $d^{\pi,\rho_0}_\bullet(s)$, P^π be the transition matrix for the corresponding Markov chain under policy π, and $V^\pi_\bullet$ (respectively, $R^\pi_\bullet$) be the column vector representation of $V^\pi_\bullet(s)$ (respectively, $R^\pi_\bullet(s)$).

Based on the Bellman equation, we obtain

$$\begin{aligned}V^{\pi'}_\gamma - V^\pi_\gamma =& \tilde R^{\pi'} + \gamma P^{\pi'} V^{\pi'}_\gamma - \tilde R^\pi - \gamma P^\pi V^\pi_\gamma\\ =& \tilde R^{\pi'} + \gamma P^{\pi'} V^{\pi'}_\gamma - \tilde R^\pi - \gamma P^\pi V^\pi_\gamma\\ &+ \underbrace{\gamma P^{\pi'} V^{\pi'}_\gamma - \gamma P^{\pi'} V^\pi_\gamma}_{\mathcal{T}}.\end{aligned}$$

Subtracting $\mathcal{T}$ on both sides, we obtain

$$\begin{aligned}&(I - \gamma P^{\pi'})(V^{\pi'}_\gamma - V^\pi_\gamma)\\ &\qquad= \tilde R^{\pi'} + \gamma P^{\pi'} V^\pi_\gamma - \tilde R^\pi - \gamma P^\pi V^\pi_\gamma. \quad (7)\end{aligned}$$

Since $(I - \gamma P^{\pi'})$ is invertible, we can derive

$$\rho_0(I - \gamma P^{\pi'})^{-1} = \rho_0 \sum_{k=0}^\infty \gamma^k (P^{\pi'})^k = d^{\pi',\rho_0}_\gamma.$$

We multiply $\rho_0(I-\gamma P^{\pi'})^{-1}$ on both sides of (7) and obtain

$$\begin{aligned}J_\gamma(\pi';\rho_0) - J_\gamma(\pi;\rho_0) &= \rho_0 V^{\pi'} - \rho_0 V^\pi\\ &= d^{\pi',\rho_0}_\gamma[\tilde R^{\pi'} + \gamma P^{\pi'} V^\pi_\gamma - \tilde R^\pi - \gamma P^\pi V^\pi_\gamma]\\ &= \sum_s d^{\pi'}_\gamma(s)\Big[(\pi'\circ Q^\pi_\gamma)(s) - (\pi\circ Q^\pi_\gamma)(s)\Big].\end{aligned}$$

This completes the proof for the discounted reward. □

Remark 3.2. Proposition 3.1 **unifies** the performance difference formulae in various setups in one single equation. It reduces to the result in (Kakade & Langford, 2002; Mei et al., 2020) since $(\pi'\circ Q^\pi_\bullet)(s) - (\pi\circ Q^\pi_\bullet)(s) = \mathbb{E}_{a\sim\pi'(a|s)}[A^\pi_\bullet(s,a)] - \tau D_{\mathrm{KL}}\Big(\pi'(\cdot|s)\|\pi(\cdot|s)\Big)$, where $A^\pi_\bullet(s,a) := Q^\pi_\bullet(s,a) - \tau\log\pi(a|s) - V^\pi_\bullet(s)$.

Derivation of Stochastic PG Consider stochastic policies $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. Let θ and $\theta' = \theta + \delta\theta$ stand for parameters of two policies. The corresponding policies are π and $\pi' = \pi + \delta\pi$, respectively. From the performance difference formula, we can obtain,

$$\begin{aligned}&J_\bullet(\theta+\delta\theta) - J_\bullet(\theta)\\ &= \sum_s d^{\pi',\rho_0}_\bullet(s)\sum_a \delta\pi(a|s)\Big(Q^\pi_\bullet(s,a) - \tau\log\pi(a|s)\Big)\\ &\quad - \tau\sum_s d^{\pi',\rho_0}_\bullet(s)\sum_a \pi'(a|s)\log\frac{\pi'(a|s)}{\pi(a|s)}.\end{aligned}$$

Plugging this into the definition eqn. (6), we obtain

$$\begin{aligned}\nabla J_\bullet(\theta) = &\lim_{\delta\theta\to0}\sum_s d^{\pi',\rho_0}_\bullet(s)\sum_a \frac{\delta\pi(a|s)}{\delta\theta}\\ &\cdot\Big(Q^\pi_\bullet(s,a) - \tau\log\pi(a|s)\Big)\\ &\underbrace{- \lim_{\delta\theta\to0}\tau\sum_s d^{\pi',\rho_0}_\bullet(s)\sum_a \frac{\pi'(a|s)}{\delta\theta}\log\frac{\pi'(a|s)}{\pi(a|s)}}_{=0}.\end{aligned}$$

The second part equals 0 because, for every s,

$$\begin{aligned}&\lim_{\delta\theta\to0}\sum_a \frac{\pi'(a|s)}{\delta\theta}\log\frac{\pi'(a|s)}{\pi(a|s)}\\ &= \lim_{\delta\theta\to0}\sum_a \frac{\delta\pi(a|s)}{\delta\theta}\frac{\pi'(a|s)}{\delta\pi(a|s)}\log\frac{\pi'(a|s)}{\pi(a|s)}\\ &= \sum_a \nabla_\theta\pi(a|s)\cdot 1 = \nabla_\theta\sum_a\pi(a|s) = \nabla_\theta 1 = 0.\end{aligned}$$

The stochastic PG derivation is complete as

$$\nabla J_\bullet(\theta) = \sum_s d^{\pi,\rho_0}_\bullet(s)\sum_a \nabla_\theta\pi(a|s)\tilde Q^\pi_\bullet(s,a),$$

where $\tilde Q^\pi_\bullet(s,a) := Q^\pi_\bullet(s,a) - \tau\log\pi(a|s)$.

Derivation of deterministic PG Apart from the standard stochastic PG, we can derive the deterministic PG (Silver et al., 2014) using the same idea. For deterministic policies $a = \mu_\theta(s)$, we temporally use μ to represent π. Accordingly, $(\pi'\circ Q^\pi_\bullet)(s)$ becomes $Q^\pi_\bullet(s,\mu(s))$. Since the entropy of μ is ill-defined, we set $\tau\equiv 0$ and obtain

$$\begin{aligned}&\frac{J_\bullet(\theta+\delta\theta) - J_\bullet(\theta)}{\delta\theta}\\ &= \sum_s d^{\mu',\rho_0}_\bullet(s)\Big[\frac{Q^\mu_\bullet(s,\mu'(s)) - Q^\mu_\bullet(s,\mu(s))}{\delta\theta}\Big],\end{aligned}$$

which yields the deterministic PG in (Silver et al., 2014),

$$\nabla J_\bullet(\mu_\theta) = \sum_s d_\bullet^{\mu',\rho_0}(s)\Big[\nabla_\theta Q_\bullet^\mu(s,a)|_{a=\mu(s)}\Big] = \sum_s d_\bullet^{\mu',\rho_0}(s)\Big[\nabla_\theta \mu(s) \nabla_a Q_\bullet^\mu(s,a)|_{a=\mu(s)}\Big].$$

We used the perturbation approach to compute the PG by following the definition. The derivations of stochastic PG in (Sutton et al., 1999) are separated for different setups, and are distinct from the deterministic PG in (Silver et al., 2014). We use a unified perspective to derive various PGs.

3.2. PG Approximation from Episodic Trajectory

The theoretical PG is a **spatial** formula since it is a summation over s and a. We can derive equivalent **temporal** forms (summation over time index k) by unrolling $d_\bullet^{\pi,\rho_0}$[5]. The corresponding temporal stochastic PGs are[6]

$$\nabla J_\gamma(\pi;\rho_0) = \mathop{\mathbb{E}}_{\pi,\rho_0}\Big[\sum_{k=0}^\infty \gamma^k \nabla \log \pi(a_k|s_k) \tilde{Q}_\gamma^\pi(s_k,a_k)\Big],$$

$$\nabla J_{\text{tot}}(\pi;\rho_0) = \mathop{\mathbb{E}}_{\pi,\rho_0}\Big[\sum_{k=0}^\infty \nabla \log \pi(a_k|s_k) \tilde{Q}_{\text{tot}}^\pi(s_k,a_k)\Big],$$

$$\nabla J_{\text{av}}(\pi;\rho_0) = \lim_{T\to\infty} \mathop{\mathbb{E}}_{\pi,\rho_0}\Big[\frac{1}{T+1}\sum_{k=0}^T \{\nabla \log \pi(a_k|s_k) \cdot \tilde{Q}_{\text{av}}^\pi(s_k,a_k)\}\Big].$$

The deterministic PGs follow similarly. The derivation details are presented in the appendix.

These temporal equations inspire us to estimate PGs from trajectories $\{s_0,a_0,s_1,a_1,\dots\}$ obtained by querying $s' \sim P(\cdot|s,a)$. We use stochastic policies for illustration. Deterministic policies follow similarly. We define

$$G_\bullet^\pi(s,a) = \nabla_\theta \log \pi(a|s)\Big(Q_\bullet^\pi(s,a) - \tau \log \pi(a|s)\Big),$$

$$\hat{G}_\bullet^\pi(s,a) = \nabla_\theta \log \pi(a|s)\Big(\hat{Q}_\bullet^\pi(s,a) - \tau \log \pi(a|s)\Big).$$

The theoretical PG in eqn. (4) and the REINFORCE algorithm in eqn. (5) becomes

$$\nabla J_\bullet(\pi;\rho_0) := \sum_s d_\bullet^\pi(s) \sum_a \pi(a|s) G_\bullet^\pi(s,a), \quad (8)$$

$$\hat{\nabla} J_\bullet(\pi;\rho_0) := \sum_{k=0}^T \hat{G}_\bullet^\pi(s_k,a_k). \quad (9)$$

[5]Apart from unrolling, one can also derive the temporal PGs by using the temporal performance difference shown in the appendix.

[6]We leverage the log-likelihood trick shown in the appendix.

Assuming that $\mathbb{E}_{\pi,\rho_0}[\hat{Q}_\bullet^\pi(s_k,a_k)]$ is an unbiased estimate of $Q_\bullet^\pi(s,a)$, we show that eqn. (9) is an unbiased estimate of eqn. (8). Details are in the appendix.

Episodic task with total reward. No adjustment is required in this case since Williams' formula is designed to solve this task. In particular, we obtain

$$\mathop{\mathbb{E}}_{\pi,\rho_0}\Big[\sum_{k=0}^T \hat{G}_{\text{tot}}^\pi(s_k,a_k)\Big] = \nabla J_{\text{tot}}(\pi;\rho_0). \quad (10)$$

Continuing task with average reward. In continuing tasks, there is no terminal state. Asymptotically, we have

$$\lim_{T\to\infty} \mathop{\mathbb{E}}_{\pi,\rho_0}\Big[\frac{1}{T+1}\sum_{k=0}^T \hat{G}_{\text{av}}^\pi(s_k,a_k)\Big] = \nabla J_{\text{av}}(\pi). \quad (11)$$

Furthermore, since the MDP is ergodic, any finite T yields an unbiased estimate. Therefore, Williams' PG is still unbiased (modulo a $1/T$ factor).

Discounted reward. For the discounted reward, the discounted scaling is required, which echos back the claim made in Nota & Thomas (2020). In particular,

$$\lim_{T\to\infty} \mathop{\mathbb{E}}_{\pi,\rho_0}\Big[\sum_{k=0}^T \gamma^k \hat{G}_\gamma^\pi(s_k,a_k)\Big] = \nabla J_\gamma(\pi;\rho_0). \quad (12)$$

In episodic tasks, the discounted REINFORCE variant is an unbiased PG estimate for a finite T. However, in continuing tasks, the variant is asymptotically unbiased. Nevertheless, without discounting, the vanilla episodic REINFORCE is always biased.

In summary, Williams' REINFORCE algorithm yields an unbiased PG estimate for both episodic and continuing tasks under their corresponding natural objectives. To adapt the formula to the discounted reward criterion, discounted factor should be multiplied. In continuing tasks, an unbiased discounted PG estimate is attainable only when $T \to \infty$.

4. Analysis of Current Implementations

Current implementations mostly maximize J_γ and approximate ∇J_γ by sampling. However, these algorithms essentially sample from the conditional distribution $s' \sim P(\cdot|s,a)$ instead of d_γ^{π,ρ_0} since the latter is not directly accessible. In general, $d_\bullet^{\pi,\rho_0}$ should be understood as the weights of a **weighted sum** instead of a state distribution. To approximate ∇J_γ as prescribed in eqn. (4), one should follow eqn. (12) by applying additional discount factors to account for d_γ^{π,ρ_0} based on the trajectory data. Mistaking the weighted sum as a probability distribution leads to the incorrect implementations observed in (Nota & Thomas, 2020).

The existing PG implementations perform well in practice although they do not strictly follow the unbiased PG estimate. The good empirical results suggest that these implementations are valid algorithm options. We dig into the issue further by showing why they work well. In particular, we quantify the approximation error bound due to the discount factor and find that the experience replay mechanism automatically fixes the implementation error.

4.1. Small Approximation Error

Most PG implementations claim to solve the discounted objective in eqn. (1). In episodic tasks, the discounted total reward reduces to the total reward for episodic tasks when $\gamma \uparrow 1$. In continuing tasks, as $\gamma \uparrow 1$, the discounted objective approaches the average reward objective due to Abelian theorem [Lemma 5.3.1 in (Hernández-Lerma & Lasserre, 1996)] as $\lim_{\gamma\uparrow 1}(1-\gamma)\sum_{k=0}^{\infty}\gamma^k R_k = \lim_{T\to\infty}\frac{1}{T+1}\sum_{k=0}^{T} R_k$. Both cases suggest that if γ is close to one, the discounted objective is a valid surrogate objective of the actual objective. The following theorems quantify the approximation error due to the discounted objective.

Assumption 4.1. There exist an integer $m > 0$ and a positive real number $\alpha < 1$ such that

$$\Pr(s_m \neq z | s_0, \pi) \leq \alpha, \quad \forall\, s_0 \in S,\ \pi.$$

Theorem 4.2 (episodic tasks)**.** *Under Assumption 4.1, both* $\|\tilde{d}^{\pi,\rho_0}_{\text{tot}} - \tilde{d}^{\pi,\rho_0}_{\gamma}\|_\infty$ *and* $\|V^{\pi}_{\text{tot}} - V^{\pi}_{\gamma}\|_\infty$ *are upper bounded by* $O\Big(\frac{m\alpha(1-\gamma^m)}{(1-\alpha)(1-\alpha\gamma^m)}\Big)$*, where* $\tilde{d}^{\pi,\rho_0}_{\bullet}$ *is* $d^{\pi,\rho_0}_{\bullet}$ *over transient states.*

Assumption 4.3 (geometric ergodicity)**.** There exist positive constants C and $\beta < 1$ such that

$$\max_{s\in\mathcal{S}} \|\rho_0 (P^\pi)^k - d^{\pi,\rho_0}_{\text{av}}\|_{\text{TV}} \leq C\beta^k,$$

where $\|\cdot\|_{\text{TV}}$ is the total variation norm in Definition C.1 in the Appendix.

Theorem 4.4 (continuing tasks)**.** *Under Assumption 4.3, both* $\|(1-\gamma)d^{\pi,\rho_0}_{\gamma} - d^{\pi,\rho_0}_{\text{av}}\|_{\text{TV}} \leq O(\frac{1-\gamma}{1-\beta\gamma})$ *and* $sp(V^{\pi}_{\text{av}} - V^{\pi}_{\gamma}) \leq O\Big(\frac{\beta(1-\gamma)}{(1-\beta)(1-\beta\gamma)}\Big)$*, where* $sp(\cdot)$ *is the span seminorm in Definition C.2 in the Appendix.*

Both Theorem 4.2 and 4.4 show that the structure of the MDP and the discount factor γ affect the approximation error. In particular, the vanishing approximation errors due to $\gamma \uparrow 1$ shows that our bound is asymptotically tight for γ. Moreover, smaller m, α, and β are roughly equivalent to faster convergence of MDP and thus a shorter "effective horizon". Therefore, a shorter "effective horizon" also leads to a smaller approximation error. By contrast, for a fixed desired approximation error, if the MDP has a long "effective horizon", the discount factor γ should be set as a large one. Since γ is typically chosen to be large values like 0.99, 0.995, and 0.999, the approximation errors are small despite that they might solve a "slow" MDP.

4.2. Benefits of Experience Replay

Experience replay in reinforcement learning (Lin, 1993) refers to using random samples from previous transitions, which breaks the transition correlations. While it first gained popularity for learning value functions (e.g., DQN (Mnih et al., 2013)), the modern of PG-based algorithms (e.g., (Silver et al., 2014; Lillicrap et al., 2016; Fujimoto et al., 2018)) apply the idea to estimating PG. We show that applying experience replay also helps improve the performance of PG-based algorithms.

Unbiased PG estimate. When the transition correlation is erased, we obtain the following equality,

$$\mathop{\mathbb{E}}_{\substack{s_k\sim\rho_{\text{replay}}(\cdot)\\ a_k\sim\pi(\cdot|s_k)}}\left[\frac{1}{T}\sum_{k=0}^{T}\hat{G}_\gamma(s_k,a_k)\right] = \mathop{\mathbb{E}}_{\substack{s\sim\rho_{\text{replay}}(\cdot)\\ a\sim\pi(\cdot|s)}}\Big[G^\pi_\gamma(s,a)\Big],$$

where ρ_{replay} is determined by the sampling mechanism of the replay buffer. As $\Delta(\mathcal{S})$ is compact, there exists some ρ_0' such that

$$\rho_{\text{replay}}(s) = (1-\gamma)\sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{s_0\sim\rho_0'(\cdot)}[P^\pi(s_k = s|s_0)].$$

Therefore, the "incorrect" empirical PG with experience replay is an unbiased estimate of $\nabla J_\gamma(\pi;\rho_0')$ (modulo a $\frac{T}{1-\gamma}$ factor).

Maximizer invariance. Let $\pi^\star = \arg\max_\pi J_\gamma(\pi;\rho_0)$. The theory of MDP shows that

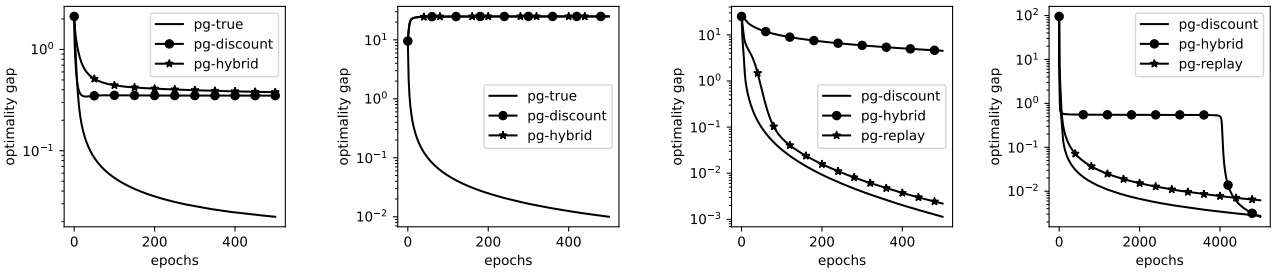
$$\pi^\star(a|s) = \arg\max_\pi \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}\Big[Q^\star_\gamma(s,a) - \tau\log\pi(a|s)\Big],$$

where $Q^\star_\gamma(s,a) := R(s,a) + \gamma\,\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^\star(s')]$ and $V^\star(s)$ satisfies the Bellman optimality equation

$$V^\star_\gamma(s) = \max_\pi \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}\left\{\tilde{R}(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P^\pi(\cdot|s,a)}\Big[V^\star(s')\Big]\right\}.$$

Since $V^\star_\gamma$ is independent of ρ_0, so is $Q^\star_\gamma$. Therefore, $\pi^\star$ is invariant with respect to ρ_0. Consequently, $\pi^\star$ is the maximizer of $J_\gamma(\pi;\rho_0')$ for any $\rho_0' \in \Delta(\mathcal{S})$ despite that $J_\gamma(\pi^\star;\rho_0')$ generally does not equal $J_\gamma(\pi^\star;\rho_0)$.

In summary, by utilizing a replay buffer, the "incorrect" PG implementation corresponds to the gradient of another objective $J_\gamma(\pi;\rho_0')$ while the maximizer of $J_\gamma(\pi;\rho_0')$ is the same as $J_\gamma(\pi;\rho_0)$. Consequently, the "incorrect" PG implementation indirectly improves $J_\gamma(\pi;\rho_0)$.



(a) Continuing task w/ J_{av}. (b) Episodic task w/ J_{tot}. (c) Continuing task w/ J_γ. (d) Episodic task w/ J_γ.

Figure 1. Convergence trajectories of optimality gap under PG algorithms in different settings.

Remark 4.5 (Scope of applicability). The maximizer invariance result applies to general MDPs. However, the invariance property does not necessarily hold in the presence of a general function approximator. If the approximator cannot represent the true maximizer, the invariance property can fail. Nevertheless, the invariance property is not limited to the tabular MDP. For example, the invariance property holds for a linear quadratic control problem although this MDP has continuous states and actions.

5. Numerical Examples

We perform numerical simulations to verify our theoretic results. In particular, we evaluate variants of PG algorithms on different settings and compare the performance **optimality gap** of every policy in each epoch during optimization. The optimal performance is obtained through the policy iteration algorithm for each setup.

Common setup. The stepsize for policy gradient algorithms are set to be 0.1 for all cases. The temperature τ for entropy regularization is 10. We did not directly implement a replay buffer but used the `pg-hybrid` gradient (to be introduced next) to illustrate our results.

5.1. Continuing Task with the Average Reward

MDP and objective. We adopt the controlled restart process (Akbarzadeh & Mahajan, 2019). The transition probability is

$$P(s'|s,a) = \begin{cases} 1, & s' = s+1, a = 0, \\ p, & s' = 0, a = 1, \\ 1-p, & s' = s+1, a = 1. \end{cases}$$

The reward function is $R(s,a) = -s^2 - \lambda \cdot \mathbb{1}(a=1)$. We use a truncated state space and let $P(s'=s|s,a) = P(s'=s+1|s,a)$ if $s = |\mathcal{S}| - 1$. We measure the policy performances with $J_{\text{av}}(\pi;\rho_0)$.

Algorithms. We compare the following PG variants, `pg-true` $\nabla J_{\text{av}}(\pi;\rho_0)$, `pg-discount` $\nabla J_\gamma(\pi;\rho_0)$, and `pg-hybrid` $\sum_{s,a} d_{\text{av}}^{\pi,\rho_0}(s)\pi(a|s)G_\gamma^\pi(s,a)$[7]. The discount factor is 0.7.

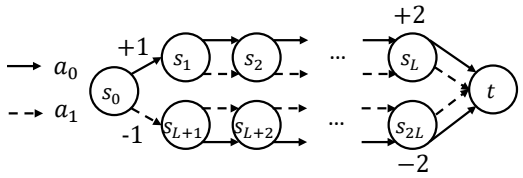


Figure 2. The binary chain example adapted from (Nota & Thomas, 2020). Taking $a = a_0$ receives $+1$ and moves into the upper chain. Taking $a = a_1$ receives -1 reward and moves into the lower chain. The final reward for the two chains are $+2$ and -2 respectively. Other states and actions incur no reward. A large γ ensures acquiring the optimal policy regarding the true total reward while a small γ yields a "short effective horizon" and thus a suboptimal policy. We set the chain length as $L = 3$.

Result. Figure 1a shows the trajectory of optimality gaps for each algorithm. Since `pg-discount` and `pg-hybrid` are tracking the incorrect objective, they fail to converge to the optimal policy and converge to a suboptimal policy instead. Moreover, `pg-discount` and `pg-hybrid` yield the same asymptotic behavior, which echos back our claim that `pg-hybrid` is equivalent to $\nabla J_\gamma(\pi;\rho_0')$ for some ρ_0'.

5.2. Episodic Task with the Total Reward

MDP and objective. We adapt the binary chain example in (Nota & Thomas, 2020) as shown in Figure 2.

Algorithms. We compare the following PG variants, `pg-true` $\nabla J_{\text{tot}}(\pi;\rho_0)$, `pg-discount` $\nabla J_\gamma(\pi;\rho_0)$, and `pg-hybrid` $\sum_{s,a} d_{\text{tot}}^{\pi,\rho_0}(s)\pi(a|s)G_\gamma^\pi(s,a)$. The discount factor is $\gamma = 0.7$.

Result. Figure 1b shows the trajectory of optimality gaps for each algorithm. Similar to the average reward case, only

[7] We use the pg-hybrid to simulate the effect of using a replay buffer for all the three case studies.

`pg-true` converges to the optimal policy, while the other two fail. This example straightforwardly illustrates that the discount factor on value functions suffers from a shorter "effective horizon" than the whole trajectory and leads to sub-optimal policies. Moreover, in this case, `pg-discount` and `pg-hybrid` yield the same transient behavior, which again echos back the equivalence of the two algorithms.

5.3. Continuing Task with the Discounted Reward

MDP and objective. We use the controlled restart process while measuring policy performances with the discounted total reward $J_\gamma(\pi;\rho_0)$ for $\gamma = 0.7$.

Algorithms. We compare the following PG variants, `pg-discount` $\nabla J_\gamma(\pi;\rho_0)$, `pg-hybrid` $\sum_{s,a} d^{\pi,\rho_0}_{\text{av}}(s)\pi(a|s)G^\pi_\gamma(s,a)$, and `pg-replay` $\sum_{s,a} \rho_{\text{replay}}(s)\pi(a|s)G^\pi_\gamma(s,a)$. We let ρ_{replay} be a uniform distribution over the state space.

Result. Figure 1c shows the trajectory of optimality gaps for each algorithm. All algorithms converge to the optimal policy. `pg-hybrid` is much slower than the others, while `pg-replay` yields almost similar performances as the true PG `pg-discount`. Note that `pg-discount` keeps the optimization objective fixed while `pg-replay` keeps the weighting factor of s fixed, which shows the advantage of keeping a reference point unchanged during optimization.

5.4. Episodic Task with the Discounted Reward

MDP and objective. We randomly generate a 10-by-10 grid world [Example 3.5 (Sutton & Barto, 2018)] with eight obstacle grids. Two ending grids are randomly placed. One corresponds to a +1 reward, and the other a -1 reward.

Algorithms. We again compare `pg-discount`, `pg-hybrid`, and `pg-replay`. `pg-hybrid` is changed into $\sum_{s,a} d^{\pi,\rho_0}_{\text{tot}}(s)\pi(a|s)G^\pi_\gamma(s,a)$ while the other two remains unchanged.

Result. Figure 1d shows the trajectory of optimality gaps for each algorithm. Similar to the previous discounted reward setup, `pg-replay` and `pg-discount` yield very similar performance while the trajectory of `pg-hybrid` is unstable. In particular, `pg-hybrid` seems trapped in a flat region while the other algorithms converge smoothly. This again demonstrates the advantage of experience replay for policy optimization.

6. Related Works

Policy gradient PG was derived simultaneously by researchers from various communities (Cao & Chen, 1997; Sutton et al., 1999; Baxter & Bartlett, 2000; Konda & Tsitsiklis, 2000; Marbach & Tsitsiklis, 2001). While they are equivalent under certain circumstances, the modern form follows from (Sutton et al., 1999). Our perturbation approach is inspired by the idea in (Cao & Chen, 1997), which addressed performance sensitivity of general Markov systems operated for continuing tasks. See (Cao, 2007) for a comprehensive overview. We extend the discussion in (Cao, 2007) to episodic task and concretize the general theory in the context of modern RL.

Theoretical analysis of PG Theoretical analysis of PG has been active recently. Agarwal et al. (2020; 2021) studied general convergence properties and sample efficiency of PG methods with function approximations. Despite that there are plentiful results on convergence rates (Mei et al., 2020; Bhandari & Russo, 2021; Li et al., 2021; Zhan et al., 2021) and sample efficiency of PG (Yuan et al., 2021; Cassel & Koren, 2021; Zhang et al., 2021a;b), connections between theory and implementations have not been addressed.

Policy regularization Williams & Peng (1991) pioneered the use of entropy to enhance performance on hierarchical tasks. Incorporating entropy of a policy as a part of the received reward has been popular recently (Haarnoja et al., 2017; 2018) as entropy regularization can 1) help with exploration as more stochastic policies are encouraged (Mnih et al., 2016) and 2) make optimization landscape smoother (Ahmed et al., 2019). Geist et al. (2019) developed the general theory for regularization in MDP.

Concerns for PG implementation A number of studies (Thomas, 2014; Engstrom et al., 2019; Liu et al., 2019; Wang et al., 2019; Nota & Thomas, 2020; Ilyas et al., 2020; Hu et al., 2020) pointed out that our understanding of PG is limited as there are mismatches between theoretical results and implementation. Ilyas et al. (2020) showed that a number commonly-accepted beliefs about PG is flawed. Wang et al. (2019) and Engstrom et al. (2019) discussed code-level caveats in implementing the famous TRPO and PPO algorithms. Thomas (2014) and Nota & Thomas (2020) showed that a number of PG implementations are inconsistent with the theoretical formulae. Wen et al. (2021) found that a class of PG variants corresponds a to optimizing an objective different from the original MDP objective. Motivated by the findings in (Nota & Thomas, 2020), our work builds a solid foundation for various PG algorithms and bridges the cognitive gap between PG theory and implementations.

7. Conclusion

We used a perturbation approach to study PG, which helped us derive PG in a conceptually straightforward manner. The alternative approach also enables us to bridge Williams' empirical PG formula to the theoretical formula in (Sutton et al., 1999), which echos back recent findings in (Nota & Thomas, 2020). Additionally, we showed that small

approximation errors under large discount factors and the experience replay mechanism contribute to the empirical success of PG-based algorithms.

References

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

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

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

Akbarzadeh, N. and Mahajan, A. Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *IEEE Conference on Decision and Control*, pp. 7294–7300, 2019.

Baxter, J. and Bartlett, P. L. Direct gradient-based reinforcement learning. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 3, pp. 271–274, 2000.

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

Cao, X.-R. *Stochastic Learning and Optimization: A Sensitivity-Based Approach*. Springer Science & Business Media, 2007.

Cao, X.-R. and Chen, H.-F. Perturbation realization, potentials, and sensitivity analysis of markov processes. *IEEE Transactions on Automatic Control*, 42(10):1382–1393, 1997.

Cassel, A. and Koren, T. Online policy gradient for model free learning of linear quadratic regulators with $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pp. 1304–1313, 2021.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., and Madry, A. Implementation matters in deep RL: A case study on PPO and TRPO. In *International Conference on Learning Representations*, 2019.

Fujimoto, S., Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pp. 1587–1596, 2018.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169, 2019.

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

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

Hernández-Lerma, O. and Lasserre, J. B. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer Science & Business Media, 1996.

Hu, K.-C., Hsieh, P.-C., Wei, T. H., and Wu, I.-C. Rethinking deep policy gradients via state-wise policy improvement. In *"I Can't Believe It's Not Better!" NeurIPS 2020 workshop*, 2020.

Ilyas, A., Engstrom, L., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., and Madry, A. A closer look at deep policy gradients. In *International Conference on Learning Representations*, 2020.

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

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pp. 1008–1014, 2000.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Softmax policy gradient methods can take exponential time to converge. In *Conference on Learning Theory*, 2021.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

Lin, L.-J. *Reinforcement Learning for Robots Using Neural Networks*. PhD thesis, School of Computer Science, Carnegie Mellon University, 1993.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Off-policy policy gradient with state distribution correction. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1180–1190, 2019.

Mahadevan, S. Optimality criteria in reinforcement learning. In *Proceedings of the AAAI Fall Symposium on Learning Complex Behaviors in Adaptive Intelligent Systems*, 1996.

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

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

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing Atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*, 2013.

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

Nota, C. and Thomas, P. S. Is the policy gradient a gradient? In *International Conference on Autonomous Agents and Multiagent Systems*, pp. 939–947, 2020.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897, 2015.

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

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *International Conference on Machine learning*, pp. 387–395, 2014.

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

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pp. 1057–1063, 1999.

Thomas, P. Bias in natural actor-critic algorithms. In *International Conference on Machine Learning*, pp. 441–448, 2014.

Wang, Y., He, H., and Tan, X. Truly proximal policy optimization. In *Conference on Uncertainty in Artificial Intelligence*, pp. 113–122, 2019.

Wen, J., Kumar, S., Gummadi, R., and Schuurmans, D. Characterizing the gap between actor-critic and policy gradient. In *International Conference on Machine Learning*, pp. 11101–11111, 2021.

Williams, R. J. Toward a theory of reinforcement-learning connectionist systems. Technical Report Technical Report NU-CCS-88-3, Northeastern University, 1988.

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

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

Yuan, R., Gower, R. M., and Lazaric, A. A general sample complexity analysis of vanilla policy gradient. In *ICML Workshop on Reinforcement Learning Theory*, 2021.

Zhan, W., Cen, S., Huang, B., Chen, Y., Lee, J. D., and Chi, Y. Policy mirror descent for regularized RL: A generalized framework with linear convergence. In *International OPT Workshop on Optimization for Machine Learning*, 2021.

Zhang, J., Kim, J., O'Donoghue, B., and Boyd, S. Sample efficient reinforcement learning with REINFORCE. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10887–10895, 2021a.

Zhang, J., Ni, C., Yu, Z., Szepesvari, C., and Wang, M. On the convergence and sample efficiency of variance-reduced policy gradient method. In *Advances in Neural Information Processing Systems*, 2021b.

Outline of the Appendix

Notations We let ρ_0 and $d_\bullet^{\pi,\rho_0}$ be row vector representations for $\rho_0(s)$ and $d_\bullet^{\pi,\rho_0}(s)$, P^π be the transition matrix for the corresponding Markov chain under policy π, and $V_\bullet^\pi$ (respectively, $R_\bullet^\pi$) be the column vector representation of $V_\bullet^\pi(s)$ (respectively, $R_\bullet^\pi(s)$).

A. Proof of Proposition 3.1 (performance difference formula)

We will prove both the spatio version and the temporal version, i.e.,

$$J_\bullet(\pi';\rho_0) - J_\bullet(\pi;\rho_0) = \sum_s d_\bullet^{\pi',\rho_0}(s)\Big[(\pi' \circ Q_\bullet^\pi)(s) - (\pi \circ Q_\bullet^\pi)(s)\Big].$$

or equivalently

$$\begin{aligned}
J_\gamma(\pi';\rho_0) - J_\gamma(\pi;\rho_0) &= \sum_{k=0}^{\infty} \gamma^k \mathop{\mathbb{E}}_{\pi',\rho_0}\Big[A_\gamma^\pi(s_k,a_k)\Big],\\
J_{\mathrm{tot}}(\pi';\rho_0) - J_{\mathrm{tot}}(\pi;\rho_0) &= \sum_{k=0}^{\infty} \mathop{\mathbb{E}}_{\pi',\rho_0}\Big[A_{\mathrm{tot}}^\pi(s_k,a_k)\Big],\\
J_{\mathrm{av}}(\pi') - J_{\mathrm{av}}(\pi) &= \lim_{T\to\infty} \frac{1}{T+1} \sum_{k=0}^{T} \mathop{\mathbb{E}}_{\pi',\rho_0}\Big[A_{\mathrm{av}}^\pi(s_k,a_k)\Big],
\end{aligned}$$

where $A_\bullet^\pi(s,a) = Q_\bullet^\pi(s,a) - \tau\log\pi(a|s) - V_\bullet^\pi(s)$.

A.1. Spatial Version

We have proven the case for the **discounted reward**. We now prove for the total reward and the average reward.

Total reward Regarding the Bellman equation, the total reward setup is the special case of the discounted reward setup. Algebraically, the results follow directly. However, $(I - P^{\pi'})$ has a zero eigenvalue and is thus not invertible. As state z contributes no reward for any action, we consider instead the transient part of the state space and the corresponding transition $\tilde{P}^\pi$. Now, $(I - \tilde{P}^{\pi'})$ is invertible and we can obtain

$$J_{\mathrm{tot}}(\pi';\rho_0) - J_{\mathrm{tot}}(\pi;\rho_0) = \sum_{s\in S\setminus\{z\}} d_{\mathrm{tot}}^{\pi';\rho_0}(s)\Big\{\mathop{\mathbb{E}}_{a\sim\pi'(\cdot|s)}[Q_{\mathrm{tot}}^\pi(s,a) - \tau\log\pi'(a|s)] - \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}[Q_{\mathrm{tot}}^\pi(s,a) - \tau\log\pi(a|s)]\Big\}.$$

Since $\mathbb{E}_{a\sim\pi'(\cdot|s)}[Q_{\mathrm{tot}}^\pi(z,a) - \tau\log\pi'(a|z)] - \mathbb{E}_{a\sim\pi(\cdot|s)}[Q_{\mathrm{tot}}^\pi(z,a) - \tau\log\pi(a|z)] \equiv 0$, the summation can be extended to all $s \in S$.

Average Reward By directly applying the Bellman equation, we obtain

$$\begin{aligned}
J_{\mathrm{av}}(\pi') - J_{\mathrm{av}}(\pi) =& \tilde{R}^{\pi'} + P^{\pi'}V_{\mathrm{av}}^{\pi'} - \tilde{R}^\pi - P^\pi V_{\mathrm{av}}^\pi + V_{\mathrm{av}}^\pi - V_{\mathrm{av}}^{\pi'}\\
=& \tilde{R}^{\pi'} + P^{\pi'}V_{\mathrm{av}}^\pi - \tilde{R}^\pi - P^\pi V_{\mathrm{av}}^\pi + \underbrace{V_{\mathrm{av}}^\pi - V_{\mathrm{av}}^{\pi'} + P^{\pi'}V_{\mathrm{av}}^{\pi'} - P^{\pi'}V_{\mathrm{av}}^\pi}_{\mathcal{T}},
\end{aligned}$$

Multiplying both sides by the row vector $d_{\mathrm{av}}^{\pi'}$, we obtain $d_{\mathrm{av}}^\pi \mathcal{T} = 0$ because $d_{\mathrm{av}}^{\pi'} = d_{\mathrm{av}}^{\pi'}P^{\pi'}$. Finally, we derive

$$\begin{aligned}
J_{\mathrm{av}}(\pi') - J_{\mathrm{av}}(\pi) =& d_{\mathrm{av}}^{\pi'}[\tilde{R}^{\pi'} + P^{\pi'}V_{\mathrm{av}}^\pi - \tilde{R}^\pi - P^\pi V_{\mathrm{av}}^\pi]\\
=& d_{\mathrm{av}}^{\pi'}\Big\{[\tilde{R}^{\pi'} - J_{\mathrm{av}}(\pi) + P^{\pi'}V_{\mathrm{av}}^\pi] - [\tilde{R}^\pi - J_{\mathrm{av}}(\pi) + P^\pi V_{\mathrm{av}}^\pi]\Big\}\\
=& \sum_s d_{\mathrm{av}}^{\pi'}(s)\Big\{\mathop{\mathbb{E}}_{a\sim\pi'(\cdot|s)}[Q_{\mathrm{av}}^\pi(s,a) - \tau\log\pi'(a|s)] - \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}[Q_{\mathrm{av}}^\pi(s,a) - \tau\log\pi(a|s)]\Big\}.
\end{aligned}$$

A.2. Temporal Version

According to the definition of $d_\bullet^{\pi,\rho_0}$, the temporal version can be directly obtained by expanding $d_\bullet^{\pi,\rho_0}$ and vice versa. Nevertheless, we provide another method to prove the temporal version from scratch.

Discounted reward The proof resembles Kakade & Langford, 2002, Lemma 6.1. We present it here for completeness.

$$\begin{aligned}
J_\gamma(\pi';\rho_0) - J_\gamma(\pi';\rho_0) &= \mathop{\mathbb{E}}_{s_0\sim\rho_0(\cdot)}[V_\gamma^{\pi'}(s_0) - V_\gamma^{\pi}(s_0)] \\
&= \mathop{\mathbb{E}}_{s_0\sim\rho_0(\cdot)}\left[\sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi',s_0}[\tilde{R}(s_k,a_k)] - V_\gamma^{\pi}(s_0)\right] \\
&= \mathop{\mathbb{E}}_{s_0\sim\rho_0(\cdot)}\left[\sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi',s_0}\left[\tilde{R}(s_k,a_k) + \gamma V^{\pi}(s_{k+1}) - \gamma V^{\pi}(s_{k+1})\right] - V_\gamma^{\pi}(s_0)\right] \\
&\quad \text{(telescoping sum \& rearranging terms)} \\
&= \mathop{\mathbb{E}}_{s_0\sim\rho_0(\cdot)}\left[\sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi',s_0}\left[\tilde{R}(s_k,a_k) + \gamma V^{\pi}(s_{k+1}) - V^{\pi}(s_k)\right]\right] \\
&= \sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi',\rho_0}\left[A_\gamma^{\pi}(s_k,a_k)\right]
\end{aligned}$$

Total reward The derivation is the same as the discounted case by letting $\gamma = 1$.

Average reward Recall the Abelian theorem from Hernández-Lerma & Lasserre, 1996, Lemma 5.3.1

$$\lim_{T\to\infty}\frac{1}{T+1}\sum_{k=0}^{T}x_k = \lim_{\gamma\uparrow 1}(1-\gamma)\sum_{k=0}^{\infty}\gamma^k x_k.$$

Therefore,

$$\begin{aligned}
J_{\mathrm{av}}(\pi') - J_{\mathrm{av}}(\pi) &= \lim_{\gamma\uparrow 1}(1-\gamma)\Big(J_\gamma(\pi';\rho_0) - J_\gamma(\pi';\rho_0)\Big) \\
&= \lim_{\gamma\uparrow 1}(1-\gamma)\sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi',\rho_0}\left[A_\gamma^{\pi}(s_k,a_k)\right] \\
&= \lim_{T\to\infty}\frac{1}{T+1}\sum_{k=0}^{T}\mathop{\mathbb{E}}_{\pi',\rho_0}\left[\lim_{\gamma\uparrow 1}A_\gamma^{\pi}(s_k,a_k)\right].
\end{aligned}$$

Note that for any fixed state s_{fixed}, we can obtain

$$\begin{aligned}
\lim_{\gamma\uparrow 1}A_\gamma^{\pi}(s,a) &= \lim_{\gamma\uparrow 1}Q_\gamma^{\pi}(s,a) - \tau\log\pi(a|s) - \lim_{\gamma\uparrow 1}V_\gamma^{\pi}(s) \\
&= \lim_{\gamma\uparrow 1}\Big(Q_\gamma^{\pi}(s,a) - V_\gamma^{\pi}(s_{\mathrm{fixed}})\Big) - \tau\log\pi(a|s) - \lim_{\gamma\uparrow 1}\Big(V_\gamma^{\pi}(s) - V_\gamma^{\pi}(s_{\mathrm{fixed}})\Big) \\
&= Q_{\mathrm{av}}^{\pi}(s,a) - \tau\log\pi(a|s) - V^{\pi}(s).
\end{aligned}$$

The proof is thus complete.

B. Bridge between PG Implementation and Theory

We first derive the temporal PG and then show unbiased empirical PG estimates.

B.1. Derivation of Temporal PG

We can derive the temporal PG by either unrolling $d_\bullet^{\pi,\rho_0}$ or leveraging the temporal performance difference formula.

B.1.1. UNROLLING APPROACH

Discounted reward The result follows directly through unrolling as

$$\begin{aligned}
\nabla J_\gamma(\pi;\rho_0) &= \sum_s d_\gamma^{\pi,\rho_0} \sum_a \nabla\pi(a|s)\tilde{Q}_\gamma^\pi(s,a) \\
&\quad [\text{from } \nabla\pi(a|s) = \pi(a|s)\nabla\log\pi(a|s)] \\
&= \sum_s d_\gamma^{\pi,\rho_0} \sum_a \pi(a|s)\nabla\log\pi(a|s)\tilde{Q}_\gamma^\pi(s,a) \\
&= \sum_s \sum_{k=0}^\infty \gamma^k \mathop{\mathbb{E}}_{s_0\sim\rho_0(\cdot)}\Big[P^\pi(s_k=s|s_0)\Big] \sum_a \pi(a|s)\nabla\log\pi(a|s)\tilde{Q}_\gamma^\pi(s,a) \\
&= \mathop{\mathbb{E}}_{\pi,\rho_0}\Big[\sum_{k=0}^\infty \gamma^k \nabla\log\pi(a_k|s_k)\tilde{Q}_\gamma^\pi(s_k,a_k)\Big].
\end{aligned}$$

Total reward The derivation is the same as the discounted case by letting $\gamma = 1$.

Average reward The result follows by using the ergodic theory as

$$\begin{aligned}
\nabla J_{\text{av}}(\pi;\rho_0) &= \sum_s d_{\text{av}}^{\pi,\rho_0} \sum_a \nabla\pi(a|s)\tilde{Q}_{\text{av}}^\pi(s,a) \\
&\quad [\text{from } \nabla\pi(a|s) = \pi(a|s)\nabla\log\pi(a|s)] \\
&= \sum_s d_{\text{av}}^{\pi,\rho_0} \sum_a \pi(a|s)\nabla\log\pi(a|s)\tilde{Q}_{\text{av}}^\pi(s,a) \\
&= \sum_s \lim_{k\to\infty} \mathop{\mathbb{E}}_{s_0\sim\rho_0(\cdot)}\Big[P^\pi(s_k=s|s_0)\Big] \sum_a \pi(a|s)\nabla\log\pi(a|s)\tilde{Q}_{\text{av}}^\pi(s,a) \\
&\quad [\text{from ergodicity}] \\
&= \sum_s \lim_{T\to\infty} \frac{1}{T+1}\sum_{k=0}^T \mathop{\mathbb{E}}_{s_0\sim\rho_0(\cdot)}\Big[P^\pi(s_k=s|s_0)\Big] \sum_a \pi(a|s)\nabla\log\pi(a|s)\tilde{Q}_{\text{av}}^\pi(s,a) \\
&= \lim_{T\to\infty} \mathop{\mathbb{E}}_{\pi,\rho_0}\Big[\frac{1}{T+1}\sum_{k=0}^T \nabla\log\pi(a_k|s_k)\tilde{Q}_{\text{av}}^\pi(s_k,a_k)\Big].
\end{aligned}$$

B.1.2. FROM TEMPORAL PERFORMANCE DIFFERENCE

Alternatively, we can leverage the temporal performance difference formula to derive the result. Recall that $\tilde{Q}_\bullet^\pi(s,a) := Q_\bullet^\pi(s,a) - \tau\log\pi(a|s)$, $V_\bullet^\pi(s) = \mathbb{E}_{a\sim\pi}$ and $A_\gamma^\pi(s,a) := \tilde{Q}_\bullet^\pi(s,a) - V_\bullet^\pi(s)$.

Discounted reward The result follows through direct computation

$$
\begin{aligned}
\nabla J_\gamma(\pi;\rho_0) &= \lim_{\delta\theta\to 0}\sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi',\rho_0}\frac{1}{\delta\theta}\Big[A^\pi_\gamma(s_k,a_k)\Big] \\
&= \lim_{\delta\theta\to 0}\sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi',\rho_0}\frac{1}{\delta\theta}\Big[\tilde{Q}^\pi_\gamma(s_k,a_k) - \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s_k)}[\tilde{Q}^\pi_\gamma(s_k,a)]\Big] \\
&= \lim_{\delta\theta\to 0}\sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi',\rho_0}\frac{1}{\delta\theta}\Big[\mathop{\mathbb{E}}_{a_k\sim\pi'(\cdot|s_k)}[\tilde{Q}^\pi_\gamma(s_k,a_k)] - \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s_k)}[\tilde{Q}^\pi_\gamma(s_k,a)]\Big] \\
&= \lim_{\delta\theta\to 0}\sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi',\rho_0}\frac{1}{\delta\theta}\Big[\mathop{\mathbb{E}}_{a\sim\pi'(\cdot|s_k)}[\tilde{Q}^\pi_\gamma(s_k,a)] - \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s_k)}[\tilde{Q}^\pi_\gamma(s_k,a)]\Big] \\
&= \lim_{\delta\theta\to 0}\sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi',\rho_0}\Big[\sum_a \frac{\pi'(a|s_k)-\pi(a|s_k)}{\delta\theta}\tilde{Q}^\pi_\gamma(s_k,a)\Big] \\
&= \lim_{\delta\theta\to 0}\sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi',\rho_0}\Big[\sum_a \pi(a|s_k)\frac{1}{\delta\theta}\Big(\frac{\pi'(a|s_k)}{\pi(a|s_k)}-1\Big)\tilde{Q}^\pi_\gamma(s_k,a)\Big] \\
&= \lim_{\delta\theta\to 0}\sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi',\rho_0}\Big[\sum_a \pi(a|s_k)\frac{1}{\delta\theta}\log\Big(\frac{\pi'(a|s_k)}{\pi(a|s_k)}\Big)\tilde{Q}^\pi_\gamma(s_k,a)\Big] \\
&\quad \Big[\lim_{x\to 1}\frac{\log x}{x-1}=1\Big] \\
&= \sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi,\rho_0}\Big[\sum_a \pi(a|s_k)\nabla_\theta\log\pi(a|s_k)\tilde{Q}^\pi_\gamma(s_k,a)\Big] \\
&= \sum_{k=0}^{\infty}\gamma^k \mathop{\mathbb{E}}_{\pi,\rho_0}\Big[\nabla_\theta\log\pi(a_k|s_k)\tilde{Q}^\pi_\gamma(s_k,a_k)\Big].
\end{aligned}
$$

Total reward and average reward Follows similarly as the discounted case.

B.2. Unbiased Temporal PG Estimates

Recall that

$$
\begin{aligned}
G^\pi_\bullet(s,a) &:= \nabla_\theta\log\pi(a|s)\Big(Q^\pi_\bullet(s,a)-\tau\log\pi(a|s)\Big), \\
\hat{G}^\pi_\bullet(s,a) &:= \nabla_\theta\log\pi(a|s)\Big(\hat{Q}^\pi_\bullet(s,a)-\tau\log\pi(a|s)\Big).
\end{aligned}
$$

The theoretical PG in eqn. (4) becomes

$$
\nabla J_\bullet(\pi,\rho_0) := \sum_s d^\pi_\bullet(s)\sum_a \pi(a|s)G^\pi_\bullet(s,a). \tag{13}
$$

The PG implementation in eqn. (5) becomes

$$
\hat{\nabla} J_\bullet(\pi,\rho_0) := \sum_{k=0}^{T}\hat{G}^\pi_\bullet(s_k,a_k). \tag{14}
$$

We now show the connection between empirical PG implementation (Williams, 1988; 1992) and the theoretical formula (Sutton et al., 1999).

Episodic task If we solve episodic task under the total reward criteria, eqn. (5) is an unbiased estimator of the exact PG because

$$\begin{aligned}\mathop{\mathbb{E}}_{\pi,\rho_0}\left[\sum_{k=0}^{T}\hat{G}^{\pi}_{\mathrm{tot}}(s_k,a_k)\right] &= \mathop{\mathbb{E}}_{\pi,\rho_0}\left[\sum_{k=0}^{T}\sum_{s,a}\mathbb{1}(s_k=s)\mathbb{1}(a_k=a)\hat{G}^{\pi}_{\mathrm{tot}}(s_k,a_k)\right]\\ &=\sum_{s,a}\mathop{\mathbb{E}}_{\pi,\rho_0}\left[\sum_{k=0}^{T}\mathbb{1}(s_k=s)\mathbb{1}(a_k=a)\hat{G}^{\pi}_{\mathrm{tot}}(s_k,a_k)\right]\\ &=\sum_{s,a}d^{\pi,\rho_0}_{\mathrm{tot}}(s)\pi(a|s)G^{\pi}_{\mathrm{tot}}(s,a)\\ &=\nabla_\theta J_{\mathrm{tot}}(\theta).\end{aligned}$$

Continuing task We consider the asymptotic average behavior of $\sum_k \hat{G}_{\mathrm{av}}(s_k,a_k)$, that is $\lim_{T\to\infty}\frac{1}{T+1}\sum_{k=0}^{T}\hat{G}_{\mathrm{av}}(s_k,a_k)$, which is proportional to $\sum_k \hat{G}_{\mathrm{av}}(s_k,a_k)$ for every finite horizon T. Similar to the episodic task, we can derive

$$\begin{aligned}\lim_{T\to\infty}\mathop{\mathbb{E}}_{\pi,\rho_0}\left[\frac{1}{T+1}\sum_{k=0}^{T}\hat{G}^{\pi}_{\mathrm{av}}(s_k,a_k)\right] &= \lim_{T\to\infty}\mathop{\mathbb{E}}_{\pi,\rho_0}\left[\sum_{k=0}^{T}\sum_{s,a}\frac{\mathbb{1}(s_k=s)\mathbb{1}(a_k=a)}{T+1}\hat{G}^{\pi}_{\mathrm{av}}(s_k,a_k)\right]\\ &=\sum_{s,a}\lim_{T\to\infty}\mathop{\mathbb{E}}_{\pi,\rho_0}\left[\frac{\sum_{k=0}^{T}\mathbb{1}(s_k=s)\mathbb{1}(a_k=a)}{T+1}\hat{G}^{\pi}_{\mathrm{av}}(s_k,a_k)\right]\\ &=\sum_{s,a}d^{\pi,\rho_0}_{\mathrm{av}}(s)\pi(a|s)G^{\pi}_{\mathrm{av}}(s,a)\\ &=\nabla_\theta J_{\mathrm{av}}(\theta).\end{aligned}$$

Discounted total reward For the discounted reward, the discounted scaling is required, which echos back the claim made in Nota & Thomas (2020. In particular,

$$\begin{aligned}\lim_{T\to\infty}\mathop{\mathbb{E}}_{\pi,\rho_0}\left[\sum_{k=0}^{T}\gamma^k\hat{G}^{\pi}_{\gamma}(s_k,a_k)\right] &=\sum_{s,a}\lim_{T\to\infty}\mathop{\mathbb{E}}_{\pi,\rho_0}\left[\sum_{k=0}^{T}\gamma^k\mathbb{1}(s_k=s)\mathbb{1}(a_k=a)\hat{G}^{\pi}_{\gamma}(s_k,a_k)\right]\\ &=\sum_{s,a}d^{\pi,\rho_0}_{\gamma}(s)\pi(a|s)G^{\pi}_{\gamma}(s,a)\\ &=\nabla_\theta J_{\gamma}(\theta).\end{aligned}$$

C. Proof of Theorem 4.2 and 4.4 (Approximation Error Bounds)

The discounted value function is often used together with the undiscounted visitation counts (episodic tasks) or frequency (continuing tasks). While this yields problems as Nota & Thomas (2020) pointed out, we can attain equivalence when $\gamma \uparrow 1$. In *episodic tasks*, the discounted sum reduces to the total sum,

$$\lim_{\gamma\uparrow 1} d_\gamma^{\pi,\rho_0}(s) = d_{\text{tot}}^{\pi,\rho_0}(s), \qquad \lim_{\gamma\uparrow 1} V_\gamma^\pi(s) = V_{\text{tot}}(s).$$

In continuing tasks, by the Abelian theorem (Hernández-Lerma & Lasserre, 1996)[Lemma 5.3.1], the discounted sum approaches the average value

$$\lim_{\gamma\uparrow 1}(1-\gamma) d_\gamma^{\pi,\rho_0}(s) = \lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T} \Pr(s_k = s|\rho_0,\pi) \overset{(i)}{=} d_{\text{av}}^{\pi,\rho_0}(s),$$

$$\lim_{\gamma\uparrow 1}(1-\gamma) V_\gamma(s) = \lim_{T\to\infty} \mathop{\mathbb{E}}_{\pi,\rho_0}\left[\frac{1}{T+1}\sum_{k=0}^{T}\tilde{R}(s_k,a_k)\right] = J_{\text{av}}(\pi),$$

where (i) follows from the ergodicity assumption. Nevertheless, only $\gamma < 1$ can be used for real problems. In this section, we show that the errors of $d_\bullet^{\pi,\rho_0}$ and $V_\bullet^\pi$ depend on γ and the structure of the Markov transition probabilities.

C.1. Episodic Task (Theorem 4.2)

We assume that there exists an upper bound of time epochs such that the probability of entering the terminal state is nonzero.

Assumption. There exist an integer $m > 0$ and a positive real number $\alpha < 1$ such that,

$$\Pr(s_m \neq z|s_0,\pi) \leq \alpha, \quad \forall s_0 \in S,\ \pi.$$

Distribution error Let $\tilde{\rho}_0$ stand for a distribution over transient states, and $\tilde{P}^\pi$ be the corresponding transition matrix. By Assumption 4.1, the accumulated visitation counts between tm and $(t+1)m-1$ is bounded by

$$\sum_{k=tm}^{(t+1)m-1} \|\tilde{\rho}_0(\tilde{P}^\pi)^k\|_\infty \leq m\alpha^t.$$

Let $\tilde{d}_\bullet^{\pi,\rho_0}$ be the corresponding expected visitation counts on transient states, we can obtain

$$\begin{aligned}
\|\tilde{d}_{\text{tot}}^{\pi,\rho_0} - \tilde{d}_\gamma^{\pi,\rho_0}\|_\infty =& \Big\|\sum_{k=0}^{\infty}(1-\gamma^k)\rho_0(\tilde{P}^\pi)^k\Big\|_\infty \\
\leq& \sum_{k=0}^{\infty}\|(1-\gamma^k)\rho_0(\tilde{P}^\pi)^k\|_\infty \\
\leq& \sum_{t=0}^{\infty} m(1-\gamma^{tm})\alpha^t \\
=& \frac{m\alpha(1-\gamma^m)}{(1-\alpha)(1-\alpha\gamma^m)}.
\end{aligned}$$

Value error According to Assumption 4.1, the accumulated reward between stage tm and $(t+1)m-1$ is bounded by

$$\mathop{\mathbb{E}}_{\rho_0,\pi}\left[\sum_{k=tm}^{(t+1)m-1} R^\pi(s_k)\right] \leq m\alpha^t\|R^\pi\|_\infty.$$

Note that the we can equivalently write V_γ^π and V_{tot}^π as

$$\begin{aligned}
V_\gamma^\pi =& \sum_{k=0}^{\infty}\gamma^k(P^\pi)^k R^\pi, \\
V_{\text{tot}}^\pi =& \sum_{k=0}^{\infty}(P^\pi)^k R^\pi.
\end{aligned}$$

The value function gap between the total reward and the discounted reward setup can be bounded by

$$\begin{aligned}\|V_{\text{tot}}^{\pi} - V_{\gamma}^{\pi}\|_{\infty} =& \Big\| \sum_{k=0}^{\infty} (1-\gamma^k)(P^{\pi})^k R^{\pi} \Big\|_{\infty} \\ \leq & \sum_{k=0}^{\infty} \|(1-\gamma^k)(P^{\pi})^k R^{\pi}\|_{\infty} \\ \leq & \sum_{t=0}^{\infty} m(1-\gamma^{tm})\alpha^t \|R^{\pi}\|_{\infty} \\ = & \frac{m\alpha(1-\gamma^m)}{(1-\alpha)(1-\alpha\gamma^m)} \|R^{\pi}\|_{\infty}.\end{aligned}$$

C.2. Continuing Task (Theorem 4.4)

Definition C.1. The total variation norm of a signed measure $\delta(\cdot)$ is

$$\|\delta\|_{\text{TV}} := \max_{X \subset S} \sum_{s \in X} \delta(s) - \min_{X \subset S} \sum_{s \in X} \delta(s).$$

We assume that, under all policies, the Markov chain under policy π is geometrically ergodic.

Assumption. There exist positive constants C and $\beta < 1$ such that

$$\max_{s \in S} \|\rho_0 (P^{\pi})^k - d_{\text{av}}^{\pi,\rho_0}\|_{\text{TV}} \leq C\beta^k.$$

Distribution error The distribution between the discounted frequency and average frequency is bounded by

$$\begin{aligned}\|(1-\gamma)d_{\gamma}^{\pi,\rho_0} - d_{\text{av}}^{\pi,\rho_0}\|_{\text{TV}} =& \Big\| (1-\gamma) \sum_{k=0}^{\infty} \gamma^k \Big(\rho_0 (P^{\pi})^k - d_{\text{av}}^{\pi,\rho_0} \Big) \Big\|_{\text{TV}} \\ \leq & (1-\gamma) \sum_{k=0}^{\infty} \gamma^k C\beta^k = \frac{1-\gamma}{1-\gamma\beta} C.\end{aligned}$$

Value error To study the error bound for the average reward, we need the following additional definition regarding the gap between the maximum and the minimum of a vector.

Definition C.2. The span semi-norm of a vector V is

$$sp(V) = \max_{s} V(s) - \min_{s} V(s).$$

The value function for the average reward is equivalent to

$$\begin{aligned}V_{\text{av}}^{\pi} =& [(I - P^{\pi} + e d_{\text{av}}^{\pi,\rho_0})^{-1} - e d_{\text{av}}^{\pi,\rho_0}] R^{\pi} \\ = & \sum_{k=0}^{\infty} \Big[I + (P^{\pi} - e d_{\text{av}}^{\pi,\rho_0})^k \Big] R^{\pi} - e d_{\text{av}}^{\pi,\rho_0} R^{\pi} \\ \overset{(i)}{=} & \sum_{k=0}^{\infty} [(P^{\pi})^k - e d_{\text{av}}^{\pi,\rho_0}] R^{\pi},\end{aligned}$$

where (i) follows from $(P^{\pi} - e d_{\text{av}}^{\pi,\rho_0})^k = (P^{\pi})^k - e d_{\text{av}}^{\pi,\rho_0}$ for $k \geq 1$, which can be proven through induction. As $e d_{\text{av}}^{\pi,\rho_0} R^{\pi} = J_{\text{av}}(\pi)$ is a constant, we can bound the value function gap between the average reward and the discounted

reward by

$$
\begin{aligned}
sp(V^\pi_{\text{av}} - V^\pi_\gamma) =& sp\Big(\sum_{k=0}^{\infty}\Big\{(1-\gamma^k)(P^\pi)^k R^\pi - J_{\text{av}}(\pi)\Big\}\Big) \\
\leq& \sum_{k=0}^{\infty} sp\Big((1-\gamma^k)(P^\pi)^k R^\pi\Big) \\
\leq& \sum_{k=0}^{\infty}(1-\gamma^k)\beta^k sp(R^\pi) \\
=& \frac{\beta(1-\gamma)}{(1-\beta)(1-\beta\gamma)} sp(R^\pi).
\end{aligned}
$$