Is the Policy Gradient a Gradient?

Chris Nota  
College of Information and Computer Science  
University of Massachusetts Amherst  
Amherst, MA 01002  
cnota@cs.umass.edu

Philip S. Thomas  
College of Information and Computer Science  
University of Massachusetts Amherst  
Amherst, MA 01002  
pthomas@cs.umass.edu

Abstract

The policy gradient theorem [Sutton et al., 2000] describes the gradient of the expected discounted return with respect to an agent’s policy parameters. However, most policy gradient methods do not use the discount factor in the manner originally prescribed, and therefore do not optimize the discounted objective. It has been an open question in RL as to which, if any, objective they optimize instead [Thomas, 2014]. We show that the direction followed by these methods is not the gradient of any objective, and reclassify them as semi-gradient methods with respect to the undiscounted objective. Further, we show that they are not guaranteed to converge to a locally optimal policy, and construct a counterexample where they will converge to the globally pessimal policy with respect to both the discounted and undiscounted objectives.

1 Introduction

Reinforcement learning (RL) is a subfield of machine learning in which computational agents learn to maximize a numerical reward signal through interaction with their environment. Policy gradient methods encode an agent’s behavior as a parameterized stochastic policy and update the policy parameters according to an estimate of the gradient of the expected sum of rewards (the expected return) with respect to those parameters. In practice, estimating the effect of a particular action on the return that follows can be difficult, so almost all state-of-the-art implementations instead consider an exponentially discounted sum of rewards (the discounted return), which causes the agent to primarily consider the short-term effects of its actions. However, they do not do so in the manner described by the original policy gradient theorem [Sutton et al., 2000]. It is an open question in RL as to whether these algorithms optimize a different, related objective [Thomas, 2014]. In this paper, we answer this question by showing that most policy gradient methods, including state-of-the-art methods, do not follow the gradient of any objective, and are therefore better understood as semi-gradient methods with respect to the undiscounted objective.

The analysis in this paper applies to all state-of-the-art policy gradient methods known to the authors. This includes the published versions of methods such as A2C/A3C [Mnih et al., 2016], TRPO [Schulman et al., 2015], PPO [Schulman et al., 2017], DDPG [Silver et al., 2014], ACKTR [Wu et al., 2017], ACER [Wang et al., 2016], and soft actor-critic [Haarnoja et al., 2018], as well as their publicly available implementations [Dhariwal et al., 2017]. We note however that the concerns of this paper long predate “deep” RL. The observation that most policy gradient methods do not conform to the policy gradient theorem presented by [Sutton et al., 2000] dates at least back to the work of Kakade [2001], while the explicit question addressed by this paper as to whether these algorithms estimate the gradient of any objective dates at least to the work of Thomas [2014]. The purpose of this paper is not to criticize any particular algorithm, or the policy gradient approach in general, nor to propose specific new algorithms, but rather to improve our understanding of the mathematical foundations of policy gradient methods.

Preprint. Under review.
2 Notation

In RL, the agent learns through interactions with an environment, which is expressed mathematically as a Markov decision process (MDP). An MDP is a tuple, \((S, A, P, R, d_0, \gamma)\), where \(S\) is the set of possible states of the environment, \(A\) is the set of actions available to the agent, \(P: S \times A \times S \rightarrow [0, 1]\) is a transition function that determines the probability of transitioning between states given an action, \(R: S \times A \rightarrow \mathbb{R}\) is the expected reward from taking an action in a particular state, \(d_0: S \rightarrow [0, 1]\) is the initial state distribution, and \(\gamma \in [0, 1]\) is the discount factor which decreases the utility of rewards received in the future.

In the episodic setting, interactions with the environment are broken into episodes. Each episode is further broken into individual timesteps. At each timestep, \(t\), the agent observes a state, \(S_t\), takes an action, \(A_t\), transitions to a new state, \(S_{t+1}\), and receives a reward, \(R_t\). The first timestep is \(0\). For finite horizon MDPs, the maximum timestep is given by \(h\). If the agent reaches a special state, called the terminal absorbing state, prior to the final timestep, it stays in this state and receives a reward of \(0\) at each timestep until \(h\), after which the process terminates.

A policy, \(\pi: S \times A \rightarrow [0, 1]\), determines the probability that an agent will choose an action in a particular state. A parameterized policy, \(\pi^\theta\), is a policy that is determined by some parameter vector, \(\theta\), which may be the weights in a neural network, values in a tabular representation, etc. The compatible features of a parameterized policy represent how \(\theta\) may be changed in order to make a particular action, \(a \in A\), more likely in a particular state, \(s \in S\), and are defined as \(\psi(s, a) := \frac{\partial}{\partial \theta} \ln \pi^\theta(s, a)\).

The value function, \(V^\theta: S \rightarrow \mathbb{R}\), represents the expected discounted sum of rewards when starting in a particular state under policy \(\pi^\theta\); that is, \(\forall t, V^\theta(s) := \mathbb{E}[\sum_{k=0}^{h-1} \gamma^k R_{t+k} | S_t = s, \theta]\), where conditioning on \(\theta\) indicates that \(\forall t, A_t \sim \pi^\theta(S_t, \cdot)\). The action-value function, \(Q^\theta: S \times A \rightarrow \mathbb{R}\), is similar, but also considers the action taken; that is, \(\forall t, Q^\theta(s, a) := \mathbb{E}[\sum_{k=0}^{h-1} \gamma^k R_{t+k} | S_t = s, A_t = a, \theta]\). The advantage function is the difference between the action-value function and the (state) value function: \(A^\theta(s, a) := Q^\theta(s, a) − V^\theta(s)\).

The objective of an RL agent is to maximize some function of its parameters, \(J(\theta)\). In the episodic setting, the two most commonly stated objectives are the discounted objective, \(J_\gamma(\theta) = \mathbb{E}[\sum_{t=0}^{h-1} \gamma^t R_t | \theta]\), and the undiscounted objective, \(J(\theta) = \mathbb{E}[\sum_{t=0}^{h-1} R_t | \theta]\). The discounted objective has some convenient mathematical properties, but it corresponds to few real-world tasks. Sutton and Barto [2018] have even argued for its depreciation. The undiscounted objective is a more intuitive match for most tasks. Both objectives and their respective gradients are discussed in this paper.

3 The policy gradient theorem

The policy gradient theorem [Sutton et al. 2000] was originally presented for two objectives: the average reward objective for the continuing setting [Mahadevan 1996] and the discounted objective \((J_\gamma)\) for the episodic setting. The episodic setting is usually more popular, as it is better suited to the types of tasks that RL researchers typically use for evaluation (e.g., many classic control tasks, and the Arcade Learning Environment [Bellemare et al. 2013]). The policy gradient, \(\nabla J_\gamma(\theta)\), tells us how to modify \(\theta\) in order to increase \(J_\gamma\), and is given by:

\[
\nabla J_\gamma(\theta) = \mathbb{E} \left[ \sum_{t=0}^{h-1} \gamma^t \psi^\theta(S_t, A_t) Q^\theta(S_t, A_t) \right] \frac{\partial}{\partial \theta}.
\]

Because \(\nabla J_\gamma\) is the true gradient of the undiscounted objective, algorithms that follow unbiased estimates of \([1]\) are given the standard guarantees of stochastic gradient descent (namely, that given an appropriate step-size schedule and smoothness assumptions, convergence to a local optimum is almost sure [Bertsekas and Tsitsiklis 2000]). However, researchers typically are not interested in the discounted objective, but rather the undiscounted objective. As a result, most conventional “policy gradient” algorithms (PPO, A3C, etc.) instead directly or indirectly estimate the expression:
\[ \nabla J^*(\theta) = E \left[ \sum_{t=0}^{h-1} \psi^\theta(S_t, A_t) Q^\theta_\gamma(S_t, A_t) \right] . \] (2)

Note that the difference between \( \nabla J_\gamma(\theta) \) and \( \nabla J^*(\theta) \) is that the latter has dropped the \( \gamma^t \) term \cite{Thomas:2014}. However, notice that this new expression is not the gradient of \( J \) either, as it uses the discounted action-value function, \( Q^\theta_\gamma \). We label this update direction \( \nabla J^*(\theta) \) because it is no longer the gradient of \( J_\gamma \) or \( J \), but is rather the gradient of some hypothetical, related objective, \( J^* \). The asterisk character, *, was chosen to represent the ambiguity of its existence. \cite{Thomas:2014} attempted to construct \( J^* \), but was only able to do so for an impractically restricted setting wherein the agent’s actions do not impact state transitions. Its value in the general case was left as an open question. We will later show that the hypothetical objective, \( J^* \), does not exist when the impractical restrictions imposed by \cite{Thomas:2014} are removed. The result is that algorithms based on (2), which are commonly referred to as “policy gradient” algorithms, cannot be interpreted as gradient algorithms with respect to any objective. In Section 5, we show that these methods are better understood as semi-gradient methods relative to the undiscounted objective. In the next section, we prove an intermediate result relating the undiscounted objective to the discounted value function.

4 Decomposing the undiscounted objective

We begin our analysis by presenting a lemma that provides a new view of the undiscounted objective by allowing us to express it in terms of the discounted value function, for any value of \( \gamma \). Writing the objective in this new form is the first step towards uncovering its relationship with \( \nabla J^* \). This is necessary, as \( \nabla J^* \) is defined in terms of the discounted action-value function.

**Lemma 1.** The undiscounted objective is related to \( V^\theta_\gamma(s) \) for all \( \gamma \) by:

\[ E \left[ \sum_{t=0}^{h-1} R_t \right] = \sum_{s \in S} d^\theta_\gamma(s) V^\theta_\gamma(s) , \]

where \( d^\theta_\gamma(s) \) is the unnormalized, weighted state distribution given by:

\[ d^\theta_\gamma(s) := d_0(s) + (1 - \gamma) \sum_{t=1}^{h-1} \Pr(S_t = s|\theta) . \]

**Proof.** See appendix.

Taking the gradient of this expression gives us a corresponding new form of the policy gradient. The first step of the derivation is given here, as it deepens our intuition as to the relationship between the discounted value function and the undiscounted objective:

\[ \nabla J(\theta) = \frac{\partial}{\partial \theta} E \left[ \sum_{t=0}^{h-1} R_t \right] = \sum_{s \in S} \frac{\partial}{\partial \theta} d^\theta_\gamma(s) V^\theta_\gamma(s) = \sum_{s \in S} \underbrace{d^\theta_\gamma(s) \frac{\partial}{\partial \theta} V^\theta_\gamma(s)}_{\text{value gradient}} + \sum_{s \in S} \underbrace{V^\theta_\gamma(s) \frac{\partial}{\partial \theta} d^\theta_\gamma(s)}_{\text{distribution gradient}} . \] (3)

We see that there are two components to this new form of the gradient. The first component optimizes the discounted value function over the current distribution of states. The second term takes into account the change in the weighted distribution over states induced by changes in the policy. Note that in the \( \gamma = 1 \) case, \( \partial d^\theta_\gamma(s) / \partial \theta \) is zero for all \( s \), and we are left with only the first term. Also note that under this formulation, \( \gamma \) is strictly a hyperparameter, and does not impact the optimal policy. In the next section, we relate this expression to \( \nabla J^* \).
5 Policy semi-gradient methods

In this section, we show that when discounting is used, most policy “gradient” methods are actually policy semi-gradient methods, and that they are not truly gradient methods with respect to any objective. First, we define the term semi-gradient. If a gradient of a function can be written with two terms, the semi-gradients are each of these terms; they represent the two “halves” of the gradient. The semi-gradients of a function are not unique, as they entirely depend on how the gradient is written. Nevertheless, they can provide useful intuition for the behavior of algorithms. The term was coined by [Sutton 2015] in order to describe the relationship between temporal difference (TD) learning methods and the mean-squared Bellman error. Similarly, we use the term to describe the relationship between $\nabla J^*$ and the undiscounted objective. In the following lemma, we show that $\nabla J^*$ may be written as the value gradient term in (3), from which the proof that it is a semi-gradient is immediate:

**Lemma 2.** The conventional policy update, $\nabla J^*(\theta)$, can be written as:

$$\nabla J^*(\theta) = \sum_{s \in S} d^\theta_s(s) \frac{\partial}{\partial \theta} V^\theta_\gamma(s).$$

(4)

**Proof.** See appendix.

**Corollary 1.** $\nabla J^*$ is a semi-gradient of $J$.

Proof. The right-hand side of (4) is the first term of the gradient, $\nabla J$, in (3). The corollary follows from the definition of a semi-gradient.

In a sense, it is remarkable that this approach to policy gradients works at all. Value function estimates are typically learned using semi-gradient TD methods [Sutton and Barto 2018]. The policy is typically learned using an approximation of $J^*(\theta)$, which we have shown is a semi-gradient. Both “actor” and “critic” follow semi-gradients, and yet implementations of actor-critic methods [Konda and Tsitsiklis 2000] that follow these update directions are able to learn effective policies. An open question in RL is whether or not these algorithms are following some related objective, and if this is the cause for their effectiveness [Thomas 2014]. We show that this is not the case.

**Theorem 1.** There does not exist any continuous objective, $J$, such that $\nabla J^*$ is its gradient. That is, for all $\gamma < 1$:

$$\neg \exists J \in C(\mathbb{R}) : \nabla J^*(\theta) = E \left[ \sum_{t=0}^{h-1} \psi^\theta(S_t, A_t) Q^\theta_\gamma(S_t, A_t) \right] \theta,$$

where $C(\mathbb{R})$ is the space of continuous functions.

Proof. We present a proof by contradiction. Assume that there exists some $J^*$ whose gradient is $\nabla J^*(\theta)$, and that $\theta_i$ and $\theta_j$ are parameters of the policy, $\pi^\theta$. If $J^*$ is a continuous function, then it is also true that its second derivative is symmetric:

$$\frac{\partial}{\partial \theta_i} \left( \frac{\partial}{\partial \theta_j} J^*(\theta) \right) = \frac{\partial}{\partial \theta_j} \left( \frac{\partial}{\partial \theta_i} J^*(\theta) \right).$$

However, it is easy to show that this is not true algebraically using Lemma 2

$$\frac{\partial^2 J^*(\theta)}{\partial \theta_i \partial \theta_j} = \frac{\partial}{\partial \theta_i} \left( \sum_{s \in S} d^\theta_s(s) \frac{\partial}{\partial \theta_j} V^\theta_\gamma(s) \right) = \sum_{s \in S} d^\theta_s(s) \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} V^\theta_\gamma(s) \right] \text{ symmetric} + \sum_{s \in S} \frac{\partial}{\partial \theta_i} d^\theta_s(s) \left[ \frac{\partial}{\partial \theta_j} V^\theta_\gamma(s) \right] \text{ asymmetric} \ .$$

(5)

Thus, we immediately arrive at an apparent contradiction. However, to formally complete the proof, we must provide a counterexample where the latter term is asymmetric. Such a counterexample is provided in Figure 1.
Figure 1: A counterexample wherein the derivative of $\nabla J^*$ is asymmetric for all $\gamma < 1$, necessary for the proof of Theorem 1. The agent begins in $s_1$, and may choose between actions $a_1$ and $a_2$, each of which produces deterministic transitions. The rewards are 0 everywhere except when the agent chooses $a_2$ in state $s_2$, which produces a reward of $+1$. The policy is assumed to be tabular over states and actions, with one parameter, $\theta_1$, determining the policy in $s_1$, and a second parameter, $\theta_2$, determining the policy in $s_2$ (e.g., the policy may be determined by the sigmoid function, $\pi(s_1, a_1) = \sigma(\theta_1)$). For intuition, consider $\gamma = 0$. (5) is asymmetric in this case because $\theta_1$ affects the state distribution, but not the value function, whereas $\theta_2$ affects the value function but not the state distribution. Therefore, the second term in (5) is non-zero when $i = 1$ and $j = 2$, and zero when $i = 2$ and $j = 1$. We show the full computation in the appendix.

6 Non-optimality of semi-gradient methods

One may wonder, why should we be concerned at all that policy gradient methods are following a semi-gradient rather than a gradient? After all, the same may be said of TD [Sutton, 2015], which remains in wide use. First, there is a large body of literature [Baird, 1995, Sutton et al., 2009, Liu et al., 2015, Mahmood et al., 2015, Liu et al., 2016] suggesting that many researchers are concerned regarding TD; even in the subfield of “deep” RL, the concept of a target network [Mnih et al., 2015] was introduced to to address the instability introduced by the use of a semi-gradient update. Second, semi-policy gradient methods suffer from weaknesses that semi-gradient TD methods do not.

Consider that semi-gradient TD is a semi-gradient of the mean-squared Bellman error [Baird, 1995, Sutton, 2015], whereas the policy semi-gradient is a semi-gradient of the undiscounted objective. In the tabular setting, semi-gradient TD will converge to the minimum of the mean-squared Bellman error [Jaakkola et al., 1994], but policy semi-gradient methods will often fail to converge to the maximum of the undiscounted objective—e.g., they may instead converge to the objective of the discounted objective (see Figure 2). There is therefore an asymmetry between semi-gradient TD and policy semi-gradient methods, in that the former converges to the objective of which it is a semi-gradient, while the latter does not. However, this situation is not particularly alarming, as it is well understood that the optimal policy under the discounted objective approximates the optimal policy under the undiscounted objective [Marbach and Tsitsiklis, 2001, Weaver and Tao, 2001, Kakade, 2001, Thomas, 2014]. The choice of the discount factor for policy semi-gradient methods in the tabular setting may therefore be considered in terms of a bias-variance tradeoff.

When a function approximator is used, a worse problem emerges that is analogous to divergence in semi-gradient TD methods. With a linear function approximator, semi-gradient TD methods only diverge in the off-policy setting [1] for instance, in Baird’s famous counterexample [Baird, 1995]. However, the following problem may emerge for policy semi-gradient methods even in the on-policy setting. It is not possible for policies to diverge in quite the same way as value estimations, as the probability of choosing an action is bounded by $[0, 1]$. The worst possible outcome is convergence to the pessimal policy within the given parameterization [2]. One may ask, pessimal with respect to which objective? In Figure 3, we give an example wherein policy semi-gradient methods will converge to the pessimal policy with respect to both the discounted and undiscounted objectives. While the behavior of policy semi-gradient methods described in Figure 2 may be justified as a myopic preference, the same justification cannot be applied to the case shown in Figure 3.

---

1 Assuming the typical bootstrapping approach is used, thus completing the “deadly triad” of function approximation, bootstrapping, and off-policy training [Sutton and Barto, 2018].

2 Convergence to a limit cycle is problematic as well, but preferable to the pessimal solution.
Figure 2: An example where semi-gradient methods using the specified value of $\gamma$ will produce the optimal policy with respect to the discounted objective but the pessimal behavior with respect to the undiscounted objective. The agent starts in state $s_1$, and can achieve a reward of $+1$ by transitioning from $s_1$ to $s_2$, and a reward of $+2$ by transitioning from $s_6$ to $s_7$. In order to maximize the undiscounted return, the agent should choose $a_2$. However, the discounted return is higher from choosing $a_1$. Thus, semi-gradient methods will choose $a_1$. If the researcher is concerned with the undiscounted objective, as we argue is generally the case, this result is problematic. Choosing a larger value of $\gamma$ trivially fixes the problem in this particular example, but for any value of $\gamma < 1$, similar problems will arise given a sufficiently long horizon, and thus the problem is not truly eliminated.

Figure 3: An example where semi-gradient methods will produce the pessimal policy with respect to both the discounted and undiscounted objectives. In this formulation, there is a single policy parameter, $\theta$, so the agent must execute the same policy in every state. It is difficult to justify any solution other than to always choose $a_1$. If the agent is completely myopic, then $a_1$ gives the superior immediate reward of $+1$ in the starting state. If the agent is somewhat farther-sighted, then always choosing $a_1$ will eventually result in the $+100$ reward. Choosing $a_2$ provides no benefit with respect to either objective. Consider three updates: (1) the policy gradient update given by Sutton et al. [2000], (2) the gradient of the undiscounted objective, and (3) the semi-gradient update that drops the $\gamma^t$ term. (1) and (2) will produce the policy that always chooses $a_1$. (3) will produce a policy that always chooses $a_2$, a solution which appears to be strictly inferior by any reasonable metric. Again, while we chose $\gamma = 0$ for simplicity, similar examples can be produced in long-horizon problems, with “reasonable” settings of $\gamma$, such as 0.99 or higher. One may also ask if sharing a policy between states is contrived, but such a situation occurs reliably when under partial observability or when a function approximator is used.

7 Conclusions

We have shown in Theorem 1 that most policy gradient methods, including state-of-the-art algorithms, do not follow the gradient of any objective function due to the way the discount factor is used. By Corollary 1 we argued that policy gradient methods are therefore better understood as semi-gradient methods. Finally, we showed that policy semi-gradients are not guaranteed to converge to a locally optimal policy (Figure 2), and in some cases even converge to the pessimal policy (Figure 3).

References

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In Machine Learning Proceedings 1995, pages 30–37. Elsevier, 1995.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research, 47: 253–279, 2013.

D. P. Bertsekas and J. N. Tsitsiklis. Gradient convergence in gradient methods with errors. SIAM J. Optim., 10:627–642, 2000.
Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. Openai baselines. https://github.com/openai/baselines, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. arXiv preprint arXiv:1801.01290, 2018.

Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. Convergence of stochastic iterative dynamic programming algorithms. In Advances in neural information processing systems, pages 703–710, 1994.

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

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In Advances in neural information processing systems, pages 1008–1014, 2000.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient td algorithms. In Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, pages 504–513. AUAI Press, 2015.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Proximal gradient temporal difference learning algorithms. In IJCAI, pages 4195–4199, 2016.

Sridhar Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. Machine learning, 22(1-3):159–195, 1996.

A Rupam Mahmood, Huizhen Yu, Martha White, and Richard S Sutton. Emphatic temporal-difference learning. arXiv preprint arXiv:1507.01569, 2015.

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

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529, 2015.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In International conference on machine learning, pages 1928–1937, 2016.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In International Conference on Machine Learning, pages 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In ICML, 2014.

Richard S Sutton. Introduction to reinforcement learning with function approximation. In Tutorial at the Conference on Neural Information Processing Systems, 2015.

Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 2018.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In Advances in neural information processing systems, pages 1057–1063, 2000.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 993–1000. ACM, 2009.
Philip Thomas. Bias in natural actor-critic algorithms. In *International Conference on Machine Learning*, pages 441–448, 2014.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016.

Lex Weaver and Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 538–545. Morgan Kaufmann Publishers Inc., 2001.

Yuhuai Wu, Elman Mansimov, Roger B Grosse, Shun Liao, and Jimmy Ba. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. In *Advances in neural information processing systems*, pages 5279–5288, 2017.
Appendix

Lemma 1. The undiscounted objective is related to $V_{\gamma}^\theta(s)$ for all $\gamma$ by:

$$
E \left[ \sum_{t=0}^{h-1} R_t \bigg| \theta \right] = \sum_{s \in S} d_{\gamma}^\theta(s) V_{\gamma}^\theta(s),
$$

where $d_{\gamma}^\theta$ is the unnormalized, weighted state distribution given by:

$$
d_{\gamma}^\theta(s) := d_0(s) + (1 - \gamma) \sum_{t=1}^{h-1} \Pr(S_t = s|\theta).
$$

Proof. Consider that for any $\gamma$ and $t$, we can rearrange the Bellman equation:

$$
E[R_t | \theta] = E[V_{\gamma}^\theta(S_t) - \gamma V_{\gamma}^\theta(S_{t+1}) | \theta].
$$

This allows us to rewrite the objective:

$$
\forall \gamma, E \left[ \sum_{t=0}^{h-1} R_t \bigg| \theta \right] = E \left[ \sum_{t=0}^{h-1} (V_{\gamma}^\theta(S_t) - \gamma V_{\gamma}^\theta(S_{t+1})) \bigg| \theta \right] = E \left[ V_{\gamma}^\theta(S_0) + \sum_{t=1}^{h-1} (V_{\gamma}^\theta(S_t) - \gamma V_{\gamma}^\theta(S_{t+1}) - \gamma V_{\gamma}^\theta(S_0)) \bigg| \theta \right] = E \left[ V_{\gamma}^\theta(S_0) + (1 - \gamma) \sum_{t=1}^{h-1} V_{\gamma}^\theta(S_t) \bigg| \theta \right] = \sum_{s \in S} \Pr(S_0 = s) V_{\gamma}^\theta(s) + (1 - \gamma) \sum_{t=1}^{h-1} \Pr(S_t = s|\theta) V_{\gamma}^\theta(s) = \sum_{s \in S} d_{\gamma}^\theta(s) V_{\gamma}^\theta(s).
$$

Lemma 2. The conventional policy update, $\nabla J^*(\theta)$, can be written as:

$$
\nabla J^*(\theta) = \sum_{s \in S} d_{\gamma}^\theta(s) \frac{\partial}{\partial \theta} V_{\gamma}^\theta(s). \tag{6}
$$

Proof. We begin by hypothesizing that $\nabla J^*(\theta)$ takes a form similar to (6), given some time-dependent weights, $w(t)$, on each term in the state distribution. That is, we hypothesize that (6) holds for some $d_{\gamma}^\theta$:

$$
d_{\gamma}^\theta(s) = \sum_{t=0}^{h-1} w(t) \Pr(S_t = s|\theta).
$$

We then must show that Equation (6) is equal to $\nabla J^*(\theta)$, for some choice of $w(t)$, and then derive the satisfying choice of $w(t)$. Sutton et al. [2000] established a lemma stating:

$$
\frac{\partial}{\partial \theta} V_{\gamma}^\theta(s) = \sum_{k=0}^{h-1} \sum_{x \in S} \gamma^k \Pr(S_{t+k} = x|S_t = s, \theta) \sum_{a \in A} Q_{\gamma}^\theta(x, a) \frac{\partial \pi^\theta(x, a)}{\partial \theta}. \tag{7}
$$

Applying (7) to (6) gives us:
\[
\sum_{s \in S} d^t_s(s) \frac{\partial}{\partial \theta} V^\theta_{\gamma}(s) = \sum_{s \in S} \sum_{t=0}^{h-1} w(t) \Pr(S_t = s|\theta) \sum_{k=0}^{h-1} \gamma^k \Pr(S_{t+k} = x|S_t = s, \theta) \sum_{a \in A} Q^\theta_{\gamma}(x, a) \frac{\partial \pi^\theta(x, a)}{\partial \theta} \\
= \sum_{s \in S} \sum_{t=0}^{h-1} \sum_{k=0}^{h-1} \sum_{x \in S} \sum_{a \in A} w(t) \gamma^k \Pr(S_t = s|\theta) \Pr(S_{t+k} = x|S_t = s, \theta) Q^\theta_{\gamma}(x, a) \frac{\partial \pi^\theta(x, a)}{\partial \theta} \\
= \sum_{s \in S} \sum_{t=0}^{h-1} \sum_{k=0}^{h-1} \sum_{x \in S} \sum_{a \in A} w(t) \gamma^k \Pr(S_t = s|\theta) \Pr(S_{t+k} = x|S_t = s, \theta) Q^\theta_{\gamma}(x, a) \pi^\theta(x, a) \frac{\partial \ln(\pi^\theta(x, a))}{\partial \theta} \\
= \sum_{s \in S} \sum_{t=0}^{h-1} \sum_{k=0}^{h-1} \sum_{x \in S} \sum_{a \in A} w(t) \gamma^k \Pr(S_t = s|\theta) \Pr(S_{t+k} = x|S_t = s, \theta) Q^\theta_{\gamma}(x, a) \psi(x, a), \\
\]

since \( \Pr(A_{t+k} = a|S_{t+k} = x, \theta) = \Pr(A_{t+k} = a|S_{t+k} = x, S_t = s, \theta) \) and by the law of total probability. Continuing, starting with the fact that \( \Pr(S_{t+k} = x|\theta) \Pr(A_{t+k} = a|S_{t+k} = x, \theta) = \Pr(S_{t+k} = x, A_{t+k} = a|\theta) \), we have that:

\[
\sum_{s \in S} d^t_s(s) \frac{\partial}{\partial \theta} V^\theta_{\gamma}(s) \\
= \sum_{t=0}^{h-1} \sum_{k=0}^{h-1} \sum_{x \in S} \sum_{a \in A} w(t) \gamma^k \Pr(S_{t+k} = x, A_{t+k} = a|\theta) Q^\theta_{\gamma}(x, a) \psi(x, a) \\
= \sum_{t=0}^{h-1} \sum_{k=0}^{h-1} \mathbb{E} \left[ w(t) \gamma^k Q^\theta_{\gamma}(S_{t+k}, A_{t+k}) \psi(S_{t+k}, A_{t+k}) | \theta \right] \\
= \sum_{t=0}^{h-1} \sum_{i=t}^{h-1+k} \mathbb{E} \left[ w(t) \gamma^{i-k} Q^\theta_{\gamma}(S_i, A_i) \psi(S_i, A_i) | \theta \right],
\]

by substitution of the variable \( i = t + k \). Since the value of the terminal state is 0, for \( i \geq h \), we have that \( Q^\theta_{\gamma}(S_i, A_i) = 0 \), and thus the sum over \( i \) can stop at \( h - 1 \) rather than \( h - 1 + t \). Continuing, we can move the summation inside the expectation and reorder the summation:

\[
\sum_{s \in S} d^t_s(s) \frac{\partial}{\partial \theta} V^\theta_{\gamma}(s) = \mathbb{E} \left[ \sum_{t=0}^{h-1} \sum_{i=t}^{h-1} w(t) \gamma^{i-t} Q^\theta_{\gamma}(S_i, A_i) \psi(S_i, A_i) | \theta \right] \\
= \mathbb{E} \left[ \sum_{t=0}^{h-1} w(t) \gamma^{i-t} Q^\theta_{\gamma}(S_i, A_i) \psi(S_i, A_i) | \theta \right] \\
= \mathbb{E} \left[ \sum_{i=0}^{h-1} Q^\theta_{\gamma}(S_i, A_i) \psi(S_i, A_i) \sum_{t=0}^{i} w(t) \gamma^{i-t} | \theta \right].
\]
In order to derive $\nabla J^*$, we simply need to choose a $w(t)$ such that $\forall i: \sum_{t=0}^i w(t)\gamma^{i-t} = 1$. This is satisfied by the choice: $w(t) = 1$ if $t = 0$, and $1 - \gamma$ otherwise. This trivially holds for $i = 0$, as $w(0)\gamma^0 = (1)(1) = 1$. For $i > 0$:

$$\sum_{t=0}^i w(t)\gamma^{i-t} = w(0)\gamma^i + \sum_{t=1}^i w(t)\gamma^{i-t}$$

$$= \gamma^i + \sum_{t=1}^i (1 - \gamma)\gamma^{i-t}$$

$$= \gamma^i + \sum_{t=1}^i (\gamma^{i-t} - \gamma^{i-t+1})$$

$$= \gamma^i + \gamma^{i-i} - \gamma^{i-1+1}$$

$$= \gamma^i + 1 - \gamma^i$$

$$= 1.$$

Thus, for this choice of $w(t)$:

$$\sum_{s \in S} d_\theta^\gamma(s) \frac{\partial}{\partial \theta} V_\gamma^\theta(s) = \mathbb{E} \left[ \sum_{t=0}^{h-1} Q_\theta^\gamma(S_t, A_t) \psi(S_t, A_t) \right]$$

$$= \nabla J^*(\theta).$$

Finally, we see that this choice of $w(t)$ also gives us the correct $d_\gamma^\theta$:

$$d_\theta^\gamma(s) = \sum_{t=0}^{h-1} w(t) \Pr(S_t = s|\theta) = w(0) \Pr(S_0 = s|\theta) + \sum_{t=1}^{h-1} w(t) \Pr(S_t = s|\theta)$$

$$= d_0(s) + (1 - \gamma) \sum_{t=1}^{h-1} \Pr(S_t = s|\gamma).$$

\[\square\]

**Theorem 1.** There does not exist any continuous objective, $J$, such that $\nabla J^*$ is its gradient. That is, for all $\gamma < 1$:

$$\neg \exists J \in C(\mathbb{R}) : \nabla J^*(\theta) = \mathbb{E} \left[ \sum_{t=0}^{h-1} \psi^\theta(S_t, A_t) Q_\theta^\gamma(S_t, A_t) \right],$$

where $C(\mathbb{R})$ is the space of continuous functions.

**Proof.** We present a proof by contradiction. Assume that there exists some $J^*$ whose gradient is $\nabla J^*(\theta)$, and that $\theta_i$ and $\theta_j$ are parameters of the policy, $\pi^\theta$. If $J^*$ is a continuous function, then it is also true that its second derivative is symmetric:

$$\frac{\partial}{\partial \theta_i} \left( \frac{\partial}{\partial \theta_j} J^*(\theta) \right) = \frac{\partial}{\partial \theta_j} \left( \frac{\partial}{\partial \theta_i} J^*(\theta) \right).$$

However, it is easy to show that this is not true algebraically using Lemma.
we must show a case where the summation produces an asymmetry. Such a case is shown in Figure

We consider the case where we do not introduce any bias due to the use of function approximation. There are two parameters, \( \theta_1 \) and \( \theta_2 \), which determine the policy in states \( s_1 \) and \( s_2 \) respectively. The probability of choosing action \( a_2 \) in state \( s_1 \) by \( \pi(s_1, a_2) = \sigma(\theta_1) \), where \( \sigma \) is the sigmoid function, \( \sigma(x) = e^x/(e^x + 1) \), and the probability of choosing \( a_1 \) is given by \( \pi(s_1, a_1) = 1 - \sigma(\theta_1) \). Finally, we initialize the policy to \( \theta_1 = \theta_2 = 0 \).

Consider the partial derivatives of \( d^\gamma_{s_1} \) and \( V^\gamma_{s_1} \). First, consider \( s_1 \). The agent cannot affect the likelihood of being in \( s_1 \) at time 0, so \( \partial d^\gamma_{s_1}(s_1)/\partial \theta = 0 \). Additionally, because there are no immediately rewards available and \( \gamma = 0 \), the value function and its gradient are zero at \( s_1 \) regardless of the policy parameters: \( V^\gamma_{s_1}(s_1) = \partial d^\gamma_{s_1}(s_1)/\partial \theta = 0 \). Next, consider \( s_2 \). The probability of entering \( s_2 \) is proportional to the probability of choosing \( a_2 \) in \( s_1 \). At \( \theta_1 = 0 \):

\[
\frac{\partial}{\partial \theta_1} d^\gamma_{s_1}(s_2) = \frac{\partial}{\partial \theta_1} \left[ d_0(s_2) + (1 - \gamma) \Pr(S_1 = s_2|\theta) \right] = \frac{\partial}{\partial \theta_1} \Pr(S_1 = s_2|\theta)
\]

\[
= \frac{\partial}{\partial \theta_1} \pi^\theta(s_0, a_2) = \frac{\partial}{\partial \theta_1} \sigma(\theta_1) = \frac{1}{4}.
\]

On the other hand, the value at \( s_2 \) does not depend on \( \theta_1 \), so \( \partial V^\gamma_{s_1}(s_2)/\partial \theta_1 = 0 \). Finally, consider \( s_2 \) and \( \theta_2 \). \( \theta_1 \) and \( \theta_2 \) cannot influence \( d^\gamma_{s_1}(s_2) \), as the action taken in \( s_2 \) has no effect on the state distribution. However:

\[
V^\gamma_{s_1}(s_2) = \pi(s_2, a_1)q^\gamma_{s_2}(s_2, a_1) + \pi(s_2, a_2)q^\gamma_{s_2}(s_2, a_2) = \pi(s_2, a_1)(0) + \pi(s_2, a_2)(1) = \pi(s_2, a_2),
\]

so:

\[
\frac{\partial}{\partial \theta_2} V^\gamma_{s_1}(s_2) = \frac{\partial}{\partial \theta_2} \pi(s_2, a_2) = \frac{\partial}{\partial \theta_2} \sigma(\theta_2) = \frac{1}{4}.
\]

Putting it all together:
\[
\sum_{s \in S} \frac{\partial}{\partial \theta_1} d_\gamma^\theta (s) \frac{\partial}{\partial \theta_2} V_\gamma^\theta (s) = \frac{\partial}{\partial \theta_1} d_\gamma^\theta (s_1) \frac{\partial}{\partial \theta_2} V_\gamma^\theta (s_1) + \frac{\partial}{\partial \theta_1} d_\gamma^\theta (s_2) \frac{\partial}{\partial \theta_2} V_\gamma^\theta (s_2)
\]
\[
= (0)(0) + \left( \frac{1}{4} \right) \left( \frac{1}{4} \right)
\]
\[
= \frac{1}{16}
\]
\[
\sum_{s \in S} \frac{\partial}{\partial \theta_2} d_\gamma^\theta (s) \frac{\partial}{\partial \theta_1} V_\gamma^\theta (s) = \frac{\partial}{\partial \theta_2} d_\gamma^\theta (s_1) \frac{\partial}{\partial \theta_1} V_\gamma^\theta (s_1) + \frac{\partial}{\partial \theta_2} d_\gamma^\theta (s_2) \frac{\partial}{\partial \theta_1} V_\gamma^\theta (s_2)
\]
\[
= (0)(0) + (0)(0)
\]
\[
= 0.
\]

We have shown that in this instance the second derivative is not symmetric. Therefore, we conclude that no general form of \( J^* \) exists. \( \square \)