Provably Convergent Off-Policy Actor-Critic with Function Approximation

Shangtong Zhang
University of Oxford
shangtong.zhang@cs.ox.ac.uk

Bo Liu
Auburn University
boliu@auburn.edu

Hengshuai Yao
Huawei Technologies
hengshuai.yao@huawei.com

Shimon Whiteson
University of Oxford
shimon.whiteson@cs.ox.ac.uk

Abstract

We present the first provably convergent off-policy actor-critic algorithm with function approximation (COF-PAC). Key to COF-PAC is the introduction of a new critic, *emphasis critic*, which is trained via Gradient Emphasis Learning (GEM), a novel combination of the key ideas of Gradient Temporal Difference Learning and Emphatic Temporal Difference Learning. With the help of the emphasis critic and the canonical value function critic, we show almost sure convergence for COF-PAC, where the policy parameterization can be nonlinear. This document is heavily outdated and we refer the reader to https://arxiv.org/abs/1911.04384 for the latest version.

1 Introduction

The policy gradient algorithm (Williams, 1992) and its actor-critic extension (Sutton et al., 2000; Konda, 2002) have recently enjoyed great success in various domains, e.g., defeating the top human player in the game Go (Silver et al., 2016), achieving human level control in Atari games (Mnih et al., 2016). The canonical actor-critic algorithm is provably convergent under function approximation (Konda, 2002). However, it is on-policy and suffers from significant data inefficiency (Mnih et al., 2016). While there have been efforts to combine actor-critic algorithms with off-policy learning (Degris et al., 2012; Imani et al., 2018; Zhang et al., 2019), none of the resulting off-policy actor-critic algorithms is provably convergent under function approximation.

In this paper, we present COF-PAC, the first provably convergent off-policy actor-critic algorithm with function approximation. COF-PAC builds on Actor-Critic with Emphatic weightings (ACE, Imani et al. 2018), which reweights policy updates with *emphasis* through the *followon trace* (Sutton et al., 2016). The emphasis corrects the state distribution and the followon trace approximates the emphasis (see Sutton et al. 2016).[1] However, the followon trace can have unbounded variance (Sutton et al., 2016). Hence its approximation to the emphasis can have arbitrarily large error, complicating convergence analysis and hindering finite-sample performance. Instead of using the followon trace, we present a novel learning-based method, Gradient Emphasis Learning (GEM), to approximate the emphasis, inspired by the Gradient TD methods (Sutton et al., 2009b,a; Maei, 2011), Emphatic TD methods (Sutton et al., 2016), and "*reversed TD*" methods (Hallak and Mannor, 2017; Gelada and Bellemare, 2019). We prove the almost sure convergence of GEM under linear function approximation. By contrast, the convergence of the followon trace is only in expectation. In previous

[1] We use emphasis to denote the limit of the expectation of the followon trace, which is slightly different from Sutton et al. (2016) and is clearly defined in the next section.

Optimization Foundations for Reinforcement Learning Workshop at NeurIPS 2019, Vancouver, Canada.

actor-critic algorithms, we have only a value function critic. In COF-PAC, we introduce a new kind of critic, the emphasis critic, which is trained via GEM. With the help of both critics, we prove the almost sure convergence of COF-PAC.

2 Background

We use $||\cdot||$ to denote ℓ_2 norm for vectors and matrices. We use $||x||_d \doteq \sqrt{\sum_i x_i^2 d_i}$ to denote weighted ℓ_2 norm for vectors. All vectors are column. We use "0" to denote an all-zero vector and an all-zero matrix when the dimension can be easily deduced from the context. The use of notation "1" is similar. For a vector x, x_i denotes its i-th component. When not causing confusion, we use vectors and functions interchangeably. Proofs of all lemmas, propositions, and theorems are in the appendix.

We consider a finite Markov Decision Process (MDP) with a finite state space $\mathcal{S}$ with $|\mathcal{S}|$ states, a finite action space $\mathcal{A}$ with $|\mathcal{A}|$ actions, a transition kernel $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0,1]$, a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and a discount factor $\gamma \in [0,1)$. At time step t, an agent at a state S_t takes an action A_t according to $\mu(\cdot|S_t)$, where $\mu : \mathcal{A} \times \mathcal{S} \to [0,1]$ is a *behavior policy*. The agent then gets a reward R_{t+1} satisfying $\mathbb{E}[R_{t+1}] = r(S_t, A_t)$ and proceeds to a new state S_{t+1} according to $p(\cdot|S_t, A_t)$. In the off-policy setting, the agent is interested in a *target policy* π. We use $G_t \doteq \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k}$ to denote the return at time step t when following π. Consequently, we define the state value function v_π and the state action value function q_π as $v_\pi(s) \doteq \mathbb{E}_\pi[G_t|S_t = s]$ and $q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t|S_t = s, A_t = a]$. We use $\rho(s,a) \doteq \frac{\pi(a|s)}{\mu(a|s)}$ to denote the importance sampling ratio and define $\rho_t \doteq \rho(S_t, A_t)$ (Assumption 1 below ensures ρ is well-defined).

2.1 Policy Evaluation

We consider linear function approximation for policy evaluation. Let $x : \mathcal{S} \to \mathbb{R}^{K_1}$ be the state feature function, and $\tilde{x} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{K_2}$ denote the state-action feature function. We use $X \in \mathbb{R}^{|\mathcal{S}| \times K_1}$ and $\tilde{X} \in \mathbb{R}^{N_{sa} \times K_2} (N_{sa} \doteq |\mathcal{S}| \times |\mathcal{A}|)$ to denote feature matrices, where each row of X is $x(s)$ and each row of $\tilde{X}$ is $\tilde{x}(s,a)$. Let $d_\mu \in \mathbb{R}^{|\mathcal{S}|}$ be the stationary distribution of μ, we define $\tilde{d}_\mu \in \mathbb{R}^{N_{sa}}$ where $\tilde{d}_\mu(s,a) \doteq d_\mu(s)\mu(a|s)$. We define $D \doteq diag(d_\mu) \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ and $\tilde{D} \doteq diag(\tilde{d}_\mu) \in \mathbb{R}^{N_{sa} \times N_{sa}}$. Assumption 1 below ensures d_μ exists and D is invertible, as well as $\tilde{D}$. Let $P_\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ be the state transition matrix and $\tilde{P}_\pi \in \mathbb{R}^{N_{sa} \times N_{sa}}$ be the state-action transition matrix, i.e., $P_\pi(s, s') \doteq \sum_a \pi(a|s) p(s'|s,a)$, $\tilde{P}_\pi((s,a),(s',a')) \doteq p(s'|s,a)\pi(a'|s')$.

We first consider Gradient TD methods. For a vector $v \in \mathbb{R}^{|\mathcal{S}|}$, we define a projection $\Pi v \doteq Xy^*, y^* \doteq \arg\min_y ||Xy - v||_{d_\mu}^2$. We have $\Pi = X(X^\top DX)^{-1}X^\top D$ (Assumption 2 below ensures the existence of $(X^\top DX)^{-1}$). Similarly, for a vector $q \in \mathbb{R}^{N_{sa}}$, we define a projection $\tilde{\Pi} \doteq \tilde{X}(\tilde{X}^\top \tilde{D}\tilde{X})^{-1}\tilde{X}^\top \tilde{D}$. The value function v_π is the unique fixed point of the Bellman operator $\mathcal{T} : \mathcal{T}v \doteq r_\pi + \gamma P_\pi v$ where $r_\pi(s) \doteq \sum_a r(s,a)\pi(a|s)$. Similarly, q_π is the unique fixed point for the operator $\tilde{\mathcal{T}} : (\tilde{\mathcal{T}}q)(s,a) \doteq r + \gamma \tilde{P}_\pi q$. GTD2 (Sutton et al., 2009a) learns an estimate v for v_π, minimizing $||\Pi\mathcal{T}v - v||_{d_\mu}^2$. GQ(0) (Maei, 2011) learns an estimate q for q_π, minimizing $||\tilde{\Pi}\hat{\mathcal{T}}q - q||_{d_\mu}^2$. The original convergence analysis of Gradient TD methods assumes i.i.d. data (Sutton et al., 2009b,a; Maei, 2011), where the transitions $\{(s_t, a_t, r_t, s_t')\}_{t=0\dots}$ are i.i.d. with $s_t \sim d_\mu(\cdot), a_t \sim \mu(\cdot|s_t), s_t' \sim p(\cdot|s_t, a_t), \mathbb{E}[r_t] = r(s_t, a_t)$ and have bounded second moments. Wang et al. (2017) showed that Gradient TD methods remain convergent when this i.i.d. assumption is relaxed to Markov data (e.g., $\{S_0, A_0, R_1, S_1, \dots\}$).

Besides Gradient TD methods, Emphatic TD (ETD, Sutton et al. 2016) is also used for off-policy policy evaluation. We use $v \doteq X\nu$ to denote an estimate for v_π, where ν is the learnable parameters. ETD(0) updates ν as

$$M_t \doteq i(S_t) + \gamma \rho_{t-1} M_{t-1}, \tag{1}$$

$$\nu_{t+1} \doteq \nu_t + \alpha M_t \rho_t (R_{t+1} + \gamma x(S_{t+1})^\top \nu_t - x(S_t)^\top \nu_t) x(S_t)^\top, \tag{2}$$

where α is a step size, M_t is the followon trace, and $i : \mathcal{S} \to \mathbb{R}_+$ is the interest function reflecting the user's preference for different states (Sutton et al., 2016).

2.2 Control

Off-policy actor-critic methods (Degris et al., 2012; Imani et al., 2018) typically aim to maximize the excursion objective,

$$J(\pi) \doteq \sum_s d_\mu(s) i(s) v_\pi(s),$$

by adapting the target policy π. We assume π is parameterized by $\theta \in \mathbb{R}^K$, and use θ and π interchangeably in the rest of this paper when this does not cause confusion.

According to the off-policy policy gradient theorem (Imani et al., 2018), the policy gradient is $\nabla_\theta J(\theta) = \sum_s \bar{m}(s) \sum_a q_\pi(s,a) \nabla_\theta \pi(a|s)$, where $\bar{m} \doteq (I - \gamma P_\pi^\top)^{-1} D i \in \mathbb{R}^{|\mathcal{S}|}$. We rewrite $\bar{m}$ as $DD^{-1}(I - \gamma P_\pi^\top)^{-1} D i$ and define

$$m_\pi \doteq D^{-1}(I - \gamma P_\pi^\top)^{-1} D i.$$

We therefore have $\bar{m} = D m_\pi$, i.e., $\bar{m}(s) = d_\mu(s) m_\pi(s)$. Alternatively, we can rewrite $\nabla_\theta J(\theta)$ as

$$\nabla_\theta J(\theta) = \sum_s d_\mu(s) m_\pi(s) \sum_a \mu(a|s) \psi_\theta(s,a) q_\pi(s,a), \tag{3}$$

where $\psi_\theta(s,a) \doteq \rho(s,a) \nabla_\theta \log \pi(a|s) \in \mathbb{R}^K$. We refer to m_π as the *emphasis* in the rest of this paper. For computing $\nabla_\theta J(\theta)$, we need m_π and q_π, to which we typically do not have access. Imani et al. (2018) approximate $m_\pi(S_t)$ with the followon trace M_t, yielding the ACE update $\theta_{t+1} \doteq \theta_t + \alpha M_t \rho_t q_\pi(S_t, A_t) \nabla \log \pi(A_t|S_t)$. Assuming $\lim_{t\to\infty} \mathbb{E}_\mu[M_t|S_t = s]$ exists, Sutton et al. (2016) show that $\lim_{t\to\infty} \mathbb{E}_\mu[M_t|S_t = s] = m_\pi(s)$. The existence of this limit is later established in Lemma 1 in Zhang et al. (2019).

2.3 Assumptions and Lemmas

Assumption 1. *The expected reward is bounded by $R_{\max}$, i.e., $\forall (s,a)$, $|r(s,a)| < R_{\max}$. The Markov Reward Process (MRP) induced by the behavior policy μ is ergodic, and $\forall (s,a)$, $\mu(a|s) > 0$.*

Let $A_\theta \doteq X^\top (I - \gamma P_\pi^\top) D X, C \doteq X^\top D X, \tilde{A}_\theta \doteq \tilde{X}^\top \tilde{D}(I - \gamma \tilde{P}_\pi)\tilde{X}, \tilde{C} \doteq \tilde{X}^\top \tilde{D} \tilde{X}$ and $\xi(\cdot)$ be the minimum singular value of a matrix, we assume

Assumption 2. $\xi(C) > 0, \xi(\tilde{C}) > 0, \inf_\theta \xi(A_\theta) > 0, \inf_\theta \xi(\tilde{A}_\theta) > 0$*, and C is positive definite.*

Remark 1. *The non-singularity for a fixed θ is essential for Gradient TD methods to ensure the problem of policy evaluation is solvable (see Sutton et al. (2009b,a); Maei (2011)). We make a slightly stronger assumption that the minimum singular value does not approach 0 during the optimization of θ. As the ℓ_2 norm of a matrix is the minimum singular value of its inverse, this assumption helps establish the boundedness of A_θ^{-1} and $\tilde{A}_\theta^{-1}$.*

Assumption 3. *(Policy Parameterization) (a) There exists a constant $C_0 < \infty$ such that $\forall (s,a)$,*

$$||\psi_\theta(s,a)|| < C_0, ||\nabla_\theta \psi_\theta(s,a)|| < C_0$$
$$|\pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s)| < C_0 ||\theta_1 - \theta_2||,\ ||\psi_{\theta_1}(s,a) - \psi_{\theta_2}(s,a)|| < C_0 ||\theta_1 - \theta_2||$$

(b) $\inf_\theta \xi(I - \gamma P_\pi) > 0$.

Remark 2. *The bounded $\nabla_\theta \psi_\theta(s,a)$ is also assumed in on-policy actor-critic algorithms (Assumption 5.4 in Konda (2002)). As $\gamma < 1$, $\xi(I - \gamma P_\pi) > 0$ holds for any fixed θ. Our assumption states the minimum singular value of $I - \gamma P_\pi$ does not approach 0 during the optimization of θ.*

Lemma 1. *Under Assumptions 1 and 3, there exists a constant $C_1 < \infty$ such that $\forall (s, a, \theta_1, \theta_2)$*

$$||\nabla_\theta J(\theta_1)|| < C_1, ||\nabla_\theta J(\theta_1) - \nabla_\theta J(\theta_2)|| < C_1 ||\theta_1 - \theta_2||, ||H(J(\theta_1))|| < C_1,$$

where $H(J(\theta))$ is the Hessian of $J(\theta)$.

3 Gradient Emphasis Learning

To motivate, we first discuss the disadvantages of the followon trace M_t. The first problem lies in the large variance. Empirically, it has been observed that the variance of M_t can be unbounded (Sutton et al., 2016), which leads to problems in real applications of ETD. For example, as pointed out in Sutton and Barto (2018), "it is nigh impossible to get consistent results in computational experiments" in Baird's counterexample, a benchmark domain in measuring RL algorithms' off-policy performance. Theoretically, this unbounded variance may preclude a convergent analysis for ACE. Under mild conditions, the on-policy actor-critic (Konda, 2002) visits regions near the local maxima (and saddle points) infinitely often, where the radius of those regions are determined by the approximation error of the critic (i.e., the distance between q and q_π) (Konda, 2002). Similarly, we would expect an off-policy actor-critic algorithm to visit the regions whose radius are determined by the approximation error of both the m_π estimate and the q_π estimate (We formalize this in Theorem 1). As M_t has unbounded variance, its approximation error for $m_\pi(S_t)$ can be arbitrarily large. The negative influence of the approximation error of the canonical critic can be eliminated if compatible features are considered (Sutton et al., 2000; Konda, 2002). However, this technique is not compatible with the followon trace M_t as it does not have any features. Consequently, those regions become arbitrarily large.

The second problem is that M_t is almost memoryless. M_t is only a scalar random variable but we expect it to track m_π, a vector in $\mathbb{R}^{|\mathcal{S}|}$. It is the expectation of M_t, not M_t itself, that converges. However, in Eq (1), M_{t+1} is bootstrapped by M_t, not its expectation, indicating this bootstrap for M_{t+1} can be poor. By contrast, in canonical learning-based methods, e.g., the ETD value update Eq (2), the approximation itself, e.g., ν_t, converges and we bootstrap via this approximation. The quality of this bootstrap is therefore likely to be high, which is particularly important when π is changing, so that the critic can adapt to the new policy quickly. The followon trace, however, can hardly provide a good bootstrap due to its lack of memory, yielding an obstacle in the convergence analysis for ACE. Moreover, the expectation of M_t tracks $m_\pi(S_t)$ only in a limiting sense for a fixed π. If π is changing, it is questionable whether the expectation of M_t can track the changing m_π, not to say M_t itself given the possibly unbounded variance. By contrast, in the canonical on-policy actor-critic, the critic's ability to track a changing actor is clearly proven (Konda, 2002). In this paper, we propose to use stochastic approximation to approximate m_π.

We now derive the Gradient Emphasis Learning (GEM) algorithm. Throughout this section, we assume π is fixed. We consider linear function approximation, and our estimate for m_π is $m \doteq Xw$, where $w \in \mathbb{R}^{K_1}$ is the learnable parameters. For a vector $y \in \mathbb{R}^{|\mathcal{S}|}$, we define an operator $\hat{\mathcal{T}}$ as $\hat{\mathcal{T}}y \doteq i + \gamma D^{-1} P_\pi^\top D y$. We have

Proposition 1. *$\hat{\mathcal{T}} m_\pi = m_\pi$ and $\forall y, \lim_{k\to\infty} \hat{\mathcal{T}}^{(k)} y = m_\pi$, where $\hat{\mathcal{T}}^{(1)} \doteq \hat{\mathcal{T}}, \hat{\mathcal{T}}^{(k+1)} \doteq \hat{\mathcal{T}}(\hat{\mathcal{T}}^{(k)})$.*

Given Proposition 1, it is tempting to compose a semi-gradient update rule for updating w:

$$w_{t+1} \leftarrow w_t + \alpha[i(S_{t+1}) + \gamma\rho(S_t, A_t)x(S_t)^\top w_t - x(S_{t+1})^\top w_t]x(S_{t+1}),$$

analogously to the semi-gradient reversed TD algorithm (discounted) COP-TD (Hallak and Mannor, 2017; Gelada and Bellemare, 2019). All semi-gradient reversed TD methods, however, can diverge under linear function approximation for the same reason as the divergence of off-policy linear TD: the A matrix (defined in Assumption 2) is not guaranteed to be negative semi-definite (see Sutton et al. (2016)). Motivated by the long-standing convergent Gradient TD methods, we seek an approximate solution m that satisfies $m = \Pi\hat{\mathcal{T}}m$ via minimizing a projected objective $J_\pi(w) \doteq \frac{1}{2}||\Pi\bar{\delta}_w||^2_{d_\mu}$, where $\bar{\delta}_w \doteq \hat{\mathcal{T}}(Xw) - Xw$. With $\bar{p}(\bar{s}, \bar{a}|s) \doteq d_\mu(s)^{-1} d_\mu(\bar{s})\mu(\bar{a}|\bar{s})p(s|\bar{s}, \bar{a})$, we have

Lemma 2. $\sum_{\bar{s},\bar{a}} \bar{p}(\bar{s}, \bar{a}|s) = 1$, $\bar{\delta}_w(s) = i(s) + \gamma \sum_{\bar{s},\bar{a}} \bar{p}(\bar{s}, \bar{a}|s)\rho(\bar{s}, \bar{a})(Xw)(\bar{s}) - (Xw)(s)$.

Intuitively, $(\bar{s}, \bar{a})$ stands for a previous state-action pair. We now compute $\nabla_w J_\pi(w)$. Similar to Gradient TD methods, we have

$$\begin{aligned}
J_\pi(w) &= \frac{1}{2}\bar{\delta}_w^\top \Pi^\top D \Pi \bar{\delta}_w = (\bar{\delta}_w^\top DX)(X^\top DX)^{-1}(X^\top D\bar{\delta}_w), \\
\nabla_w J_\pi(w) &= \nabla_w (X^\top D\bar{\delta}_w)^\top (X^\top DX)^{-1}(X^\top D\bar{\delta}_w), \\
X^\top D\bar{\delta}_w &= \mathbb{E}_{s\sim d(s),(\bar{s},\bar{a})\sim\bar{p}(\bar{s},\bar{a}|s)}[i(s) + \gamma\rho(\bar{s}, \bar{a})x(\bar{s})^\top w - x(s)^\top w]x(s), \\
\nabla_w X^\top D\bar{\delta}_w &= \mathbb{E}_{s\sim d(s),(\bar{s},\bar{a})\sim\bar{p}(\bar{s},\bar{a}|s)}[\gamma\rho(\bar{s}, \bar{a})x(s)x(\bar{s})^\top - x(s)x(s)^\top].
\end{aligned}$$

Here $(X^\top DX)^{-1}(X^\top D\bar{\delta}_w)$ is the solution to the supervised learning problem of predicting $\bar{\delta}_w$ with features X. Similar to GTD2 (Sutton et al., 2009a), we use another set of parameters $\kappa \in \mathbb{R}^{K_1}$ to approximate $(X^\top DX)^{-1}(X^\top D\bar{\delta}_w)$.

In this section we consider the same i.i.d. transitions $\{(\bar{s}_t, \bar{a}_t, s_t)\}_{t=0,...}$ for analysis as Sutton et al. (2009b,a), where $\bar{s}_t \sim d_\mu(\cdot), \bar{a}_t \sim \mu(\cdot|\bar{s}_t), s_t \sim p(\cdot|\bar{s}_t, \bar{a}_t)$ and the transitions have bounded second moments. The joint distribution is therefore $p(\bar{s}_t, \bar{a}_t, s_t) = d_\mu(\bar{s}_t)\mu(\bar{a}_t|\bar{s}_t)p(s|\bar{a}_t, \bar{s}_t)$. A convergent analysis for sequential Markov data can be done with similar techniques as Wang et al. (2017) or Yu (2017), which we leave for future work. As d_μ is the stationary distribution, the marginalized distribution of s_t is therefore also d_μ. Consequently, we have $p(\bar{s}_t, \bar{a}_t|s_t) = p(\bar{s}_t, \bar{a}_t, s_t)d_\mu(s_t)^{-1} = \bar{p}(\bar{s}_t, \bar{a}_t|s_t)$, indicating we can now use samples $(\bar{s}_t, \bar{a}_t, s_t)$ to estimate $\bar{\delta}_w$ and $\nabla_w X^\top D\bar{\delta}_w$, as well as $\nabla_w J_\pi(w)$. Now we are ready to present GEM, which updates κ and w recursively as

$$\begin{aligned} \kappa_{t+1} &\doteq \kappa_t + \alpha_t(i_t + \gamma\bar{\rho}_t\bar{x}_t^\top w_t - x_t^\top w_t - x_t^\top \kappa_t)x_t, \\ w_{t+1} &\doteq w_t + \alpha_t(x_t - \gamma\bar{\rho}_t\bar{x}_t)x_t^\top \kappa_t, \end{aligned} \tag{4}$$

where $\bar{x}_t \doteq x(\bar{s}_t), \bar{\rho}_t \doteq \rho(\bar{s}_t, \bar{a}_t), i_t \doteq i(s_t), x_t \doteq x(s_t)$, and α_t is a deterministic sequence satisfying the Robbins-Monro condition (Robbins and Monro, 1951), i.e., $\{\alpha_t\}$ is non-increasing positive and $\sum_t \alpha_t = \infty, \sum_t \alpha_t^2 < \infty$. We now characterize the asymptotic behavior of GEM. With $b \doteq X^\top Di, w^* \doteq A_\theta^{-1}b$, we have

Proposition 2. $\Pi\hat{\mathcal{T}}(Xw^*) = Xw^*$.

Proposition 3. *(Convergence of GEM) Under Assumptions (1, 2), the iterates $\{w_t\}$ generated by* (4) *converges to w^* almost surely.*

We use similar techniques as Hallak and Mannor (2017) in proving Proposition 2 and the proof of Proposition 3 is similar to Sutton et al. (2009a) up to a change of notations. Although reversed TD has been explored by Hallak and Mannor (2017); Gelada and Bellemare (2019), GEM is the first provably convergent reversed TD method under linear function approximation.

4 Convergent Off-Policy Actor Critic

We drop the subscript θ in ∇_θ for simplicity in this section. To estimate $\nabla J(\theta)$, we need both m_π and q_π. The former can be learned via GEM. For the latter, we consider GQ(0) with linear function approximation, Our estimate for q_π is $q \doteq \tilde{X}u$, where $u \in \mathbb{R}^{K_2}$ is the learnable parameters. GQ(0) minimizes the objective $J_\pi(u) \doteq ||\tilde{\Pi}\hat{\mathcal{T}}q - q||_{d_\mu}^2$. Under Assumptions 1 and 2, GQ(0) converges to $u^* \doteq \tilde{A}_\theta^{-1}\tilde{b}$ almost surely, where $\tilde{b} \doteq \tilde{X}^\top\tilde{D}r$ (Maei, 2011).

Algorithm 1: Convergent Off-Policy Actor-Critic (COF-PAC)

Input:
θ: parameters of π
β_t: a sequence of deterministic step sizes
$x, \tilde{x}$: feature functions

Get S_0 and set $t \leftarrow 0$
while *True* **do**
    Get $A_t \sim \mu(\cdot|S_t)$
    Execute A_t and get R_{t+1}, S_{t+1}
    $w_t \leftarrow \arg\min_w J_{\theta_t}(w)$ // e.g., using GEM with historical transitions
    $u_t \leftarrow \arg\min_u J_{\theta_t}(u)$ // e.g., using GQ(0) with historical transitions
    $\theta_{t+1} \leftarrow \theta_t + \beta_t(w_t^\top x_t)(u_t^\top \tilde{x}_t)\rho_t \nabla \log \pi(A_t|S_t)$
    $t \leftarrow t + 1$
end

With a slight abuse of notation, we now present the Convergent Off-Policy Actor-Critic (COF-PAC) algorithm in Algorithm 1, where $x_t \doteq x(S_t), \tilde{x}_t \doteq \tilde{x}(S_t, A_t)$, and β_t is a deterministic sequence satisfying the Robbins-Monro condition (Robbins and Monro, 1951), i.e., $\{\beta_t\}$ is non-increasing positive and $\sum_t \beta_t = \infty, \sum_t \beta_t^2 < \infty$. In Algorithm 1, w_t and u_t are uniquely determined by θ_t due to Assumption 2, so we use w_t, u_t and $w_{\theta_t}, u_{\theta_t}$ interchangeably. Particularly, we define $w_\theta \doteq \arg\min_w J_\theta(w), u_\theta \doteq \arg\min_u J_\theta(u), m(s;\theta) \doteq w_\theta^\top x(s), q(s, a;\theta) \doteq u_\theta^\top \tilde{x}(s, a)$. We first show the Lipschitz continuity of GEM and GQ(0) solutions.

Proposition 4. *(Lipschitz continuity of GEM and GQ(0) solutions) Under Assumptions (1,2, 3), there exists a constant $C_1 < \infty$ such that $\forall \theta_1, \theta_2$*

$$\max(||w_{\theta_1}||, ||u_{\theta_1}||) < C_1, \, \max(||w_{\theta_1} - w_{\theta_2}||, ||u_{\theta_1} - u_{\theta_2}||) \leq C_1 ||\theta_1 - \theta_2||.$$

Before proceeding to a convergence analysis of COF-PAC, we first analyze a noise term introduced by the estimates m and q. With $\hat{g}(\theta) \doteq \sum_s d_\mu(s) m(s;\theta) \sum_a \mu(a|s) \psi_\theta(s,a) q(s,a;\theta)$, $\psi_t \doteq \rho(S_t, A_t) \nabla \log \pi(A_t|S_t)$, $m_t \doteq w_t^\top x_t$, $q_t \doteq u_t^\top \tilde{x}_t$, we have

Lemma 3. *Under Assumptions (1,2, 3), $|\sum_t \beta_t \nabla J(\theta_t)^\top (m_t q_t \psi_t - \hat{g}(\theta_t))| < \infty \quad a.s.$* [2]

Proof. (Sketch) This lemma plays a central role in the following Theorem 1. The proof is inspired by Konda (2002). We first construct several auxiliary MDPs. The differential state-action value function of those MDPs help make a transformation of the original noise. We then decompose the transformed noise into four components. The first component is a Martingale with bounded second moments thus converges. We then verify the boundedness of the remaining three components, which involve Proposition 4. Details are in the appendix. □

The bias introduced by the estimates m and q is $\nabla_\theta J(\theta) - \hat{g}(\theta) = b^{(1)}(\theta) + b^{(2)}(\theta)$ where

$$b^{(1)}(\theta) \doteq \sum_s d_\mu(s)(m_\pi(s) - m(s;\theta)) \sum_a \mu(a|s) \psi_\theta(s,a) q(s,a;\theta),$$
$$b^{(2)}(\theta) \doteq \sum_s d_\mu(s) m_\pi(s) \sum_a \mu(a|s) \psi_\theta(s,a)(q_\pi(s,a) - q(s,a;\theta)).$$

If the estimate m is accurate, $b^{(1)}(\theta)$ will be 0. If the estimate q is accurate, $b^{(2)}(\theta)$ will be 0. The accuracy of m and q determines where COF-PAC converges to.

Theorem 1. *(Convergence of COF-PAC) Under Assumptions 1-3, the iterates $\{\theta_t\}$ generated by COF-PAC (Algorithm 1) satisfy*

$$\liminf_t \left[||\nabla J(\theta_t)|| - \left(||b^{(1)}(\theta_t)|| + ||b^{(2)}(\theta_t)|| \right) \right] \leq 0,$$

almost surely, i.e., $\{\theta_t\}$ visits any neighborhood of the set $\{\theta : ||\nabla J(\theta)|| \leq ||b^{(1)}(\theta)|| + ||b^{(2)}(\theta)||\}$ infinitely often almost surely.

The proof of Theorem 1 is standard and follows the same routine as Konda (2002). Noise is dealt with Lemma 3. According to Theorem 1, COF-PAC has reached the same convergence level as the canonical on-policy actor-critic (Konda, 2002). We present experimental results and related work in the appendix.

5 Conclusion

In this paper, we present COF-PAC, the first provably convergent off-policy actor-critic algorithm under function approximation. Key to COF-PAC is GEM, which can be combined with any emphatic algorithm (e.g., ETD). The GEM algorithm presented in this paper is of its simplest form. A possibility for future work is to extend GEM to GEM(λ, β) in analogue to ETD(λ, β) (Hallak et al., 2016) and GTD(λ) (Yu, 2015). Our COF-PAC is presented in a bi-level optimization form, similar to Sutton et al. (2000). A possibility for future work is to adapt it into a two-timescale form as Konda (2002). Furthermore, we consider only general features for the emphasis critic and the value function critic. A natural extension is to specify compatible features as used by Sutton et al. (2000); Konda (2002), so that the bias $b^{(1)}(\theta)$ and $b^{(2)}(\theta)$ can be reduced to 0 even if there is still approximation error in m and q. COF-PAC optimizes the excursion objective. Developing convergent off-policy actor-critic algorithms under some potentially better objectives, e.g., the counterfactual objective (Zhang et al., 2019), is also worth further investigation.

[2] By $|\sum_t x_t| < \infty$ we mean there exists a constant $C < \infty$ such that $\forall T, |\sum_{t=0}^T x_t| < C$.

6 Acknowledgments

SZ is generously funded by the Engineering and Physical Sciences Research Council (EPSRC). This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement number 637713). The experiments were made possible by a generous equipment grant from NVIDIA. BL's research is funded by the National Science Foundation (NSF) under grant NSF IIS1910794 and Amazon Research Award.

References

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. *Machine Learning*.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific Belmont, MA.

Ciosek, K. and Whiteson, S. (2017). Expected policy gradients. *arXiv preprint arXiv:1706.05374*.

Degris, T., White, M., and Sutton, R. S. (2012). Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. (2018). Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*.

Gelada, C. and Bellemare, M. G. (2019). Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*.

Gu, S. S., Lillicrap, T., Turner, R. E., Ghahramani, Z., Schölkopf, B., and Levine, S. (2017). Interpolated policy gradient: Merging on-policy and off-policy gradient estimation for deep reinforcement learning. In *Advances in Neural Information Processing Systems*.

Hallak, A. and Mannor, S. (2017). Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning*.

Hallak, A., Tamar, A., Munos, R., and Mannor, S. (2016). Generalized emphatic temporal difference learning: Bias-variance analysis. In *Proceedins of 30th AAAI Conference on Artificial Intelligence*.

Imani, E., Graves, E., and White, M. (2018). An off-policy policy gradient theorem using emphatic weightings. In *Advances in Neural Information Processing Systems*.

Konda, V. R. (2002). *Actor-critic algorithms*. PhD thesis, Massachusetts Institute of Technology.

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

Maei, H. R. (2011). *Gradient temporal-difference learning algorithms*. PhD thesis, University of Alberta.

Maei, H. R. (2018). Convergent actor-critic algorithms under off-policy training and function approximation. *arXiv preprint arXiv:1802.07842*.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction (2nd Edition)*. MIT press.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. (2009a). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning*.

Sutton, R. S., Maei, H. R., and Szepesvári, C. (2009b). A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*.

Sutton, R. S., Mahmood, A. R., and White, M. (2016). An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*.

Wang, Y., Chen, W., Liu, Y., Ma, Z.-M., and Liu, T.-Y. (2017). Finite sample analysis of the gtd policy evaluation algorithms in markov setting. In *Advances in Neural Information Processing Systems*.

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., and de Freitas, N. (2016). Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*.

Yu, H. (2015). On convergence of emphatic temporal-difference learning. In *Conference on Learning Theory*.

Yu, H. (2017). On convergence of some gradient-based temporal-differences algorithms for off-policy learning. *arXiv preprint arXiv:1712.09652*.

Zhang, S., Boehmer, W., and Whiteson, S. (2019). Generalized off-policy actor-critic. In *Advances in Neural Information Processing Systems*.

A Proofs

A.1 Proof of Lemma 1

Proof. (1) According to Assumption 3, there exists a constant $\epsilon > 0$ such that $\forall \theta, \xi(I - P_\pi^\top) > \epsilon$. We have

$$||m_\pi|| \leq ||D^{-1}|| \, ||(I - \gamma P_\pi^\top)^{-1}|| \, ||D|| \, ||i|| < \frac{1}{\epsilon} ||D^{-1}|| \, ||D|| \, ||i||, \tag{5}$$

$$||v_\pi|| = ||(I - \gamma P_\pi)^{-1} r_\pi|| < \frac{\sqrt{|\mathcal{S}|} R_{\max}}{\epsilon}. \tag{6}$$

According to the analytical expression of $\nabla_\theta J(\theta)$ in Eq (3), it follows easily that there exists a constant C_1 such that $||\nabla_\theta J(\theta)|| < C_1$ holds for all θ.

(ii) As π is Lipschitz continuous in θ, it follows easily that P_π is also Lipschitz continuous in θ. To show the Lipschitz continuity of $\nabla_\theta J(\theta)$ in θ, it suffices to show m_π and v_π are Lipschitz continuous. Using the fact $||Y_1^{-1} - Y_2^{-1}|| = ||Y_1^{-1}(Y_1 - Y_2)Y_2^{-1}|| \leq ||Y_1^{-1}|| \, ||Y_1 - Y_2|| \, ||Y_2^{-1}||$, we have

$$\begin{aligned} ||(I - \gamma P_{\theta_1}^\top)^{-1} - (I - \gamma P_{\theta_2}^\top)^{-1}|| &\leq ||(I - \gamma P_{\theta_1}^\top)^{-1}|| \gamma ||P_{\theta_1}^\top - P_{\theta_2}^\top|| \, ||(I - \gamma P_{\theta_2}^\top)^{-1}|| \\ &\leq \frac{\gamma}{\epsilon^2} ||P_{\theta_1}^\top - P_{\theta_2}^\top||. \end{aligned}$$

$(I - \gamma P_\theta^\top)^{-1}$ is therefore Lipschitz continuous in θ. It follows easily that m_π is Lipschitz continuous in θ. As r_π is bounded and Lipschitz continuous in θ, v_π is therefore also Lipschitz continuous in θ as it is a product of two bounded Lipschitz function.

(iii) For the sake of clarity, in this part use ∇_θ to denote the gradient w.r.t. one dimension of θ. We first show $\nabla_\theta v_\pi(s)$ is bounded. As $v_\pi = r_\pi + \gamma P_\pi v_\pi$, we have

$$\begin{aligned} \nabla_\theta v_\pi &= \nabla_\theta r_\pi + \gamma P_\pi \nabla_\theta v_\pi + \gamma \nabla_\theta P_\pi v_\pi, \\ \nabla_\theta v_\pi &= (I - \gamma P_\pi)^{-1}(\nabla_\theta r_\pi + \gamma \nabla_\theta P_\pi v_\pi). \end{aligned}$$

According to Assumptions (1, 3) and Eq (6), there exists a constant $C_1 < \infty$ such that

$$||\nabla_\theta r_\pi + \gamma \nabla_\theta P_\pi v_\pi|| < C_1.$$

Consequently, $||\nabla_\theta v_\pi|| < \frac{C_1}{\epsilon}$. It follows easily that $||\nabla_\theta q_\pi||$ is bounded.

We then show $\nabla_\theta m_\pi(s)$ is bounded. We have

$$\begin{aligned} i + \gamma D^{-1} P_\pi^\top D m_\pi &= i + \gamma D^{-1} P_\pi^\top (I - P_\pi^\top)^{-1} D i \\ &= \Big(D^{-1}(I - \gamma P_\pi^\top) + \gamma D^{-1} P_\pi^\top\Big)(I - P_\pi^\top)^{-1} D i \\ &= D^{-1}(I - P_\pi^\top)^\top D i = m_\pi. \end{aligned} \tag{7}$$

Consequently,

$$\begin{aligned} \nabla_\theta m_\pi &= \gamma D^{-1} \nabla_\theta P_\pi^\top D m_\pi + \gamma D^{-1} P_\pi^\top D \nabla_\theta m_\pi \\ \nabla_\theta m_\pi &= (I - \gamma D^{-1} P_\pi^\top D)^{-1} \gamma D^{-1} \nabla_\theta P_\pi^\top D m_\pi \\ &= \Big(D^{-1}(I - \gamma P_\pi^\top) D\Big)^{-1} \gamma D^{-1} \nabla_\theta P_\pi^\top D m_\pi \\ &= D^{-1}(I - \gamma P_\pi^\top)^{-1} D \gamma D^{-1} \nabla_\theta P_\pi^\top D m_\pi \\ &= \gamma D^{-1}(I - \gamma P_\pi^\top)^{-1} \nabla_\theta P_\pi^\top D m_\pi. \end{aligned}$$

According to Assumption 3 and Eq (5), it follows easily that there exists a constant $C_1 < \infty$ such that $||\nabla_\theta m_\pi|| < C_1$.

We now take gradients w.r.t. θ in both sides of Eq (3) and use the product rule of calculus, it follows easily that there exists a constant $C_1 < \infty$ such that $||H(J(\theta))|| < C_1$.

Note we have considered three different constants to establish the boundedness in (i)(iii) and the Lipschitz continuity in (ii). The C_1 in the statement of this Lemma can be simply set to the largest one. $\square$

A.2 Proof of Proposition 1

Proof. $\hat{\mathcal{T}} m_\pi = m_\pi$ follows directly from Eq (7). As P_π is a stochastic matrix, we have

$$I + \gamma P_\pi^\top + (\gamma P_\pi^\top)^2 + \cdots = (I - \gamma P_\pi^\top)^{-1}.$$

For any y,

$$\begin{aligned}
\hat{\mathcal{T}}^{(2)} y &= i + \gamma D^{-1} P_\pi^\top D(i + \gamma D^{-1} P_\pi^\top D y) \\
&= i + D^{-1} \gamma P_\pi^\top D i + \gamma D^{-1} (\gamma P_\pi^\top)^2 D y \\
\hat{\mathcal{T}}^{(\infty)} y &= D^{-1} I D i + D^{-1} \gamma P_\pi^\top D i + D^{-1} (\gamma P_\pi^\top)^2 D i + \dots \\
&= D \Big(\sum_{k=0}^{\infty} (\gamma P_\pi^\top)^k \Big) D^{-1} \\
&= D^{-1} (I - \gamma P_\pi^\top)^{-1} D i \\
&= m_\pi
\end{aligned}$$

□

A.3 Proof of Lemma 2

Proof. As d_μ is the stationary distribution under μ and p, we have

$$\sum_{\bar{s}, \bar{a}} \bar{p}(\bar{s}, \bar{a} | s) = d_\mu(s)^{-1} \sum_{\bar{s}, \bar{a}} d_\mu(\bar{s}) \mu(\bar{a} | \bar{s}) p(s | \bar{s}, \bar{a}) = d_\mu(s)^{-1} d_\mu(s) = 1.$$

According to the definition of $\bar{\delta}_w$ and $\hat{\mathcal{T}}$, we have

$$\bar{\delta}_w(s) = i(s) + \gamma d_\mu(s)^{-1} \sum_{\bar{s}, \bar{a}} d_\mu(\bar{s}) \pi(\bar{a} | \bar{s}) p(s | \bar{s}, \bar{a}) (Xw)(\bar{s}) - (Xw)(s)$$

The rest follows immediately from the definition of $\bar{p}(\bar{s}, \bar{a} | s)$. □

A.4 Proof of Proposition 2

Proof. Using similar techniques as Hallak and Mannor (2017), we have

$$\begin{aligned}
\Pi \hat{\mathcal{T}}(X w^*) &= X (X^\top D X)^{-1} X^\top D \Big(i + \gamma D^{-1} P_\pi^\top D X w^* \Big) \\
&= X (X^\top D X)^{-1} b + \gamma X (X^\top D X)^{-1} X^\top P_\pi^\top D X A_\theta^{-1} b \\
&= X (X^\top D X)^{-1} \Big(A_\theta + \gamma X^\top P_\pi^\top D X \Big) A_\theta^{-1} b \\
&= X (X^\top D X)^{-1} X^\top D X A_\theta^{-1} b \quad \text{(Definition of } A_\theta) \\
&= X w^*
\end{aligned}$$

□

A.5 Proof of Proposition 3

Proof. This proof is very similar to Sutton et al. (2009a). We first define $d_t^\top \doteq [\kappa_t^\top, w_t^\top]$, which can be expressed in a recursive form as

$$d_{t+1} = d_t + \alpha_t (G_{t+1} d_t + g_{t+1}),$$

where

$$G_{t+1} = \begin{bmatrix} -x_t x_t^\top & x_t (\gamma \bar{\rho}_t \bar{x}_t - x_t)^\top \\ (x_t - \gamma \bar{\rho}_t \bar{x}_t) x_t^\top & 0 \end{bmatrix}, \quad g_{t+1} = \begin{bmatrix} i_t x_t \\ 0 \end{bmatrix}.$$

Note

$$
\begin{aligned}
\mathbb{E}[x_t(x_t - \gamma \bar{\rho}_t \bar{x}_t)^\top] &= X^\top D(X - \gamma D^{-1} P_\pi^\top DX) = X^\top (I - \gamma P_\pi^\top) DX = A_\theta, \\
\mathbb{E}[i_t x_t] &= X^\top D i = b, \\
\mathbb{E}[x_t x_t^\top] &= C.
\end{aligned}
$$

We therefore have

$$
G \doteq \mathbb{E}[G_{t+1}] = \begin{bmatrix} -C & -A_\theta \\ A_\theta^\top & 0 \end{bmatrix}, \quad g \doteq \mathbb{E}[g_{t+1}] = \begin{bmatrix} b \\ 0 \end{bmatrix}
$$

Now we rewrite the update for d_t as

$$
\begin{aligned}
d_{t+1} &= d_t + \alpha_t [G d_t + g + (G_{t+1} - G) d_t + (g_{t+1} - g)] \\
&= d_t + \alpha_t [h(d_t) + M_{t+1}],
\end{aligned}
$$

where

$$
\begin{aligned}
h(d) &\doteq Gd + g \\
M_{t+1} &\doteq (G_{t+1} - G) d_t + (g_{t+1} - g)
\end{aligned}
$$

We can now imitate the proof of the convergence of GTD2 in Section 5 of Sutton et al. (2009a) directly, up to a change of notations. Particularly, our A is the transpose of their A, our b is defined with i while their b is defined with r_π, and we set their η to 1. It is straightforward to verify that these differences do not influence their convergent arguments and the iterates $\{d_t\}$ converges to $-G^{-1}b$ almost surely. It can be easily verified by block matrix inversion that the second half of $-G^{-1}b$ is indeed $A_\theta^{-1} b$. □

A.6 Proof of Proposition 4

Proof. From Assumption 2, there exists a constant $C_1 < \infty$ such that $\forall \theta, ||A_\theta^{-1}|| < C_1, ||\tilde{A}_\theta^{-1}|| < C_1$. As both b and $\tilde{b}$ are bounded and independent of θ, $w_\theta = A_\theta^{-1} b, u_\theta = \tilde{A}_\theta^{-1} \tilde{b}$, it follows easily that $||w_\theta||$ and $||u_\theta||$ are bounded. We show only the Lipschitz continuity of w_θ here. The Lipschitz continuity of u_θ can be established with the same routine. We have

$$
\begin{aligned}
||A_{\theta_1}^{-1} - A_{\theta_2}^{-1}|| &= ||A_{\theta_1}^{-1}(A_{\theta_1} - A_{\theta_2}) A_{\theta_2}^{-1}|| \\
&\leq ||A_{\theta_1}^{-1}|| \, ||A_{\theta_2}^{-1}|| \, ||A_{\theta_1} - A_{\theta_2}|| \\
&\leq \gamma ||A_{\theta_1}^{-1}|| \, ||A_{\theta_2}^{-1}|| \, ||X^\top|| \, ||D_\mu|| \, ||X|| \, ||P_{\theta_1}^\top - P_{\theta_2}^\top||.
\end{aligned}
$$

Due to Assumption 3, $||P_{\theta_1}^\top - P_{\theta_2}^\top|| < C_1 ||\theta_1 - \theta_2||$ for some constant $C_1 < \infty$. The rest follows easily. □

A.7 Proof of Lemma 3

Proof. We first make a transformation of the original noise. We define

$$
\hat{r}^\theta(s, a) \doteq w_\theta^\top x(s) u_\theta^\top \tilde{x}(s, a) \psi_\theta(s, a) \in \mathbb{R}^K.
$$

Proposition 4 and Assumption 3 imply that there exists a constant $C_1 < \infty$ such that

$$
\forall (\theta, s, a), \; ||\hat{r}^\theta(s, a)|| < C_1.
$$

For each $i \in \{1, \dots, K\}$, we consider an MDP where the state space is $\mathcal{S}$, the action space is $\mathcal{A}$, the transition kernel is p, and the reward function is $\hat{r}_i^\theta(s, a)$. Under the i-th MDP, the average reward of the behavior policy μ is

$$
\bar{r}_i(\theta) \doteq \sum_s d_\mu(s) \sum_a \mu(a|s) \hat{r}_i^\theta(s, a) = \hat{g}_i(\theta).
$$

We consider the differential state-action value function $\hat{q}_i^\theta(s, a)$ of this MDP, where

$$
\hat{q}_i^\theta(s, a) \doteq \mathbb{E}_\mu \Big[\sum_{k=0}^{\infty} \Big(\hat{r}_i^\theta(S_k, A_k) - \bar{r}_i(\theta) \Big) \mid S_0 = s, A_0 = a \Big].
$$

According to the standard MDP theory (e.g., Section 8.2.1 in Puterman 2014), we have

$$\hat{q}_i^\theta = H_\mu \hat{r}_i^\theta, \tag{8}$$

where $\hat{q}_i^\theta \in \mathbb{R}^{N_{sa}}$, $\hat{r}_i^\theta \in \mathbb{R}^{N_{sa}}$, and $H_\mu \doteq (I - P_\mu + P_\mu^*)^{-1}(I - P_\mu^*) \in \mathbb{R}^{N_{sa} \times N_{sa}}$ refers to the fundamental matrix which depends only on μ and p. Here $P_\mu[(s,a),(s',a')] \doteq p(s'|s,a)\mu(a'|s')$ and each row of P_μ^* is $\tilde{d}_\mu$. The corresponding Bellman equation for $\hat{q}_i^\theta$ is

$$\hat{q}_i^\theta(s,a) = \left(\hat{r}_i^\theta(s,a) - \bar{r}_i(\theta)\right) + \sum_{s',a'} \tilde{p}(s',a'|s,a)\hat{q}_i^\theta(s',a'). \tag{9}$$

From Eq (8), we have

$$|\hat{q}^\theta(s,a)| < C_1 < \infty, \quad \forall (\theta, s, a) \tag{10}$$

for some constant C_1. For a fixed (s,a), $\hat{r}^\theta(s,a)$ is a product of three bounded Lipschitz continuous functions (Assumption 3 and Proposition 4). It is, therefore, also Lipschitz continuous. Eq (8) and the fact we only have finite states and actions imply

$$||\hat{q}^{\theta_1}(s,a) - \hat{q}^{\theta_2}(s,a)|| \leq C_1 ||\theta_1 - \theta_2|| \quad \forall (s, a, \theta_1, \theta_2) \tag{11}$$

for some constant C_1.

Now we are ready to decompose the noise $\nabla J(\theta_t)^\top (m_t q_t \psi_t - \hat{g}(\theta_t))$ as

$$\begin{aligned}
&\nabla J(\theta_t)^\top (m_t q_t \psi_t - \hat{g}(\theta_t)) \\
=&\nabla J(\theta_t)^\top (\hat{r}^{\theta_t}(S_t, A_t) - \bar{r}(\theta_t)) \quad \text{(Definition of } \hat{r}^{\theta_t} \text{ and } \bar{r}(\theta_t)) \\
=&\nabla J(\theta_t)^\top \Big(\hat{q}^{\theta_t}(S_t, A_t) - \sum_{s',a'} p(s'|S_t, A_t)\mu(a'|s')\hat{q}^{\theta_t}(s', a')\Big) \quad \text{(Eq (9))} \\
=&\sum_{i=1}^4 \epsilon_t^{(i)},
\end{aligned}$$

where

$$\begin{aligned}
\epsilon_t^{(1)} &\doteq \nabla J(\theta_t)^\top \Big(\hat{q}^{\theta_t}(S_{t+1}, A_{t+1}) - \sum_{s',a'} p(s'|S_t, A_t)\mu(a'|s')\hat{q}^{\theta_t}(s', a')\Big), \\
\epsilon_t^{(2)} &\doteq \frac{\beta_{t-1}\nabla J(\theta_{t-1})^\top \hat{q}^{\theta_{t-1}}(S_t, A_t) - \beta_t \nabla J(\theta_t)^\top \hat{q}^{\theta_t}(S_{t+1}, A_{t+1})}{\beta_t}, \\
\epsilon_t^{(3)} &\doteq \frac{\beta_t - \beta_{t-1}}{\beta_t} \nabla J(\theta_{t-1})^\top \hat{q}^{\theta_{t-1}}(S_t, A_t), \\
\epsilon_t^{(4)} &\doteq \nabla J(\theta_t)^\top \hat{q}^{\theta_t}(S_t, A_t) - \nabla J(\theta_{t-1})^\top \hat{q}^{\theta_{t-1}}(S_t, A_t).
\end{aligned}$$

We now show $|\sum_t \beta_t \epsilon_t^{(i)}| < \infty \quad a.s.$ for $i = 1, 2, 3, 4$.

(1) We first state a Martingale Convergence Theorem (see Proposition 4.3 in Bertsekas and Tsitsiklis 1996).

Lemma 4. *Assuming $\{M_l\}_{l=1,...}$ is a Martingale sequence and there exists a constant $C_1 < \infty$ such that $\forall l, \mathbb{E}[|M_l|^2] < C_1$, then $\{M_l\}$ converges almost surely.*

Let $\mathcal{F}_l \doteq \sigma(S_0, A_0, \theta_0, \dots, S_l, A_l, \theta_l, S_{l+1}, A_{l+1})$ be the σ-algebra and $M_l \doteq \sum_{t=0}^l \beta_t \epsilon_t^{(1)}$. It is easy to see that M_l is adapted to $\mathcal{F}_l$. Due to Lemma 1 and Eq (10), $|\epsilon_t^{(1)}| < C_1$, implying $\mathbb{E}[|M_l|] < \infty$ holds for any fixed l. Moreover,

$$\begin{aligned}
\mathbb{E}[M_{l+1}|\mathcal{F}_l] &= M_l + \mathbb{E}_{\theta_{l+1}, S_{l+2}, A_{l+2}}[\beta_{l+1}\epsilon_{l+1}^{(1)}|\mathcal{F}_l] \\
&= M_l + \beta_{l+1}\mathbb{E}_{\theta_{l+1}}\Big[\mathbb{E}_{S_{l+2}, A_{l+2}}[\epsilon_{l+1}^{(1)}|\theta_{l+1}, \mathcal{F}_l]\Big] \\
&= M_l + \beta_{l+1}\mathbb{E}_{\theta_{l+1}}[0] = M_l
\end{aligned}$$

M_l is therefore a Martingale. We now verify that M_l has bounded second moments, then $\{M_l\}$ converges according to Lemma 4. For any $t_1 < t_2$, we have

$$\mathbb{E}[\epsilon_{t_1}^{(1)}\epsilon_{t_2}^{(1)}] = \mathbb{E}\Big[\mathbb{E}[\epsilon_{t_1}^{(1)}\epsilon_{t_2}^{(1)}|\mathcal{F}_{t_2-1}]\Big] = \mathbb{E}\Big[\epsilon_{t_1}^{(1)}\mathbb{E}[\epsilon_{t_2}^{(1)}|\mathcal{F}_{t_2-1}]\Big] = \mathbb{E}\Big[\epsilon_{t_1}^{(1)}0\Big] = 0.$$

Consequently,

$$\forall l, \quad \mathbb{E}[|M_l|^2] = \mathbb{E}[\sum_{t=0}^{l}\beta_t^2\big(\epsilon_t^{(1)}\big)^2] \leq C_1\sum_{t=0}^{\infty}\beta_t^2 < \infty$$

for some constant C_1. Therefore, $\{M_l\}$ indeed converges and $|\sum_t \beta_t\epsilon_t^{(1)}| < \infty \quad a.s.$

(2) $\sum_{t=1}^{l}\beta_t\epsilon_t^{(2)} = \beta_0\nabla J(\theta_0)^\top\hat{q}^{\theta_0}(S_1, A_1) - \beta_l\nabla J(\theta_l)^\top\hat{q}^{\theta_l}(S_{l+1}, A_{l+1})$. The rest follows from the boundedness of $\nabla J(\theta)$ and $\hat{q}^\theta(s, a)$ (Lemma 1 and Eq (10)).

(3)

$$\begin{aligned}|\sum_{t=1}^{l}\beta_t\epsilon_t^{(3)}| &\leq \sum_{t=1}^{l}|\beta_t - \beta_{t-1}|\,|\nabla J(\theta_{t-1})^\top\hat{q}^{\theta_{t-1}}(S_t, A_t)| \\ &\leq C_1\sum_{t=1}^{l}(\beta_{t-1} - \beta_t) \leq C_1(\beta_0 - \beta_l) < C_1\beta_0 \quad a.s.\end{aligned}$$

(4) Eq (11), Eq (10) and Lemma 1 imply $\nabla J(\theta)^\top\hat{q}^\theta(S_t, A_t)$ is Lipschitz continuous in θ, yielding

$$|\epsilon_t^{(4)}| < C_1||\theta_t - \theta_{t-1}|| = C_1||\beta_t m_t q_t\psi_t|| \leq \beta_t C_2,$$

where the last inequality comes from the Assumption 3 and Proposition 4. Consequently,

$$|\sum_{t=1}^{l}\beta_t\epsilon_t^{(4)}| < C_2\sum_{t=1}^{l}\beta_t^2 < C_2\sum_{t=1}^{\infty}\beta_t^2 < \infty \quad a.s.$$

□

A.8 Proof of Theorem 1

Proof. This proof is standard and follows the same routine as Konda (2002). We first rewrite the update as

$$\begin{aligned}\theta_{t+1} &= \theta_t + \beta_t m_t q_t\psi_t \\ &= \theta_t + \beta_t\Big(\nabla J(\theta_t) - \hat{g}(\theta_t) - b^{(1)}(\theta_t) - b^{(2)}(\theta_t)\Big) + \beta_t m_t q_t\psi_t\end{aligned}$$

Using the second order Taylor expansion and $y_1^\top y_2 \leq ||y_1||\,||y_2||$, we have

$$\begin{aligned}J(\theta_{t+1}) \geq &J(\theta_t) + \beta_t||\nabla J(\theta_t)||^2 \\ &- \beta_t||\nabla J(\theta_t)||\,||b^{(1)}(\theta_t)|| - \beta_t||\nabla J(\theta_t)||\,||b^{(2)}(\theta_t)|| \\ &+ \beta_t\nabla J(\theta_t)^\top(m_t q_t\psi_t - \hat{g}(\theta_t)) \\ &- \frac{1}{2}C_1||\beta_t m_t q_t\psi_t||^2,\end{aligned}$$

where C_1 reflects the bound of the Hessian. Due to Assumption 3 and Proposition 4, $|m_t q_t\psi_t|^2$ is bounded by some constant $C_2 < \infty$ for all t. Therefore,

$$\sum_t||\beta_t m_t q_t\psi_t||^2 \leq C_2^2\sum_t\beta_t^2 < \infty \quad a.s. \tag{12}$$

Lemma 3 states

$$|\sum_t\beta_t\nabla J(\theta_t)^\top(m_t q_t\psi_t - \hat{g}(\theta_t))| < \infty \quad a.s. \tag{13}$$

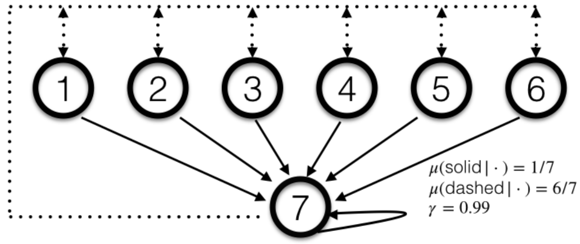

Figure 1: A variant of Baird's counterexample. This figure is adapted from Sutton and Barto (2018). The `solid` action always leads to the state 7 and a reward 0, and the `dashed` action leads to states 1 - 6 with equal probability and a reward $+1$.

If Theorem 1 does not hold, there must exist $t_0 > 0, \epsilon > 0$ such that

$$||\nabla J(\theta_t)|| - \Big(||b^{(1)}(\theta_t)|| + ||b^{(2)}(\theta_t)||\Big) > \epsilon$$

holds for all $t \geq t_0$. Consequently,

$$\begin{aligned} J(\theta_{t+1}) \geq & J(\theta_t) + \beta_t \epsilon^2 \\ & + \beta_t \nabla J(\theta_t)^\top (m_t q_t \psi_t - \hat{\nabla}(\theta_t)) \\ & - \frac{1}{2} C_3 ||\beta_t m_t q_t \psi_t||^2. \end{aligned} \tag{14}$$

holds for all $t \geq t_0$. Together with Eq (12), Eq (13) and $\sum_t \beta_t = \infty$, iterating Eq (14) implies

$$\lim_{t\to\infty} J(\theta_t) = \infty,$$

which contradicts the fact that $J(\theta)$ is always bounded for all θ as $\gamma < 1$ and $|r(s,a)| < R_{\max}$. $\square$

B Experiments

We design experiments to answer the following questions:

- Can GEM approximate the emphasis as promised?
- Can the learned emphasis boost performance?

We consider variants of Baird's counterexample (Baird, 1995; Sutton and Barto, 2018) as shown in Figure 1. In Baird's counterexample, there are two actions and the behavior policy μ always chooses the `dashed` action with probability $\frac{6}{7}$. The initial state is chosen from all the states with equal probability, and the interest i is 1 for all states. We consider four different sets of features: original features, one-hot features, zero-hot features, and aliased features. Original features are the features used by Sutton and Barto (2018), where the feature for each state lies in $\mathbb{R}^8$ (We will detail these features in the end of the appendix). This set of features is somehow uncommon as the number of states is usually much larger than the number of features in practice. One-hot features use one-hot encoding, where each feature lies in $\mathbb{R}^7$, which indeed degenerates to a tabular setting. Zero-hot features are the complements of one-hot features, e.g., the feature of the state 1 is $[0,1,1,1,1,1,1]^\top \in \mathbb{R}^7$. The quantities of interest, e.g., m_π and v_π, can always be expressed accurately under all the three sets of features. In the fourth set of features, we consider state aliasing, which is common in practical settings. In Baird's counterexample, the states 1-6 are equivalent. We therefore alias the state 7 to the state 6. To be more specific, we still consider the original features but now the feature of the state 7 is modified to be identical as the feature of the state 6. The last two dimensions of features then become identical for all states and we therefore removed them, resulting in features lying in $\mathbb{R}^6$. Now there is no guarantee that the quantities of interest still lie in the feature space.

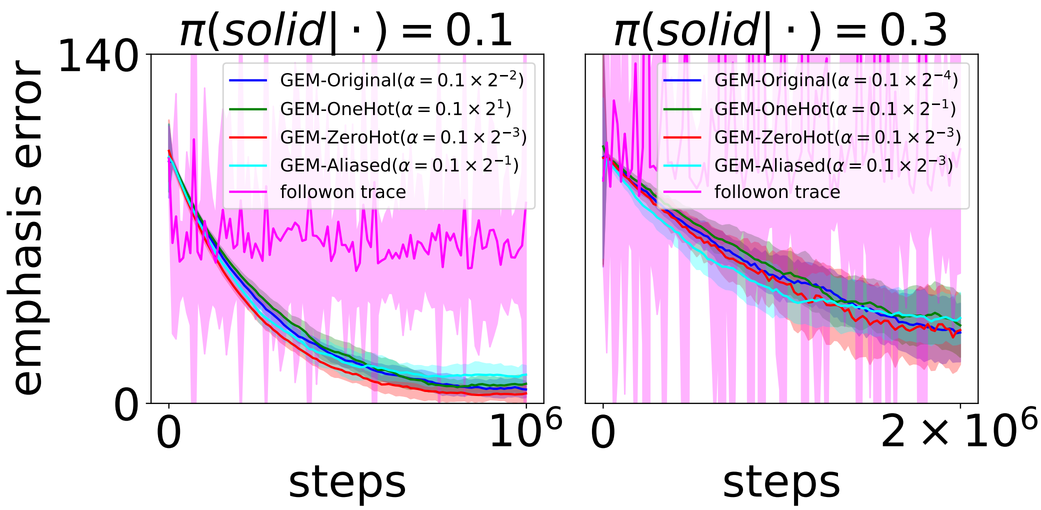

Figure 2: Averaged emphasis approximation error in recent 1000 steps for the followon trace and GEM with different features. Curves are averaged over 30 independent runs. Shadowed regions indicate one standard derivation. Learning rates are bracketed.

B.1 Approximating Emphasis

In this section, we compare the accuracy of approximating the emphasis m_π with GEM (Eq. (4)) and the followon trace (Eq. (1)). We report the emphasis approximation error in Figure 2. At time step t, the emphasis approximation error is computed as $|M_t - m_\pi(S_t)|$ and $|w_t^\top x(S_t) - m_\pi(S_t)|$ for the followon trace and GEM respectively, where the oracle m_π is computed analytically, $M_{-1} = 0$, and w_0 is drawn from a unit normal distribution. For GEM, we tune the learning rate α from $\{0.1 \times 2^1, \ldots, 0.1 \times 2^{-6}\}$. We consider two target policies: $\pi(\texttt{solid}|\cdot) = 0.1$ and $\pi(\texttt{solid}|\cdot) = 0.3$.

As shown in Figure 2, the GEM approximation enjoys lower variance than the followon trace approximation and has lower approximation error under all four sets of features. It is interesting to note that when the original features are used, the C matrix is indeed singular, which violates the Assumption 2. However, the algorithm does not diverge. This may suggest that the Assumption 2 can be relaxed in practice.

B.2 Policy Evaluation with GEM

The followon trace M_t is originally used in ETD to reweigh updates (Eq (1) and Eq (2)). Here we compare two algorithms, ETD(0) and GEM-ETD(0). In GEM-ETD(0) (Algorithm 2), instead of using M_t, we use $w_t^\top x(S_t)$ to reweigh updates. To make a fair comparison with ETD(0), we formulate GEM-ETD(0) in a two-timescale form in Algorithm 2. If we assume m_π lies in the spanning space of X, a convergent analysis of a bi-level optimization version of GEM-ETD(0) will be straightforward.

We consider a target policy $\pi(\texttt{solid}|\cdot) = 0.05$. We report the root mean squared value error (RMSVE) at each time step during training in Figure 3. RMSVE is computed as $||v - v_\pi||_{d_\mu}$, where v_π is computed analytically. For ETD(0), we tune the learning rate α from $\{0.1 \times 2^0, \ldots, 0.1 \times 2^{-19}\}$. For GEM-ETD(0), we set $\alpha_1 = 0.025$ and tune α_2 in the same range as the α. For both algorithms, we report the results with learning rates that minimized the area under curve (AUC) in solid lines in Figure 3.

In our policy evaluation experiments, GEM-ETD(0) has a clear win over ETD(0) under all four sets of features. Note the AUC-minimizing learning rate for ETD(0) is usually several orders smaller than that of GEM-ETD(0), which explains why ETD(0) curves tend to have smaller variance than GEM-ETD(0) curves. When we decrease the learning rate of GEM-ETD(0) (as indicated by the red dashed lines in Figure 3), the variance of GEM-ETD(0) can be reduced and the AUC is still smaller than that of ETD(0).

Algorithm 2: GEM-ETD(0) with linear function approximation

Input:
ν: parameters for approximating v_π
κ, w: parameters of GEM
α_1, α_2: learning rates

Get S_0 and set $t \leftarrow 0$
while *True* **do**
  Get $A_t \sim \mu(\cdot|S_t)$
  Execute A_t and get R_{t+1}, S_{t+1}
  $\kappa_{t+1} \leftarrow \kappa_t + \alpha_1(i(S_{t+1}) + \gamma \rho_t x_t^\top w_t - x_{t+1}^\top w_t - x_{t+1}^\top \kappa_t)x_{t+1}$
  $w_{t+1} \leftarrow w_t + \alpha_1(x_{t+1} - \gamma \rho_t x_t)x_{t+1}^\top \kappa_t$
  $\nu_{t+1} \leftarrow \nu_t + \alpha_2 x_t^\top w_{t+1} \rho_t (R_{t+1} + \gamma x_{t+1}^\top \nu_t - x_t^\top \nu_t)x_t$
  $t \leftarrow t + 1$
end

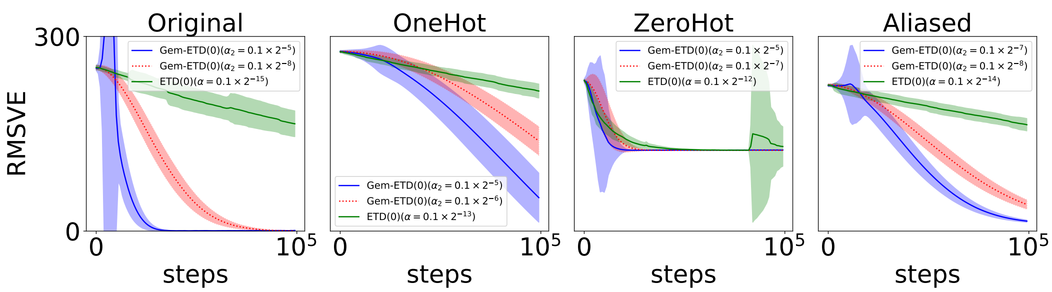

Figure 3: Averaged RMSVE in recent 1000 steps for GEM-ETD(0) and ETD(0) with four different sets of features. Curves are averaged over 30 independent runs. Shadowed regions indicate one standard derivation.

ETD(0) is the simplest emphatic algorithm and is a special case of ETD(λ, β) (Hallak et al., 2016), where λ and β are used for bias-variance trade-off. Similarly, we can have GEM-ETD(λ, β) by introducing λ and β to our GEM operator $\hat{\mathcal{T}}$ in the same manner as ETD(λ, β). A comparison between ETD(λ, β) and GEM-ETD(λ, β) is a possibility for future work.

GEM-ETD is indeed a way for bias-variance trade-off. If the states are heavily aliased, the GEM emphasis estimation may be heavily biased, so does GEM-ETD. We do not aim to claim that GEM-ETD is always better than ETD. For example, if we consider the original Baird's counterexample, where the target policy is $\pi(\texttt{solid}|\cdot) = 1$, there is no observable progress for both GEM-ETD(0) and ETD(0) with reasonable computation resources (This target policy is problematic for GEM-ETD(0) mainly because the corresponding $\bar{\delta}_w$ tends to be highly imbalanced, i.e., one dimension can be much larger than the others. Consequently, the supervised learning process of κ becomes problematic.). When it comes to bias-variance trade-off, the optimal choice is usually task-dependent. And our empirical results do suggest GEM-ETD is a promising approach for this trade-off.

C Related Work

Off-Policy Actor-Critic (Degris et al., 2012) is the first provably convergent off-policy actor-critic algorithm in the tabular setting [3] and inspired the invention of many other off-policy actor-critic algorithms, e.g., (Deep) Deterministic Policy Gradient (Silver et al., 2014; Lillicrap et al., 2015), Actor-Critic with Experience Replay (Wang et al., 2016), Interpolated Policy Gradient (IPG, Gu et al. 2017), Off-policy Expected Policy Gradients (Ciosek and Whiteson, 2017), and IMPALA (Espeholt et al., 2018). However, none of them has a convergent analysis under function approximation.

[3] See Errata in Degris et al. (2012)

Previously, Maei (2018) proposes the Gradient Actor Critic under a different objective $\sum_s d_\mu(s)v(s)$ for off-policy training with function approximation. This objective is different from the excursion objective in that it replaces the true value function v_π with an estimate v. Furthermore, the policy gradient estimator proposed by Maei (2018) tracks the true gradient only in a limiting sense for a fixed π (see Theorem 2 in Maei (2018)) and has potentially unbounded variance, similar to how M_t tracks $m_\pi(S_t)$. It is questionable whether that policy gradient estimator can track the true policy gradient under a changing π.

D Original Features of Baird's Counterexample

According to Sutton and Barto (2018), we have

$$
\begin{aligned}
x(s_1) &\doteq [2, 0, 0, 0, 0, 0, 0, 1]^\top \\
x(s_2) &\doteq [0, 2, 0, 0, 0, 0, 0, 1]^\top \\
x(s_3) &\doteq [0, 0, 2, 0, 0, 0, 0, 1]^\top \\
x(s_4) &\doteq [0, 0, 0, 2, 0, 0, 0, 1]^\top \\
x(s_5) &\doteq [0, 0, 0, 0, 2, 0, 0, 1]^\top \\
x(s_6) &\doteq [0, 0, 0, 0, 0, 2, 0, 1]^\top \\
x(s_7) &\doteq [0, 0, 0, 0, 0, 0, 1, 2]^\top
\end{aligned}
$$