Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Generalized Emphatic Temporal Difference Learning: Bias-Variance Analysis

Assaf Hallak
Technion
Israel
ifogph@gmail.com

Aviv Tamar
UC Berkeley
USA
avivt@berkeley.edu

Remi Munos
Google DeepMind
UK
munos@google.com

Shie Mannor
Technion
Israel
shie@ee.technion.ac.il

Abstract

We consider the off-policy evaluation problem in Markov decision processes with function approximation. We propose a generalization of the recently introduced *emphatic temporal differences* (ETD) algorithm (Sutton, Mahmood, and White 2015), which encompasses the original ETD(λ), as well as several other off-policy evaluation algorithms as special cases. We call this framework ETD(λ, β), where our introduced parameter β controls the decay rate of an importance-sampling term. We study conditions under which the projected fixed-point equation underlying ETD(λ, β) involves a contraction operator, allowing us to present the first asymptotic error bounds (bias) for ETD(λ, β). Our results show that the original ETD algorithm always involves a contraction operator, and its bias is bounded. Moreover, by controlling β, our proposed generalization allows trading-off bias for variance reduction, thereby achieving a lower total error.

1 Introduction

In Reinforcement Learning (RL; Sutton and Barto, 1998), *policy-evaluation* refers to the problem of evaluating the value function – a mapping from states to their long-term discounted return under a given policy, using sampled observations of the system dynamics and reward. Policy-evaluation is important both for assessing the quality of a policy, but also as a sub-procedure for policy optimization.

For systems with large or continuous state-spaces, an exact computation of the value function is often impossible. Instead, an *approximate* value-function is sought using various function-approximation techniques (a.k.a. approximate dynamic-programming; Bertsekas, 2012). In this approach, the parameters of the value-function approximation are tuned using machine-learning inspired methods, often based on *temporal-differences* (TD;Sutton and Barto, 1998).

The source generating the sampled data divides policy evaluation into two cases. In the *on-policy* case, the samples are generated by the *target-policy* – the policy under evaluation; In the *off-policy* setting, a different *behavior-policy* generates the data. In the on-policy setting, TD methods are well understood, with classic convergence guarantees and approximation-error bounds, based on a contraction property of the projected Bellman operator underlying TD (Bertsekas and Tsitsiklis 1996). These bounds guarantee that the asymptotic error, or *bias*, of the algorithm is contained. For the off-policy case, however, standard TD methods no longer maintain this contraction property, the error bounds do not hold, and these methods might even diverge (Baird 1995).

The standard error-bounds may be shown to hold for an *importance-sampling* TD method (IS-TD), as proposed by Precup, Sutton, and Dasgupta (2001). However, this method is known to suffer from a high variance of its importance-sampling estimator, limiting its practicality.

Lately, Sutton, Mahmood, and White (2015) proposed the *emphatic TD* (ETD) algorithm: a modification of the TD idea, which converges off-policy (Yu 2015), and has a reduced variance compared to IS-TD. This variance reduction is achieved by incorporating a certain decay factor over the importance-sampling ratio. However, to the best of our knowledge, there are no results that bound the bias of ETD. Thus, while ETD is assured to converge, it is not known how good its limit actually is.

In this paper, we propose the ETD(λ, β) framework – a modification of the ETD(λ) algorithm, where the decay rate of the importance-sampling ratio, β, is a free parameter, and λ is the same bootstrapping parameter employed in TD(λ) and ETD(λ). By varying the decay rate, one can smoothly transition between the IS-TD algorithm, through ETD, to the standard TD algorithm.

We investigate the bias of ETD(λ, β), by studying the conditions under which its underlying projected Bellman operator is a contraction. We show that the original ETD possesses a contraction property, and present the first error bounds for ETD and ETD(λ, β). In addition, our error bound reveals that the decay rate parameter balances between the bias and variance of the learning procedure. In particular, we show that selecting a decay equal to the discount factor as in the original ETD may be suboptimal in terms of the mean-squared error.

The main contributions of this work are therefore a unification of several off-policy TD algorithms under the ETD(λ, β) framework, and a new error analysis that reveals the bias-variance trade-off between them.

Related Work: In recent years, several different off-policy policy-evaluation algorithms have been studied, such as importance-sampling based least-squares TD (Yu 2012), and gradient-based TD (Sutton et al. 2009; Liu et al. 2015). These algorithms are guaranteed to converge, however, their asymptotic error can be bounded only when the target and behavior policies are similar (Bertsekas and Yu 2009), or when their induced transition matrices satisfy a certain matrix-inequality suggested by Kolter (2011), which limits the discrepancy between the target and behavior policies. When these conditions are not satisfied, the error may be arbitrarily large (Kolter 2011). In contrast, the approximation-error bounds in this paper hold for *general target and behavior policies*.

2 Preliminaries

We consider an MDP $M = (S, A, P, R, \gamma)$, where S is the state space, A is the action space, P is the transition probability matrix, R is the reward function, and $\gamma \in [0, 1)$ is the discount factor.

Given a target policy π mapping states to a distribution over actions, our goal is to evaluate the *value function*:

$$V^{\pi}(s) \doteq \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} R(s_t, a_t)\middle| s_0 = s\right].$$

Linear temporal difference methods (Sutton and Barto 1998) approximate the value function by

$$V^{\pi}(s) \approx \theta^{\top}\varphi(s),$$

where $\varphi(s) \in \mathbb{R}^n$ are state features, and $\theta \in \mathbb{R}^n$ are weights, and use sampling to find a suitable θ. Let μ denote a behavior policy that generates the samples $s_0, a_0, s_1, a_1, \dots$ according to $a_t \sim \mu(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$. We denote by ρ_t the ratio $\pi(a_t|s_t)/\mu(a_t|s_t)$, and we assume, similarly to Sutton, Mahmood, and White (2015), that μ and π are such that ρ_t is well-defined[1] for all t.

Let T denote the Bellman operator for policy π, given by

$$T(V) \doteq R^{\pi} + \gamma P^{\pi} V,$$

where R^{π} and P^{π} are the reward vector and transition matrix induced by policy π, and let Φ denote a matrix whose columns are the feature vectors for all states. Let d_{μ} and d_{π} denote the stationary distributions over states induced by the policies μ and π, respectively. For some $d \in \mathbb{R}^{|S|}$ satisfying $d > 0$ element-wise, we denote by Π_d a projection to the subspace spanned by $\varphi(s)$ with respect to the d-weighted Euclidean-norm.

For $\lambda = 0$, the ETD(0, β) (Sutton, Mahmood, and White 2015) algorithm seeks to find a good approximation of the value function by iteratively updating the weight vector θ:

$$\begin{aligned} \theta_{t+1} &= \theta_t + \alpha F_t \rho_t (R_{t+1} + \gamma \theta_t^{\top} \varphi_{t+1} - \theta_t^{\top} \varphi_t)\varphi_t \\ F_t &= \beta \rho_{t-1} F_{t-1} + 1, \quad F_0 = 1, \end{aligned} \tag{1}$$

where F_t is a decaying trace of the importance-sampling ratios, and $\beta \in (0, 1)$ controls the decay rate.

Remark 1. *The algorithm of Sutton, Mahmood, and White (2015) selects the decay rate equal to the discount factor, i.e., $\beta = \gamma$. Here, we provide more freedom in choosing the decay rate. As our analysis reveals, the decay rate controls a bias-variance trade-off of ETD, therefore this freedom is important. Moreover, we note that for $\beta = 0$, we obtain the standard TD in an off-policy setting (Yu 2012), and when $\beta = 1$ we obtain the full importance-sampling TD algorithm (Precup, Sutton, and Dasgupta 2001).*

[1]Namely, if $\mu(a|s) = 0$ then $\pi(a|s) = 0$ for all $s \in S$.

Remark 2. *The ETD(0, γ) algorithm of Sutton, Mahmood, and White (2015) also includes a state-dependent emphasis weight $i(s)$, and a state-dependent discount factor $\gamma(s)$. Here, we analyze the case of a uniform weight $i(s) = 1$ and constant discount factor γ for all states. While our analysis can be extended to their more general setting, the insights from the analysis remain the same, and for the purpose of clarity we chose to focus on this simpler setting.*

An important term in our analysis is the emphatic weight vector f, defined by

$$f^{\top} = d_{\mu}^{\top}(I - \beta P)^{-1}. \tag{2}$$

It can be shown (Sutton, Mahmood, and White 2015; Yu 2015), that ETD(0, β) converges to θ^* - a solution of the following *projected fixed point equation*:

$$V = \Pi_f T V, \qquad V \in \mathbb{R}^{|S|}. \tag{3}$$

For the fixed point equation (3), a contraction property of $\Pi_f T$ is important for guaranteeing both a unique solution, and a bias bound (Bertsekas and Tsitsiklis 1996).

It is well known that T is a γ-contraction with respect to the d_{π}-weighted Euclidean norm (Bertsekas and Tsitsiklis 1996), and by definition Π_f is a non-expansion in f-norm, however, it is not immediate that the composed operator $\Pi_f T$ is a contraction in any norm. Indeed, for the TD(0) algorithm (Sutton and Barto, 1998; corresponding to the $\beta = 0$ case in our setting), a similar representation as a projected Bellman operator holds, but it may be shown that in the off-policy setting the algorithm might diverge (Baird 1995). In the next section, we study the contraction properties of $\Pi_f T$, and provide corresponding bias bounds.

3 Bias of ETD(0, β)

In this section we study the bias of the ETD(0, β) algorithm. Let us first introduce the following measure of discrepancy between the target and behavior policies:

$$\kappa \doteq \min_s \frac{d_{\mu}(s)}{f(s)}.$$

Lemma 1. *The measure κ obtains values ranging from $\kappa = 0$ (when there is a state visited by the target policy, but not the behavior policy), to $\kappa = 1 - \beta$ (when the two policies are identical).*

The technical proof is available in an extended version of the paper (Hallak et al. 2015). The following theorem shows that for ETD(0, β) with a suitable β, the projected Bellman operator $\Pi_f T$ is indeed a contraction.

Theorem 1. *For $\beta > \gamma^2(1-\kappa)$, the projected Bellman operator $\Pi_f T$ is a $\sqrt{\frac{\gamma^2}{\beta}(1-\kappa)}$-contraction with respect to the Euclidean f-weighted norm, namely, $\forall v_1, v_2 \in \mathbb{R}^{|S|}$:*

$$\|\Pi_f T v_1 - \Pi_f T v_2\|_f \leq \sqrt{\frac{\gamma^2}{\beta}(1-\kappa)}\|v_1 - v_2\|_f .$$

Proof. Let $F = diag(f)$. We have

$$\begin{aligned}
\|v\|_f^2 - \beta \|Pv\|_f^2 &= v^\top F v - \beta v^\top P^\top F P v \\
&\geq^{(a)} v^\top F v - \beta v^\top diag(f^\top P) v \\
&= v^\top [F - \beta diag(f^\top P)] v \\
&= v^\top \left[diag\left(f^\top (I - \beta P)\right)\right] v \\
&=^{(b)} v^\top diag(d_\mu) v = \|v\|_{d_\mu}^2 ,
\end{aligned}$$

where (a) follows from Jensen inequality:

$$\begin{aligned}
v^\top P^\top F P v &= \sum_s f(s)(\sum_{s'} P(s'|s) v(s'))^2 \\
&\leq \sum_s f(s) \sum_{s'} P(s'|s) v^2(s') \\
&= \sum_{s'} v^2(s') \sum_s f(s) P(s'|s) \\
&= v^\top diag(f^\top P) v,
\end{aligned}$$

and (b) is by the definition of f in (2).

Notice that for every v:

$$\|v\|_{d_\mu}^2 = \sum_s d_\mu(s) v^2(s) \geq \sum_s \kappa f(s) v^2(s) = \kappa \|v\|_f^2$$

Therefore:

$$\begin{aligned}
&\|v\|_f^2 \geq \beta \|Pv\|_f^2 + \|v\|_{d_\mu}^2 \geq \beta \|Pv\|_f^2 + \kappa \|v\|_f^2 , \\
&\Rightarrow \quad \beta \|Pv\|_f^2 \leq (1-\kappa) \|v\|_f^2
\end{aligned}$$

and:

$$\begin{aligned}
\|Tv_1 - Tv_2\|_f^2 &= \|\gamma P(v_1 - v_2)\|_f^2 \\
&= \gamma^2 \|P(v_1 - v_2)\|_f^2 \\
&\leq \frac{\gamma^2}{\beta}(1-\kappa) \|v_1 - v_2\|_f^2 .
\end{aligned}$$

Hence, T is a $\sqrt{\frac{\gamma^2}{\beta}(1-\kappa)}$-contraction. Since Π_f is a non-expansion in the f-weighted norm (Bertsekas and Tsitsiklis 1996), $\Pi_f T$ is a $\sqrt{\frac{\gamma^2}{\beta}(1-\kappa)}$-contraction as well. □

Recall that for the original ETD algorithm (Sutton, Mahmood, and White 2015), we have that $\beta = \gamma$, and the contraction modulus is $\sqrt{\gamma(1-\kappa)} < 1$, thus the contraction of $\Pi_f T$ always holds.

Also note that in the on-policy case, the behavior and target policies are equal, and according to Lemma 1 we have $1 - \kappa = \beta$. In this case, the contraction modulus in Theorem 1 is γ, similar to the result for on-policy TD (Bertsekas and Tsitsiklis 1996).

We remark that Kolter (2011) also used a measure of discrepancy between the behavior and the target policy to bound the TD-error. However, Kolter (2011) considered the standard TD algorithm, for which a contraction could be guaranteed only for a class of behavior policies that satisfy a certain matrix inequality criterion. Our results show that for ETD(0, β) with a suitable β, a contraction is guaranteed for *general* behavior policies. We now show in an example that our contraction modulus bounds are tight.

Example 1. *Consider an MDP with two states: Left and Right. In each state there are two identical actions leading to either Left or Right deterministically. The behavior policy will choose Right with probability ε, and the target policy will choose Left with probability ε, hence $1 - \kappa \approx 1$. Calculating the quantities of interest:*

$$P = \begin{pmatrix} \varepsilon & 1-\varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}, \quad d_\mu = (1-\varepsilon, \varepsilon)$$

$$f = \frac{1}{1-\beta}\left(1 + 2\varepsilon\beta - \varepsilon - \beta, -2\varepsilon\beta + \varepsilon + \beta\right)^\top .$$

So for $v = (0,1)^\top$:

$$\|v\|_f^2 = \frac{\varepsilon + \beta - 2\varepsilon\beta}{1-\beta}, \quad \|Pv\|_f^2 = \frac{(1-\varepsilon)^2}{1-\beta},$$

and for small ε we obtain that $\frac{\|\gamma P v\|^2}{\|v\|_f^2} \approx \frac{\gamma^2}{\beta}$.

An immediate consequence of Theorem 1 is the following error bound based on Lemma 6.9 of Bertsekas and Tsitsiklis (1996):

Corollary 1. *We have*

$$\left\|\Phi^\top \theta^* - V^\pi\right\|_f \leq \frac{1}{\sqrt{1 - \frac{\gamma^2}{\beta}(1-\kappa)}} \|\Pi_f V^\pi - V^\pi\|_f ,$$

$$\left\|\Phi^\top \theta^* - V^\pi\right\|_{d_\mu} \leq \frac{1}{\sqrt{\gamma\left(1 - \frac{\gamma^2}{\beta}(1-\kappa)\right)}} \|\Pi_f V^\pi - V^\pi\|_f .$$

Up to the weights in the norm, the error $\|\Pi_f V^\pi - V^\pi\|_f$ is the best approximation we can hope for, within the capability of the linear approximation architecture. Corollary 1 guarantees that we are not too far away from it.

Notice that the error $\left\|\Phi^\top \theta^* - V^\pi\right\|_{d_\mu}$ uses a measure d_μ which is independent of the target policy; This could be useful in further analysis of a policy iteration algorithm, which iteratively improves the target policy using samples from a single behavior policy. Such an analysis may proceed similarly to that in Munos (2003) for the on-policy case.

3.1 Numerical Illustration

We illustrate the importance of the ETD(0, β) bias bound in a numerical example. Consider the 2-state MDP example of Kolter (2011), with transition matrix $P = (1/2)\underline{\underline{\mathbf{1}}}$ (where $\underline{\underline{\mathbf{1}}}$ is an all 1 matrix), discount factor $\gamma = 0.99$, and value function $V = [1, 1.05]^\top$ (with $R = (I - \gamma P)V$). The features are

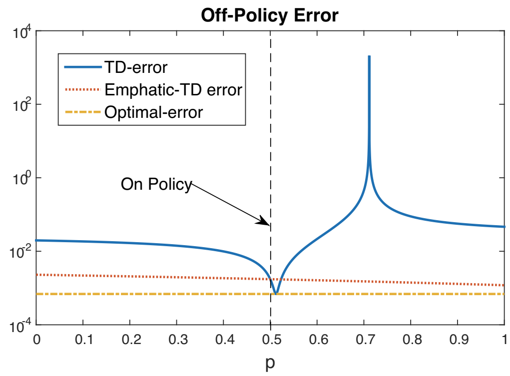

Figure 1: Mean squared error in value function approximation for different behavior policies.

$\Phi = [1, 1.05+\varepsilon]^\top$, with $\varepsilon = 0.001$. Clearly, in this example we have $d_\pi = [0.5, 0.5]$. The behavior policy is chosen such that $d_\mu = [p, 1-p]$.

In Figure 1 we plot the mean-squared error $\left\|\Phi^\top\theta^* - V^\pi\right\|_{d_\pi}$, where θ^* is either the fixed point of the standard TD equation $V = \Pi_{d_\mu}TV$, or the ETD(0, β) fixed point of (3), with $\beta = \gamma$. We also show the optimal error $\|\Pi_{d_\pi}V - V^\pi\|_{d_\pi}$ achievable with these features. Note that, as observed by Kolter (2011), for certain behavior policies the bias of standard TD is infinite. This means that algorithms that converge to this fixed point, such as the GTD algorithm (Sutton et al. 2009), are hopeless in such cases. The ETD algorithm, on the other hand, has a bounded bias *for all behavior policies*.

4 The Bias-Variance Trade-Off of ETD(0, β)

From the results in Corollary 1, it is clear that increasing the decay rate β decreases the bias bound. Indeed, for the case $\beta = 1$ we obtain the importance sampling TD algorithm (Precup, Sutton, and Dasgupta 2001), which is known to have a bias bound similar to on-policy TD. However, as recognized by Precup, Sutton, and Dasgupta (2001) and Sutton, Mahmood, and White (2015), the importance sampling ratio F_t suffers from a high variance, which increases with β. The quantity F_t is important as it appears as a multiplicative factor in the definition of the ETD learning rule, so its amplitude directly impacts the stability of the algorithm. In fact, the asymptotic variance of F_t may be infinite, as we show in the following example:

Example 2. *Consider the same MDP given in Example 1, only now the behavior policy chooses Left or Right with probability* 0.5*, and the target policy chooses always Right. For ETD(0, β) with $\beta \in [0, 1)$, we have that when $S_t = Left$ then $F_t = 1$ (since $\rho_{t-1} = 0$). When $S_t = Right$, F_t may take several values depending on how many steps, $\tau(t)$, was the last transition from Left to Right, i.e. $\tau(t) \overset{\text{def}}{=} \min\{i \geq 0 : S_{t-i} = Left\}$. We can write this value as $F^{\tau(t)}$ where:*

$$F^\tau \doteq \sum_{i=0}^{\tau}(2\beta)^i = \frac{(2\beta)^{\tau+1} - 1}{2\beta - 1},$$

if $2\beta \neq 1$. Let us assume that $2\beta > 1$ since interesting cases happen when β is close to 1.

Let's compute F_t's average over time: Following the stationary distribution of the behavior policy, $S_t = Left$ with probability $1/2$. Now, conditioned on $S_t = Right$ (which happens with probability $1/2$), we have $\tau(t) = i$ with probability 2^{-i-1}. Thus the average (over time) value of F_t is

$$\mathbb{E}F_t = \frac{1}{2}\sum_{i=0}^{\infty}2^{-i-1}F^i = \frac{\sum_i \beta^{i+1} - 1}{2(2\beta - 1)} = \frac{1}{2(1-\beta)}.$$

Thus F_t amplifies the TD update by a factor of $\frac{1}{2(1-\beta)}$ in average. Unfortunately, the actual values of the (random variable) F_t does not concentrate around its expectation, and actually F_t does not even have a finite variance. Indeed the average (over time) of F_t^2 is

$$\mathbb{E}F_t^2 = \frac{1}{4}\sum_{i=0}^{\infty}2^{-i}(F^i)^2 = \frac{\sum_i 2^{-i}\left((2\beta)^{i+1} - 1\right)^2}{4(2\beta - 1)^2} = \infty,$$

as soon as $2\beta^2 \geq 1$.

So although ETD(0, β) converges almost surely (as shown by Yu, 2015), the variance of the estimate may be infinite, which suggests a prohibitively slow convergence rate.

In the following proposition we characterize the dependence of the variance of F_t on β.

Proposition 1. *Define the mismatch matrix $\tilde{P}_{\mu,\pi}$ such that $[\tilde{P}_{\mu,\pi}]_{\bar{s}s} = \sum_a p(s|\bar{s}, \bar{a})\frac{\pi^2(a|\bar{s})}{\mu(a|\bar{s})}$ and write $\alpha(\mu, \pi)$ the largest magnitude of its eigenvalues. Then for any $\beta < 1/\sqrt{\alpha(\mu,\pi)}$ the average variance of F_t (conditioned on any state) is finite, and*

$$\mathbb{E}_\mu\left[Var[F_t|S_t = s]\right] \leq \frac{\beta^2}{1-\beta}\left(2 + \frac{(1+\beta)\left\|\tilde{P}_{\mu,\pi}\right\|_\infty}{1 - \beta^2\left\|\tilde{P}_{\mu,\pi}\right\|_\infty}\right),$$

where $\left\|\tilde{P}_{\mu,\pi}\right\|_\infty$ is the l_∞-induced norm which is the maximum absolute row sum of the matrix.

Proof. (Partial) Following the same derivation that Sutton, Mahmood, and White (2015) used to prove that $f(s) = d_\mu(s)\lim_{t\to\infty}\mathbb{E}[F_t|S_t = s]$, we have

$$\begin{aligned} q(s) &\doteq d_\mu(s)\lim_{t\to\infty}\mathbb{E}[F_t^2|S_t = s] \\ &= d_\mu(s)\lim_{t\to\infty}\mathbb{E}[(1 + \rho_{t-1}\beta F_{t-1})^2|S_t = s] \\ &= d_\mu(s)\lim_{t\to\infty}\mathbb{E}[1 + 2\rho_{t-1}\beta F_{t-1} + \rho_{t-1}^2\beta^2F_{t-1}^2|S_t = s]. \end{aligned}$$

For the first summand, we get $d_\mu(s)$. For the second summand, we get:

$$2\beta d_\mu(s)\lim_{t\to\infty}\mathbb{E}[\rho_{t-1}F_{t-1}|S_t = s] = 2\beta\sum_{\bar{s}}[P_\pi]_{\bar{s}s}f(\bar{s}).$$

The third summand equals

$$\beta^2 \sum_{\bar{s},\bar{a}} d_\mu(\bar{s})\mu(\bar{a}|\bar{s})p(s|\bar{s},\bar{a})\frac{\pi^2(\bar{a}|\bar{s})}{\mu^2(\bar{a}|\bar{s})} \lim_{t\to\infty} \mathbb{E}[F_{t-1}^2|S_{t-1}=\bar{s}]$$
$$= \beta^2 \sum_{\bar{s},\bar{a}} p(s|\bar{s},\bar{a})\frac{\pi^2(\bar{a}|\bar{s})}{\mu(\bar{a}|\bar{s})} q(\bar{s}) = \beta^2 \sum_{\bar{s}} [\tilde{P}_{\mu,\pi}]_{\bar{s}s} q(\bar{s}).$$

Hence $q = d_\mu + 2\beta P_\pi^\top f + \beta^2 \tilde{P}_{\mu,\pi}^\top q$. Thus for any $\beta < 1/\sqrt{\alpha(\mu,\pi)}$, all eigenvalues of the matrix $\beta^2 \tilde{P}_{\mu,\pi}^\top$ have magnitude smaller than 1, and the vector q has finite components. The rest of the proof is very technical and is available in an extended version of the paper (Hallak et al. 2015).

□

Proposition 1 and Corollary 1 show that the decay rate β *acts as an implicit trade-off parameter* between the bias and variance in ETD. For large β, we have a low bias but suffer from a high variance (possibly infinite if $\beta \geq 1/\sqrt{\alpha(\mu,\pi)}$), and vice versa for small β. Notice that for the on-policy case, $\alpha(\mu,\pi) = 1$ thus for any $\beta < 1$ the variance is finite.

Originally, ETD(0, β) was introduced with $\beta = \gamma$, and from our perspective, it may be seen as a specific choice for the bias-variance trade-off. However, there is no intrinsic reason to choose $\beta = \gamma$, and other choices may be preferred in practice, depending on the nature of the problem. In the following numerical example, we investigate the bias-variance dependence on β, and show that the optimal β in term of mean-squared error may be quite different from γ.

4.1 Numerical Illustration

We revisit the 2-state MDP described in Section 3.1, with $\gamma = 0.9$, $\varepsilon = 0.2$ and $p = 0.95$. For these parameter settings, the error of standard TD is 42.55 (p was chosen to be close to a point of infinite bias for these parameters).

In Figure 2 we plot the mean-squared error $\left\|\Phi^\top\theta^* - V^\pi\right\|_{d_\pi}$, where θ^* was obtained by running ETD(0, β) with a step size $\alpha = 0.001$ for $10,000$ iterations, and averaging the results over $10,000$ different runs.

First of all, note that for all β, the error is smaller by two orders of magnitude than that of standard TD. Thus, algorithms that converge to the standard TD fixed point such as GTD (Sutton et al. 2009) are significantly outperformed by ETD(0, β) in this case. Second, note the dependence of the error on β, demonstrating the bias-variance trade-off discussed above. Finally, note that the minimal error is obtained for $\gamma = 0.8$, and is considerably smaller than that of the original ETD with $\beta = \gamma = 0.9$.

5 Contraction Property for ETD(λ, β)

We now extend our results to incorporate eligibility traces, in the style of the ETD(λ) algorithm (Sutton, Mahmood, and White 2015), and show similar contraction properties and error bounds.

The ETD(λ, β) algorithm iteratively updates the weight

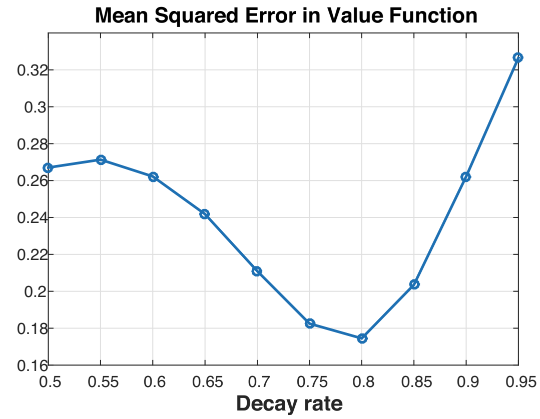

Figure 2: Mean squared error in value function approximation for different decay rates β.

vector θ according to

$$\begin{aligned}
\theta_{t+1} &:= \theta_t + \alpha(R_{t+1} + \gamma\theta_t^\top\varphi_{t+1} - \theta_t^\top\varphi_t)e_t \\
e_t &= \rho_t(\gamma\lambda e_{t-1} + M_t\varphi_t), \quad e_{-1} = 0 \\
M_t &= \lambda + (1-\lambda)F_t \\
F_t &= \beta\rho_{t-1}F_{t-1} + 1, \quad F_0 = 1,
\end{aligned}$$

where e_t is the eligibility trace (Sutton, Mahmood, and White 2015). In this case, we define the emphatic weight vector m by

$$m^\top = d_\mu^\top (I - P^{\lambda,\beta})^{-1}, \tag{4}$$

where $P^{a,b}$ for some $a, b \in \mathbb{R}$ denotes the following matrix:

$$P^{a,b} = I - (I - baP)^{-1}(I - bP).$$

The Bellman operator for general λ and γ is given by:

$$T^{(\lambda)}(V) = (I - \gamma\lambda P)^{-1}R + P^{\lambda,\gamma}V, \quad V \in \mathbb{R}^{|S|}.$$

For $\lambda = 0$ we have $P_\pi^{\lambda,\beta} = \beta P$, $P_\pi^{\lambda,\gamma} = \gamma P$, and $m = f$ so we recover the definitions of ETD(0, β).

Recall that our goal is to estimate the value function V^π. Thus, we would like to know how well the ETD(λ, β) solution approximates V^π. Mahmood et al. (2015) show that, under suitable step-size conditions, ETD converges to some θ^* that is a solution of the *projected fixed-point equation*:

$$\theta^\top\Phi = \Pi_m T^{(\lambda)}(\theta^\top\Phi).$$

In their analysis, however, Mahmood et al. (2015) did not show how well the solution $\Phi^\top\theta^*$ approximates V^π. Next, we establish that the projected Bellman operator $\Pi_m T^{(\lambda)}$ is a contraction. This result will then allow us to bound the error $\left\|\Phi^\top\theta^* - V^\pi\right\|_m$.

Theorem 2. *$\Pi_m T^{(\lambda)}$ is an ω-contraction with respect to the Euclidean m-weighted norm where:*

$$\begin{aligned}
\beta \geq \gamma: \quad \omega &= \sqrt{\frac{\gamma^2(1+\lambda\beta)^2(1-\lambda)}{\beta(1+\gamma\lambda)^2(1-\lambda\beta)}}, \\
\beta \leq \gamma: \quad \omega &= \sqrt{\frac{\gamma^2(1-\beta\lambda)(1-\lambda)}{\beta(1-\gamma\lambda)^2}}.
\end{aligned} \tag{5}$$

Proof. (sketch) The proof is almost identical to the proof of Theorem 1, only now we cannot apply Jensen's inequality directly, since the rows of $P^{\lambda,\beta}$ do not sum to 1. However:

$$P^{\lambda,\beta}\mathbf{1} = \left(I - (I - \beta\lambda P)^{-1}(I - \beta P)\right)\mathbf{1} = \zeta\mathbf{1},$$

where $\zeta = \frac{\beta(1-\lambda)}{1-\lambda\beta}$. Notice that each entry of $P^{\lambda,\beta}$ is positive. Therefore $\frac{P^{\lambda,\beta}}{\zeta}$ will hold for Jensen's inequality. Let $M = diag(m)$, we have

$$\begin{aligned}
\|v\|_m^2 - \frac{1}{\zeta}\left\|P^{\lambda,\beta}v\right\|_m^2 &= v^\top Mv - \zeta v^\top \frac{P^{\lambda,\beta}}{\zeta}^\top M \frac{P^{\lambda,\beta}}{\zeta}v \\
&\geq^{(a)} v^\top Mv - \beta v^\top diag(m^\top \frac{P^{\lambda,\beta}}{\zeta})v \\
&= v^\top[M - diag(m^\top P^{\lambda,\beta})]v \\
&= v^\top \left[diag\left(m^\top(I - P^{\lambda,\beta})\right)\right]v \\
&=^{(b)} v^\top diag(d_\mu)v = \|v\|_{d_\mu}^2,
\end{aligned}$$

where (a) follows from the Jensen inequality and (b) from Equation (4). Therefore:

$$\|v\|_m^2 \geq \frac{1}{\zeta}\left\|P^{\lambda,\beta}v\right\|_m^2 + \|v\|_{d_\mu}^2 \geq \frac{1}{\zeta}\left\|P^{\lambda,\beta}v\right\|_m^2,$$

and:

$$\begin{aligned}
&\left\|T^{(\lambda)}v_1 - T^{(\lambda)}v_2\right\|_m^2 = \left\|P^{\lambda,\gamma}(v_1 - v_2)\right\|_m^2 \\
&\text{(Case A:} \quad \beta \geq \gamma) \leq \left\|\frac{\gamma(1+\beta\lambda)}{\beta(1+\gamma\lambda)}P^{\lambda,\beta}(v_1 - v_2)\right\|_m^2 \\
&\qquad \leq \frac{\gamma^2(1+\lambda\beta)^2(1-\lambda)}{\beta(1+\gamma\lambda)^2(1-\lambda\beta)}\|v_1 - v_2\|_m^2, \\
&\text{(Case B:} \quad \beta \leq \gamma) \leq \left\|\frac{\gamma(1-\beta\lambda)}{\beta(1-\gamma\lambda)}P^{\lambda,\beta}(v_1 - v_2)\right\|_m^2 \\
&\qquad \leq \frac{\gamma^2(1-\beta\lambda)(1-\lambda)}{\beta(1-\gamma\lambda)^2}\|v_1 - v_2\|_m^2.
\end{aligned}$$

The inequalities depending on the two cases originate from the fact that the two matrices $P^{\lambda,\beta}, P^{\lambda,\gamma}$ are polynomials of the same matrix P_π, and mathematical manipulation on the corresponding eigenvalues decomposition of $(v_1 - v_2)$. The details are given in an extended version of the paper (Hallak et al. 2015).

Now, for a proper choice of β, the operator $T^{(\lambda)}$ is a contraction, and since Π_m is a non-expansion in the m-weighted norm, $\Pi_m T^{(\lambda)}$ is a contraction as well. □

In Figure 3 we illustrate the dependence of the contraction moduli bound on λ and β. In particular, for $\lambda \to 1$, the contraction modulus diminishes to 0. Thus, for large enough λ, a contraction can always be guaranteed (this can also be shown mathematically from the contraction results of Theorem 2). We remark that a similar result for standard TD(λ) was established by Yu, 2012. However, as is well-known (Bertsekas 2012), increasing λ also increases the variance of the algorithm, and we therefore obtain a bias-variance trade-off in λ as well as β. Finally, note that for $\beta = \gamma$, the contraction modulus equals $\sqrt{\frac{\gamma(1-\lambda)}{1-\gamma\lambda}}$, and that for $\lambda = 0$ the result is the same as in Theorem 1.

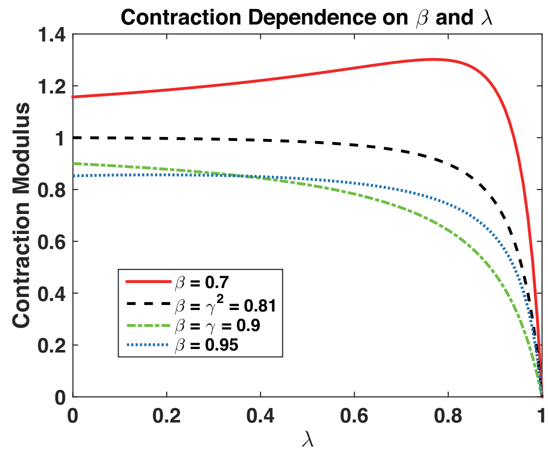

Figure 3: Contraction moduli of $\Pi_m T^{(\lambda)}$ for different β's, as a function of the bootstrapping parameter λ. Notice that we see a steep decrease in the moduli only for λ close to 1.

6 Conclusion

In this work we unified several off-policy TD algorithms under the ETD(λ, β) framework, which flexibly manages the bias and variance of the algorithm by controlling the decay-rate of the importance-sampling ratio. From this perspective, we showed that several different methods proposed in the literature are special instances of this bias-variance selection.

Our main contribution is an error analysis of ETD(λ, β) that quantifies the bias-variance trade-off. In particular, we showed that the recently proposed ETD algorithm of Sutton, Mahmood, and White (2015) has bounded bias for *general* behavior and target policies, and that by controlling the decay-rate in the ETD(λ, β) algorithm, an improved performance may be obtained by reducing the variance of the algorithm while still maintaining a reasonable bias.

Possible future extensions of our work includes finite-time bounds for off-policy ETD(λ, β), an error propagation analysis of off-policy *policy improvement*, and solving the bias-variance trade-off adaptively from data.

7 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. Aviv Tamar is also partially funded by the Viterbi Scholarship, Technion.

References

Baird, L. 1995. Residual algorithms: Reinforcement learning with function approximation. In *ICML*.

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

Bertsekas, D., and Yu, H. 2009. Projected equation methods for approximate solution of large linear systems. *Journal of Computational and Applied Mathematics* 227(1):27–50.

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

Hallak, A.; Tamar, A.; Munos, R.; and Mannor, S. 2015. Generalized emphatic temporal difference learning: Bias-variance analysis. *arXiv preprint arXiv:1509.05172*.

Kolter, J. Z. 2011. The fixed points of off-policy TD. In *NIPS*.

Liu, B.; Liu, J.; Ghavamzadeh, M.; Mahadevan, S.; and Petrik, M. 2015. Finite-sample analysis of proximal gradient TD algorithms. In *UAI*.

Mahmood, A. R.; Yu, H.; White, M.; and Sutton, R. S. 2015. Emphatic temporal-difference learning. *arXiv:1507.01569*.

Munos, R. 2003. Error bounds for approximate policy iteration. In *ICML*.

Precup, D.; Sutton, R. S.; and Dasgupta, S. 2001. Off-policy temporal-difference learning with function approximation. In *ICML*.

Sutton, R. S., and Barto, A. 1998. *Reinforcement learning: An introduction*. Cambridge Univ Press.

Sutton, R. S.; Maei, H. R.; Precup, D.; Bhatnagar, S.; Silver, D.; Szepesvári, C.; and Wiewiora, E. 2009. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *ICML*.

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

Yu, H. 2012. Least squares temporal difference methods: An analysis under general conditions. *SIAM Journal on Control and Optimization* 50(6):3310–3343.

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