Consistent On-Line Off-Policy Evaluation

Assaf Hallak IFOGPH@GMAIL.COM
Technion Institute of Technology
Haifa, Israel

Shie Mannor IFOGPH@GMAIL.COM
Technion Institute of Technology
Haifa, Israel

Abstract

The problem of on-line off-policy evaluation (OPE) has been actively studied in the last decade due to its importance both as a stand-alone problem and as a sub-module in a policy improvement scheme. However, most Temporal Difference (TD) based solutions ignore the discrepancy between the stationary distribution of the behavior and target policies and its effect on the convergence limit when linear function approximation is applied. In this paper we propose Consistent Off-Policy TD (COP-TD) that addresses this issue directly and enables reducing this bias at some computational expense. We show that COP-TD(λ, β) can be designed to converge to the same value that would have been obtained by using on-policy TD(λ) with the target policy. Subsequently, the proposed scheme leads to a related and promising heuristic we call log-COP-TD(λ, β). Both algorithms have favorable empirical results to the current state of the art on-line OPE algorithms.

1. Introduction

In this paper we address the problem of assessing the performance of a complex strategy without trying it known as off-policy evaluation (OPE). OPE formulation is often considered in domains with limited sampling capability. For example, marketing and recommender systems (Theocharous and Hallak, 2013; Theocharous et al., 2015) directly relate policies to revenue. A more extreme example is drug administration, where there are only few patients in the testing population, and sub-optimal policies can have life threatening effects (Hochberg et al., 2016).

We consider the OPE problem in an on-line setup where each new sample is immediately used to update our current value estimate of some previously unseen policy. We propose a new algorithm called COP-TD(λ,β) for estimating the value of the target policy which is consistent with sufficient resources: it converges to the value on-policy learning with the target policy would have converged to. In addition, we introduce a related heuristic called Log-COP-TD(λ,β) and explain its motivation.

2. Notations and Background

We consider the standard discounted Markov Decision Process (MDP) formulation Bertsekas and Tsitsiklis (1996) with a single long trajectory. Let $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \zeta, \gamma)$ be an MDP where $\mathcal{S}$ is the finite state space and $\mathcal{A}$ is the finite action space. The parameter $\mathcal{P}$ sets the transition probabilities $P(s'|s,a)$ given the previous state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, where the first state is determined by the distribution ζ. The parameter $\mathcal{R}$ sets the reward distribution $r(s,a)$ obtained by taking action a in state sn and γ is the discount factor specifying the exponential reduction in reward with time.

The process advances as follows: A state s_0 is sampled according to the distribution $\zeta(s)$. Then, each time step t starting from $t = 0$ the agent chooses an action a_t according to some distribution (defined as the policy) $\mu(a|s_t)$ (the behavior policy), a reward $r_t \doteq r(s_t, a_t)$ is accumulated by the agent, and the next state s_{t+1} is sampled using the transition probability $P(s'|s_t, a_t)$.

The expected discounted accumulated reward starting from a specific state and choosing an action by some policy π is called the value function, which is also known to hold the Bellman equation in a vector form: $V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t\right], T_\pi V \doteq R_\pi + \gamma P_\pi V$, where $[R_\pi]_s \doteq \mathbb{E}_\pi\left[r(s, \pi(s))\right]$ and $[P_\pi]_{s,s'} \doteq \mathbb{E}_\pi\left[P(s'|s, \pi(s))\right]$ are the policy induced reward vector and transition probability matrix respectively; T_π is called the Bellman operator. The problem of estimating $V^\pi(s)$ from samples is called policy evaluation. If the target policy π is different than the behavior policy μ which generated the samples, the problem is called off-policy evaluation (OPE).

The TD(λ) (Sutton, 1988) algorithm is a standard solution to on-line on-policy evaluation: Each time step the TD error updates the current value function estimate, so eventually the process will converge to the true value function. The standard form of TD(λ) is given by:

$$\begin{gathered} R_{t,s_t}^{(n)} = \sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n \hat{V}_t(s_{t+n}), \qquad R_{t,s_t}^{\lambda} = (1-\lambda) \sum_{n=0}^{\infty} \lambda^n R_{s_t}^{(n+1)}, \\ \hat{V}_{t+1}(s_t) = \hat{V}_t(s_t) + \alpha_t \left(R_{t,s_t}^{\lambda} - \hat{V}_t(s_t) \right), \end{gathered} \tag{1}$$

where α_t is the learning rate. The value $R_{t,s_t}^{(n)}$ is an estimate of $V(s_t)$, looking forward n steps, and R_{t,s_t}^{λ} is an exponentially weighted average of all of these estimates going forward till infinity.

We denote by $d_\mu(s)$ the stationary distribution over states induced by taking the policy μ and mark $D_\mu = diag(d_\mu)$. Since we are concerned with the behavior at the infinite horizon, we assume $\zeta(s) = d_\mu(s)$. In addition, we assume the MDP is ergodic for the two specified policies μ, π so $\forall s \in \mathcal{S} : d_\mu(s) > 0, d_\pi(s) > 0$ and that the OPE problem is proper - $\mu(a|s) > 0 \iff \pi(a|s) > 0$.

When the state space is too large to hold $V^\pi(s)$, a linear function approximation scheme is used: $V^\pi(s) \approx \theta_\pi^\top \phi(s)$, where θ is the optimized weight vector and $\phi(s)$ is the feature vector of state s composed of k features; in this paper we assume the features are linearly independent. TD(λ) can be adjusted accordingly to find the fixed point of $\Pi_{d_\pi} T_\pi^\lambda$ where Π_{d_π} is the projection to the subspace spanned by the features with respect to the d_π-weighted norm:

$$\begin{gathered} R_{t,s_t}^{(n)} = \sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n \theta_t^\top \phi(s_{t+n}), \qquad R_{t,s_t}^{\lambda} = (1-\lambda) \sum_{n=0}^{\infty} \lambda^n R_{s_t}^{(n+1)} \\ \theta_{t+1} = \theta_t + \alpha_t \left(R_{t,s_t}^{\lambda} - \theta_t^\top \phi(s_t) \right) \phi(s_t). \end{gathered} \tag{2}$$

Finally, we define OPE-related quantities: $\rho_t \doteq \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}, \Gamma_t^n \doteq \prod_{i=0}^{n-1} \rho_{t-1-i}, \rho_d(s) \doteq \frac{d_\pi(s)}{d_\mu(s)}$, we call ρ_d the stationary distribution ratio.

3. Previous Work

We can roughly categorize previous OPE algorithms to two main families: Gradient based methods that perform stochastic gradient descent on error terms they want to minimize (Sutton et al., 2009a,b; White and White, 2016), and Importance Sampling (IS) methods that correct the gains between on-policy and off-policy updates using the IS-ratios ρ_t (Precup et al., 2001; Sutton et al., 2015).

There are also a few algorithms that fall between the two, (van Hasselt et al., 2014; Mahmood and Sutton, 2015). Our work falls within the IS family.

A comparison of these algorithms in terms of convergence rate is available in (White and White, 2016; Geist and Scherrer, 2014). We focus on the limit point of the convergence. For most of the TD algorithms, the process was shown to converge almost surely to the fixed point of the projected Bellman operator $\Pi_d T_\pi$ where d is some stationary distribution (usually d_μ), however the d in question was never[1] d_π as we would have obtained from running on-policy TD with the target policy. The algorithm achieving the closest result is ETD(λ,β) which replaced d with $f = \left(I - \beta P_\pi^\top\right)^{-1} d_\mu$, where β trades-off some of the process' variance with the bias in the limit point.

4. COP-TD(λ, β)

Most off-policy algorithms multiply the TD summand of TD(λ) with some value that depends on the history and the current state. For example, full IS-TD by Precup et al. (2001) examines the ratio between the trajectory's probabilities under both policies: $\frac{P_\pi(s_0,a_0,s_1,\dots,s_t,a_t)}{P_\mu(s_0,a_0,s_1,\dots,s_t,a_t)} = \prod_{m=0}^{t} \rho_m \;= \Gamma_t^t \rho_t$.

In problems with a long horizon, or these that start from the stationary distribution, we suggest using a time-invariant term which is the stationary distribution ratio ρ_d multiplied by the current ρ_t. This leads us to the following update equations:

$$\theta_{t+1} = \theta_t + \alpha_t \rho_d(s_t)\rho_t \left(r_t + \theta_t^\top \left(\gamma\phi(s_{t+1}) - \phi(s_t)\right)\right)\phi(s_t). \tag{3}$$

Lemma 1 *If the step sizes α_t satisfy $\sum_{t=0}^{\infty} \alpha_t = \infty, \sum_{t=0}^{\infty} \alpha_t^2 < \infty$ then the process described by Equation 3 converges almost surely to the fixed point of $\Pi_\pi T_\pi V = V$.*

In practice, we have no way of knowing $\rho_d(s)$, therefore we suggest estimating it using an additional stochastic approximation process. In order to do so, we note that the quantity Γ_t^n that is commonly used in IS-based algorithm satisfies the following, for any function on the state space $u(s)$: $\mathbb{E}_\mu\left[\Gamma_t^n u(s_{t-n}) | s_t\right] = u^\top D_\mu P_\pi^n D_\mu^{-1} e_{s_t}$, where e_{s_t} is the unit vector of state s_t. So, for an unbiased estimate of ρ_d denoted $\widehat{\rho_d}$ we can define and derive: $\tilde{\Gamma}_t^n \doteq \widehat{\rho_d}(s_{t-n})\Gamma_t^n = \widehat{\rho_d}(s_{t-n}) \prod_{i=0}^{n-1} \rho_{t-i-1} \qquad \Rightarrow \qquad \mathbb{E}_\mu\left[\tilde{\Gamma}_t^n | s_t\right] = \rho_d(s_t)$.

For any state s_t there are $t \to \infty$ such quantities $\{\tilde{\Gamma}_t^n\}_{n=0}^t$, where we propose to weight them similarly to TD(λ): $\tilde{\Gamma}^\beta = (1-\beta)\sum_{n=0}^{\infty} \beta^n \tilde{\Gamma}_t^{n+1}$.

Note that $\rho_d(s)$, unlike $V(s)$, is restricted to a close set since its d_μ-weighted linear combination is equal to 1 and all of its entries are non-negative; We denote this d_μ-weighted simplex by Δ_{d_μ}, and $\Pi_{\Delta_{d_\mu}}$ is the (non-linear) projection to this set with respect to the Euclidean norm ($\Pi_{\Delta_{d_\mu}}$ can be calculated efficiently, Chen and Ye (2011)). Now, we can devise a TD algorithm which estimates ρ_d and use it to find θ, which we call COP-TD(0, β) (Consistent Off-Policy TD).

Theorem 2 *If the step sizes satisfy $\sum_t \alpha_t = \sum_t \alpha_t^d = \infty, \sum_t (\alpha_t^2 + (\alpha_t^d)^2) < \infty, \frac{\alpha_t}{\alpha_t^d} \to 0, t\alpha_t^d \to 0$, and $\mathbb{E}\left[(\beta^n \Gamma_t^n)^2 | s_t\right] \le C$ for some constant C and every t, n, then after applying COP-TD(0, β), $\widehat{\rho}_{d,t}$ converges to ρ_d almost surely, and θ_t converges to the fixed point of $\Pi_\pi T_\pi V$.*

Notice that COP-TD(0, β) in its current form is infeasible in problems with large state spaces. Like TD(λ), we can introduce function approximation: represent $\rho_d(s) \approx \theta_\rho^\top \phi_\rho(s)$ where θ_ρ is a

[1] Except full IS, however its variance is too high to be applicable in practice.

Algorithm 1 COP-TD(0,β), Input: θ_0, $\widehat{\rho}_{d,0}$,

```
Init: $F_0 = 0, \quad n_0^\beta = 1, \quad N(s) = 0$
for $t = 1, 2, \ldots$ do
    Observe $s_t, a_t, r_t, s_{t+1}$
    Update normalization terms
    $n_t^\beta = \beta n_t^\beta + 1, \quad N(s_t) = N(s_t) + 1, \quad \forall s \in \mathcal{S} : \hat{d}_\mu(s) = \frac{N(s)}{t}$
    Update $\Gamma_t^n$'s weighted average:
    $F_t = \rho_{t-1}(\beta F_{t-1} + e_{s_{t-1}})$
    Update & project by $\rho_d$'s TD error:
    $\delta_t^d = \frac{F_t^\top \widehat{\rho}_{d,t}}{n_t^\beta} - \widehat{\rho}_{d,t}(s_t), \qquad \widehat{\rho}_{d,t+1} = \Pi_{\Delta_{\hat{d}_\mu}}\left(\widehat{\rho}_{d,t} + \alpha_t^d \delta_t^d e_{s_t}\right)$
    Off-policy TD(0):
    $\delta_t = r_t + \theta_t^\top(\gamma\phi(s_{t+1}) - \phi(s_t)), \qquad \theta_{t+1} = \theta_t + \alpha_t \widehat{\rho}_{d,t+1}(s_t)\rho_t\delta_t\phi(s_t)$
end for
```

weight vector and $\phi_\rho(s)$ is the off-policy feature vector and adjust the algorithm accordingly. For $\widehat{\rho}_d$ to still be contained in the set Δ_{d_μ}, we pose the requirement on the feature vectors: $\phi_\rho(s) \in \mathbb{R}_+^k$, and $\sum_s d_\mu(s)\theta_\rho^\top\phi_\rho(s) = 1$ (noted as the simplex projection $\Pi_{\Delta_{\mathbb{E}_\mu[\phi_\rho(s)]}}$). We provide the details in Algorithm 2, which also incorporates non-zero λ (similarly to ETD(λ,β)).

Algorithm 2 COP-TD(λ,β) with Function Approximation, Input: θ_0, $\theta_{\rho,0}$

```
Init: $F_0 = \underline{0}, \quad n_0^\beta = 1, \quad N_\phi = \underline{0}, \quad e_0 = \underline{0}$
for $t = 1, 2, \ldots$ do
    Observe $s_t, a_t, r_t, s_{t+1}$
    Update normalization terms:
    $n_t^\beta = \beta n_t^\beta + 1, \quad N_\phi = N_\phi + \phi_\rho(s_t), \quad \hat{d}_{\phi_\rho} = \frac{N_\phi}{t}$
    Update $\Gamma_t^n$'s weighted average:
    $F_t = \rho_{t-1}(\beta F_{t-1} + \phi_\rho(s_{t-1}))$
    Update & project by $\rho_d$'s TD error:
    $\delta_t^d = \theta_{\rho,t-1}^\top\left(\frac{F_t}{n_t^\beta} - \phi_\rho(s_t)\right), \qquad \theta_{\rho,t+1} = \Pi_{\Delta_{\hat{d}_{\phi_\rho}}}\left(\theta_{\rho,t} + \alpha_t^d\delta_t^d\phi_\rho(s_t)\right)$
    Off-policy TD($\lambda$):
    $M_t = \lambda + (1-\lambda)\theta_{\rho,t+1}^\top\phi_\rho(s_t), \qquad e_t = \rho_t\left(\lambda\gamma e_t + M_t\phi(s_{t+1})\right)$
    $\delta_t = r_t + \theta_t^\top(\gamma\phi(s_{t+1}) - \phi(s_t)) \qquad \theta_{t+1} = \theta_t + \alpha_t\delta_t e_t$
end for
```

Theorem 3 *If the step sizes hold $\sum_t \alpha_t = \sum_t \alpha_t^d = \infty$, $\sum_t(\alpha_t^2 + (\alpha_t^d)^2) < \infty$, $\frac{\alpha_t}{\alpha_t^d} \to 0, t\alpha_t^d \to 0$, and $\mathbb{E}\left[(\beta^n\Gamma_t^n)^2|s_t\right] \le C$ for some constant C and every t, n, then after applying COP-TD(0, β) with function approximation satisfying $\phi(s) \in \mathbb{R}_+^k$, $\widehat{\rho}_{d,t}$ converges to the fixed point of $\Pi_{\Delta_{\mathbb{E}_\mu[\phi_\rho(s)]}}\Pi_{\phi_\rho}Y^\beta$ denoted by ρ_d^{COP} almost surely, and θ_t converges to the fixed point of $\Pi_{d_\mu\circ\rho_d^{COP}}T_\pi V$.*

A possible criticism on COP-TD(0,β) is that it is not actually consistent, since in order to be consistent the original state space needs to be small, in which case every off-policy algorithm is consistent as well. Still, the dependence on another set of features allows to trade-off accuracy with computational power in estimating ρ_d and subsequently V.

5. The Logarithm Approach for Handling Long Products

Konidaris et al. (2011) suggested a statistical interpretation of TD(λ). They show that under several assumptions the TD(λ) estimate $R_{s_t}^{\lambda}$ is the maximum likelihood estimator of $V(s_t)$ given $R_{s_t}^{n}$: (1) Each $R_{s_t}^{n}$ is an unbiased estimator of $V(s_t)$ (2) The random variables $R_{s_t}^{n}$ are independent and specifically uncorrelated (3) The random variables $R_{s_t}^{n}$ are jointly normally distributed and (4) The variance of each $R_{s_t}^{n}$ is proportional to λ^n.

Assumption 3 is that the sampled estimators $(R^{(n)}, \Gamma_t^n)$ are normally distributed. For on policy TD(λ), this assumption might seem not too harsh as the estimators $R^{(n)}$ represent growing **sums** of random variables. However, in our case the estimators Γ_t^n are growing **products** of random variables. To correct this issue we can define new estimators using a logarithm on each $\tilde{\Gamma}_t^n$:

$$\log\left[\rho_d(s_t)\right] = \log\left[\mathbb{E}\left[\widehat{\rho_d}(s_{t-m})\prod_{k=t-m}^{t-1}\rho_k \;\middle|\; s_t\right]\right] \approx \log\left[\widehat{\rho_d}(s_{t-m})\right] + \sum_{k=t-m}^{t-1}\mathbb{E}\left[\log\left[\rho_k\right]|s_t\right]. \quad (4)$$

This approximation is crude – hence, we can relate to this method mainly as a well-motivated heuristic. Notice that this formulation resembles the standard MDP formulation, only with the corresponding "reward" terms $\log[\rho_t]$ going backward instead of forward, and no discount factor. Unfortunately, without a discount factor we cannot expect the estimated value to converge, so we propose using an artificial one $\gamma_{\log}$. We can incorporate function approximation for this formulation as well. Unlike COP-TD(λ, β), we can choose the features and weights as we wish with no restriction, besides the linear constraint on the resulting ρ_d through the weight vector θ_ρ. This can be approximately enforced by normalizing θ_ρ using $\frac{X}{t} \doteq \frac{1}{t}\sum_t \exp(\theta_{\rho,t}^\top \phi(s_t))$ (which should equal 1 if we were exactly correct). We call the resulting algorithm Log-COP-TD(λ,β).

Algorithm 3 Log-COP-TD(λ,β) with Function Approximation, Input: θ_0,$\theta_{\rho,0}$

```
Init: $F_0 = 0, \quad n_0(\beta) = 1, \quad N(s) = 0$
for $t = 1, 2, ...$ do
    Observe $s_t, a_t, r_t, s_{t+1}$
    Update normalization terms:
    $n_t^\beta = \beta n_t^\beta + 1, \quad N_\phi = \gamma_{\log}(\beta N_\phi + \phi_\rho(s_t)), \quad X = X + \exp(\theta_{\rho,t}^\top \phi(s_t))$
    Update $\log(\Gamma_t^n)$'s weighted average:
    $F_t = \beta\gamma_{\log}F_{t-1} + n_t^\beta \log[\rho(s_{t-1})]$
    Update & project by $\log(\rho_d)$'s TD error:
    $\delta_t^d = \frac{F_t}{n_t^\beta} + \theta_{\rho,t}^\top\left(\frac{N_\phi}{n_t^\beta} - \phi_\rho(s_t)\right), \theta_{\rho,t+1} = \theta_{\rho,t} + \alpha_t^d \delta_t^d \phi_\rho(s_t)$
    Off-policy TD($\lambda$):
    $M_t = \lambda + (1-\lambda)\exp\left(\theta_{\rho,t+1}^\top \phi_\rho(s_t)\right)/(X/t), \quad e_t = \rho_t\left(\lambda\gamma e_t + M_t\phi(s_{t+1})\right)$
    $\delta_t = r_t + \theta_t^\top(\gamma\phi(s_{t+1}) - \phi(s_t)), \quad \theta_{t+1} = \theta_t + \alpha_t\delta_t e_t$
end for
```

6. Experiments

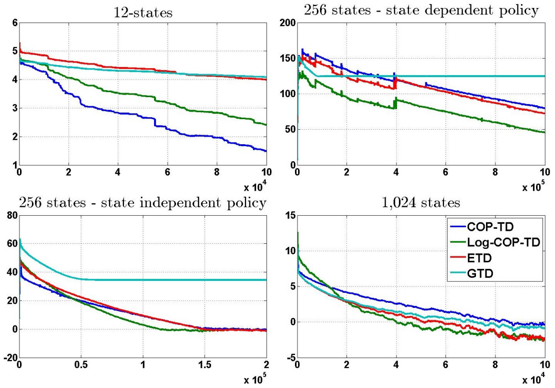
We have compared our algorithm to ETD(λ, β) and GTD(λ). Our method of comparison estimates the value function and finds the d_π weighted average of the error between V and the on-policy fixed point $\Pi_\pi T V_\pi$: $\|\hat{V} - \Pi_\pi T V_\pi\|^2_{d_\pi} = \sum_s d_\pi(s) \left[(\theta^* - \hat{\theta})^\top \phi(s)\right]^2$, where θ^* is the optimal θ obtained by on-policy TD using the target policy, and we apply log on the output to emphasize the results.

We build two MDPs with $256/1024$ states and 2 actions with uniformly distributed transition probabilities, where half of the states are with reward 0 and half with reward 1. The behavior / target policy chooses action 1 with probability 0.25 / 0.75 correspondingly, regardless of the state. The feature vectors encode each state using its binary 8-bit vector along with a free parameter, the same feature vector was used for estimating ρ_d.

Next, we build a simple 12-states MDP designed especially to emphasize COP-TD(λ, β)'s performance and the randomized 256 states MDP with feature-dependent policy: $\frac{\mu(a=1|s)}{\mu(a=2|s)} = \exp\left(\mathbf{1}^\top \phi(s)/3\right), \pi(a|s) = 1 - \mu(a|s)$, which was designed to emphasize Log-COP-TD(λ, β)'s performance (since $\phi(s)$ is the binary representation of s, this means ρ can have values as high as e^8). For each algorithm we have manually chosen the best λ,β parameters from the set $\{0, 0.25, 0.5, 0.75, 0.99\}$.

The specialized example is a 12 states chain MDP where the agent can move either right or left. In one policy the probability to move left is 0.85 and in the other it is 0.15. This kind of setting causes distinct differences in the stationary distribution since for each policy the process is much more concentrated on a different part of the chain. The rewards were set to be 0 for the left half of the chain and 1's for the other half. The feature space aggregates each 3 consecutive states to have the same 1 in one entry and 0 in the other 3 entries.

The experiments show that on the largest MDPs COP-TD(λ, β) and Log-COP-TD(λ, β) have comparable performance to ETD(λ, β) and GTD(λ) for the random MDPs. On the smaller problem and the 256 state dependent policy MDP, however, COP-TD(λ, β) and Log-COP-TD(λ, β) show better results due to the high discrepancy in the stationary distributions, or very high ρ values.



7. Conclusion

Research on off-policy evaluation has flourished in the last decade. While a plethora of algorithms were suggested so far, these does not converge to the same point achieved by on-line TD when function approximation is required. We address this issue with COP-TD(λ,β) and proved it can achieve consistency. Despite requiring a new set of features and calibrating an additional update function, COP-TD(λ,β)'s performance does not depend as much on β as ETD(λ,β), and shows promising empirical results.

8. Acknowledgments

This Research was supported in part by the Israel Science Foundation (grant No. 920/12) and by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ ERC Grant Agreement n.306638.

References

Hasan AA Al-Rawi, Ming Ann Ng, and Kok-Lim Alvin Yau. Application of reinforcement learning to routing in distributed wireless networks: a review. *Artificial Intelligence Review*, 43(3):381–416, 2015.

Enda Barrett, Enda Howley, and Jim Duggan. Applying reinforcement learning towards automating resource allocation and application scalability in the cloud. *Concurrency and Computation: Practice and Experience*, 25(12):1656–1674, 2013.

D. Bertsekas. *Dynamic Programming and Optimal Control, Vol II*. Athena Scientific, 4th edition, 2012.

D. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Yunmei Chen and Xiaojing Ye. Projection onto a simplex. *arXiv preprint arXiv:1101.6081*, 2011.

Matthieu Geist and Bruno Scherrer. Off-policy learning with eligibility traces: A survey. *The Journal of Machine Learning Research*, 15(1):289–333, 2014.

Hirotaka Hachiya, Masashi Sugiyama, and Naonori Ueda. Importance-weighted least-squares probabilistic classifier for covariate shift adaptation with application to human activity recognition. *Neurocomputing*, 80:93–101, 2012.

Assaf Hallak, Aviv Tamar, Remi Munos, and Shie Mannor. Generalized emphatic temporal difference learning: Bias-variance analysis. *arXiv preprint arXiv:1509.05172*, 2015.

Irit Hochberg, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Elad Yom-Tov. Encouraging physical activity in patients with diabetes through automatic personalized feedback via reinforcement learning improves glycemic control. *Diabetes care*, 39(4):e59–e60, 2016.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, page 0278364913495721, 2013.

George Konidaris, Scott Niekum, and Philip S Thomas. Td-gamma: Re-evaluating complex backups in temporal difference learning. In *Advances in Neural Information Processing Systems 24*, pages 2402–2410. Curran Associates, Inc., 2011. URL `http://papers.nips.cc/paper/4472-td_gamma-re-evaluating-complex-backups-in-temporal-difference-learning.pdf`.

A Rupam Mahmood and Richard S Sutton. Off-policy learning based on weighted importance sampling with linear computational complexity. In *Conference on Uncertainty in Artificial Intelligence*, 2015.

Doina Precup, Richard S Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In *ICML*, 2001.

R. S. Sutton, A. R. Mahmood, and M White. An emphatic approach to the problem of off-policy temporal-difference learning. *arXiv:1503.04269*, 2015.

Rich Sutton, Ashique R Mahmood, Doina Precup, and Hado V Hasselt. A new q (lambda) with interim forward view and monte carlo equivalence. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 568–576, 2014.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1):9–44, 1988.

Richard S Sutton, Hamid R Maei, and Csaba Szepesvári. A convergent o(n) temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in neural information processing systems*, pages 1609–1616, 2009a.

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, 2009b.

Georgios Theocharous and Assaf Hallak. Lifetime value marketing using reinforcement learning. *RLDM 2013*, page 19, 2013.

Georgios Theocharous, Philip S Thomas, and Mohammad Ghavamzadeh. Personalized ad recommendation systems for life-time value optimization with guarantees. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-15)*, 2015.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2380–2388, 2015a.

Philip S Thomas, Scott Niekum, Georgios Theocharous, and George Konidaris. Policy evaluation using the omega-return. In *Advances in Neural Information Processing Systems 28*, pages 334–342. Curran Associates, Inc., 2015b. URL `http://papers.nips.cc/paper/5807-policy-evaluation-using-the-return.pdf`.

Hado van Hasselt, A Rupam Mahmood, and Richard S Sutton. Off-policy td (λ) with a true online equivalence. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence, Quebec City, Canada*, 2014.

Adam White and Martha White. Investigating practical, linear temporal difference learning. *arXiv preprint arXiv:1602.08771*, 2016.

H. Yu. On convergence of emphatic temporal-difference learning. In *COLT*, 2015.