Journal of Machine Learning Research 17 (2016) 1-36 Submitted 8/14; Revised 12/15; Published 3/16

Learning the Variance of the Reward-To-Go

Aviv Tamar AVIVT@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94709, USA

Dotan Di Castro DOT@YAHOO-INC.COM
Yahoo! Research Labs
MATAM, Haifa 31905, Israel

Shie Mannor SHIE@EE.TECHNION.AC.IL
Department of Electrical Engineering
The Technion - Israel Institute of Technology
Haifa 32000, Israel

Editor: Peter Auer

Abstract

In Markov decision processes (MDPs), the variance of the reward-to-go is a natural measure of uncertainty about the long term performance of a policy, and is important in domains such as finance, resource allocation, and process control. Currently however, there is no tractable procedure for calculating it in large scale MDPs. This is in contrast to the case of the expected reward-to-go, also known as the value function, for which effective simulation-based algorithms are known, and have been used successfully in various domains. In this paper[1] we extend temporal difference (TD) learning algorithms to estimating the variance of the reward-to-go for a fixed policy. We propose variants of both TD(0) and LSTD(λ) with linear function approximation, prove their convergence, and demonstrate their utility in an option pricing problem. Our results show a dramatic improvement in terms of sample efficiency over standard Monte-Carlo methods, which are currently the state-of-the-art.

Keywords: Reinforcement learning, Markov decision processes, variance estimation, simulation, temporal differences

1. Introduction

In sequential decision making within the Markov Decision Process (MDP) framework, whether in a planning setting (Puterman, 1994; Powell, 2011) or a reinforcement learning setting (RL; Bertsekas and Tsitsiklis, 1996, Sutton and Barto, 1998), the decision maker ultimately obtains a policy π, often with some guarantees on its expected long-term performance. This typical conclusion of the policy optimization process is the starting point of our work.

We consider the policy π to be fixed[2], and we are interested in understanding how π performs in practice, with the natural quantity of interest being the reward-to-go from each state of the system. The *expected* reward-to-go J, also known as the *value function*, is often

[1] This paper extends an earlier work by the authors (Tamar et al., 2013).

[2] This setting is also known as a Markov reward process.

a part of an optimization process, and efficient methods for learning it are well known. In many applications, however, looking at expectations is not enough, and it seems only reasonable to estimate other statistics of the reward-to-go, such as its *variance*, denoted by V. Quite surprisingly, this topic has received very little attention; at the current state-of-the-art, the only solution for large-scale MDPs is a naive Monte-Carlo approach, demanding extensive simulations of the long-term outcomes *from each system state*. In this paper we explore much more efficient alternatives.

We further motivate policy evaluation with respect to the variance of the reward-to-go. The variance is an intuitive measure of uncertainty, and common practice in many domains such as finance, process control, and clinical decision making (Sharpe, 1966; Shortreed et al., 2011). As we show in the paper, in an option pricing domain, the uncertainty captured by the variance of the reward-to-go highlights important properties of the policy, that are not visible by looking at the value function alone.

The variance may also be used for policy selection. In some practical situations, a full policy-optimization procedure is not an option, and the agent can only select between several predefined policies. For example, it may be that each policy is designed by an expert (e.g., a private equity or investment fund), and the agent can simply select between several policies, given the current features of the system (e.g., current economic indicators). A related financial experiment in a non-sequential setting was reported by Moody and Saffell (2001), which selected between several investment types, and which emphasized the importance of incorporating variance-based objectives in the policy selection.

Finally, the value function has proved to be a fundamental ingredient in many policy optimization algorithms. The variance of the reward-to-go may thus prove valuable for *risk aware* optimization algorithms, a topic that has gained significant interest recently (Filar et al., 1995; Mihatsch and Neuneier, 2002; Geibel and Wysotzki, 2005; Mannor and Tsitsiklis, 2013). Therefore, the policy evaluation methods in this work may be used as a sub-procedure in policy optimization. Since the conference publication of this work, this idea has already been explored by Tamar and Mannor (2013) and Prashanth and Ghavamzadeh (2013). Both Tamar and Mannor (2013) and Prashanth and Ghavamzadeh (2013) suggested actor-critic algorithms, in which the critic uses the policy evaluation ideas introduced in this paper. We are certain that risk-aware policy evaluation would play a major role in future risk-aware optimization algorithms as well.

The principal challenge in policy evaluation arises when the state space is large, or continuous. Then, solving Bellman's equation for the value or its extension (Sobel, 1982) for the variance becomes intractable. This difficulty is even more pronounced in the learning setting, when a model of the process is not available, and the evaluation has to be *estimated* from a limited amount of samples. Fortunately, for the case of the value function, effective learning approaches are known.

Temporal Difference methods (TD; Sutton, 1988) typically employ *function approximation* to represent the value function in a lower dimensional subspace, and *learn* the approximation parameters efficiently, by fitting the spatiotemporal relations in Bellman's equation to the observed (or simulated) data. TD methods have been studied extensively, both theoretically (Bertsekas, 2012, Lazaric et al., 2010) and empirically (e.g., Tesauro, 1995, Powell, 2011, Section 14.5), and are considered to be the state-of-the-art in policy evaluation.

However, when it comes to evaluating additional statistics of the reward-to-go, such as its variance, little is known. This may be due to the fact that the linearity of the expectation in Bellman's equation plays a key role in TD algorithms.

In this paper we present a TD framework for learning the variance of the reward-to-go, using function approximation, in problems where a model is not available, or too large to solve. To our knowledge, this is the first work that addresses the challenge of large state spaces, by considering an approximation scheme for the variance. Our approach is based on the following observation: the second moment of the reward-to-go, denoted by M, together with the value function J, satisfies a linear 'Bellman-like' equation. By extending TD methods to jointly estimate J and M with linear function approximation, we obtain a solution for estimating the variance, using the relation $V = M - J^2$.

We propose both a variant of Least Squares Temporal Difference (LSTD; Boyan 2002) and of TD(0) (Sutton and Barto, 1998) for jointly estimating J and M with linear function approximation. For these algorithms, we provide convergence guarantees and error bounds. In addition, we introduce novel methods for enforcing the approximate variance to be positive, through a constrained TD equation or through an appropriate choice of features. An empirical evaluation of our approach on an American-style option pricing problem demonstrates a dramatic improvement in terms of sample efficiency compared to Monte Carlo techniques—the current state of the art.

A previous study by Sato et al. (2001) suggested TD equations for J and V, without function approximation. Their approach relied on a non-linear equation for V, and it is not clear how it may be extended to handle large state spaces. More recently, Morimura et al. (2010) proposed TD learning rules for a parametric distribution of the return, albeit without function approximation nor formal guarantees. In the Bayesian Gaussian process temporal difference framework of Engel et al. (2005), the reward-to-go is assumed to have a Gaussian posterior distribution, and its mean and variance are estimated. However, the resulting variance is a product of both stochastic transitions and model uncertainty, and is thus different than the variance considered here. For average reward MDPs, several studies (e.g., Filar et al., 1989) considered the variation of the reward from its average. This measure of variability is not suitable for the discounted and episodic settings considered here. A different line of work considers MDPs with a *dynamic-risk* measure (Ruszczyński, 2010). In dynamic risk, instead of considering the reward-to-go as the random variable of interest, the risk is defined iteratively over the possible future trajectories. In an optimization setting, dynamic-risk has some favorable properties such as time-consistency, and a dynamic programming formulation (Ruszczyński, 2010). However, the variance of the reward-to-go considered here is considerably more intuitive, and leads to a much simpler approach.

This paper is organized as follows. In Section 2 we present our formal MDP setup. In Section 3 we derive the fundamental equations for jointly approximating J and M, and discuss their properties. A solution to these equations may be obtained by sampling, through the use of TD algorithms, as presented in Section 4. As it turns out, our approximation scheme may result in cases where the approximate variance is negative. We discuss this in Section 5, and propose methods for avoiding it. Section 6 presents an empirical evaluation on an option pricing problem, and Section 7 concludes, and discusses future directions.

2. Framework and Background

We consider an episodic MDP[3] (also known as a stochastic shortest path problem; Bertsekas 2012) in discrete time with a finite state space $X \triangleq \{1, \ldots, n\}$ and a terminal state x^*. A *fixed* policy π determines, for each $x \in X$, a stochastic transition to a subsequent state $x' \in \{X \cup x^*\}$ with probability $P(x'|x)$. We consider a deterministic and bounded reward function $r : X \to \mathbb{R}$, and assume zero reward at the terminal state. We denote by x_k the state at time k, where $k = 0, 1, 2, \ldots$.

A policy is said to be *proper* (Bertsekas, 2012) if there is a positive probability that the terminal state x^* will be reached after at most n transitions, from any initial state. Throughout this paper we make the following assumption:

Assumption 1 *The policy π is proper.*

Let $\gamma \in (0, 1]$ denote a discount factor. We emphasize that the case $\gamma = 1$, corresponding to a non-discounted setting, is allowed, and much of our effort in the sequel is to handle this special and important case. Let $\tau \triangleq \min\{k > 0 | x_k = x^*\}$ denote the first visit time to the terminal state, and let the random variable B denote the accumulated (and possibly discounted) reward along the trajectory until that time

$$B \triangleq \sum_{k=0}^{\tau-1} \gamma^k r(x_k).$$

In this work, we are interested in the mean-variance tradeoff in B, represented by the *value function*

$$J(x) \triangleq \mathbb{E}\left[B | x_0 = x\right], \quad x \in X,$$

and the *variance of the reward-to-go*

$$V(x) \triangleq \mathrm{Var}\left[B | x_0 = x\right], \quad x \in X.$$

We will find it convenient to define also the *second moment of the reward-to-go*

$$M(x) \triangleq \mathbb{E}\left[B^2 | x_0 = x\right], \quad x \in X.$$

Our goal is to estimate the functions $J(x)$ and $V(x)$ from trajectories obtained by simulating the MDP with policy π.

3. Approximation of the Variance of the Reward-To-Go

In this section we derive a projected equation method for approximating $J(x)$ and $M(x)$ using linear function approximation. The approximation of $V(x)$ will then follow from the relation $V(x) = M(x) - J(x)^2$.

Our starting point is a system of equations for $J(x)$ and $M(x)$, first derived by Sobel (1982) for a discounted infinite horizon case, and extended here to the episodic case. The equation for J is the well known Bellman equation for a fixed policy, and independent of the equation for M.

[3] In particular, any finite horizon MDP is an episodic MDP, for which our results apply. Extending these results to the infinite horizon discounted setting is straightforward.

Proposition 2 *The following equations hold for $x \in X$*

$$
\begin{aligned}
J(x) &= r(x) + \gamma \sum_{x' \in X} P(x'|x) J(x'), \\
M(x) &= r(x)^2 + 2\gamma r(x) \sum_{x' \in X} P(x'|x) J(x') + \gamma^2 \sum_{x' \in X} P(x'|x) M(x').
\end{aligned}
\tag{1}
$$

Furthermore, under Assumption 1 a unique solution to Eq. (1) *exists.*

A straightforward proof is given in Appendix A.

At this point the reader may wonder why an equation for V is not presented. While such an equation may be derived (see, e.g., Sobel 1982, Tamar et al. 2012), it is not linear. The linearity of (1) in J and M is the key to our approach. As we show in the next subsection, the solution to (1) may be expressed as the fixed point of a linear mapping in the joint space of J and M. We will then show that a projection of this mapping onto a linear feature space is contracting, thus allowing us to use the TD methodology to estimate J and M.

3.1 A Projected Fixed Point Equation in the Joint Space of J and M

For the sequel, we introduce the following vector notations. We denote by $P \in \mathbb{R}^{n \times n}$ and $r \in \mathbb{R}^n$ the episodic MDP transition matrix and reward vector, i.e., $P_{x,x'} = P(x'|x)$ and $r_x = r(x)$, where $x, x' \in X$. Also, we define the diagonal matrix $R \triangleq diag(r)$.

For a vector $z \in \mathbb{R}^{2n}$ we let $z_J \in \mathbb{R}^n$ and $z_M \in \mathbb{R}^n$ denote its leading and ending n components, respectively. Thus, such a vector belongs to the joint space of J and M.

We define the mapping $T : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ by

$$
\begin{aligned}
[Tz]_J &= r + \gamma P z_J, \\
[Tz]_M &= Rr + 2\gamma R P z_J + \gamma^2 P z_M.
\end{aligned}
\tag{2}
$$

It may easily be verified that a fixed point of T is a solution to (1), and by Proposition 2 such a fixed point exists and is unique.

When the state space X is large, however, a direct solution of (1) is not feasible. A popular approach in this case is to approximate $J(x)$ by restricting it to a lower dimensional subspace, and use simulation based TD algorithms to *learn* the approximation parameters (Bertsekas, 2012). In this paper we extend this approach to the approximation of $M(x)$ as well.

We consider a linear approximation architecture of the form

$$
\tilde{J}(x) = \phi_J(x)^\top w_J, \quad \tilde{M}(x) = \phi_M(x)^\top w_M,
$$

where $w_J \in \mathbb{R}^l$ and $w_M \in \mathbb{R}^m$ are the approximation parameter vectors, $\phi_J(x) \in \mathbb{R}^l$ and $\phi_M(x) \in \mathbb{R}^m$ are state dependent features, and $(\cdot)^\top$ denotes the transpose of a vector. The low dimensional subspaces are therefore

$$
S_J = \{\Phi_J w | w \in \mathbb{R}^l\}, \quad S_M = \{\Phi_M w | w \in \mathbb{R}^m\},
$$

where Φ_J and Φ_M are matrices whose rows are $\phi_J(x)^\top$ and $\phi_M(x)^\top$, respectively. We make the following standard independence assumption on the features.

Assumption 3 *The matrix Φ_J has rank l and the matrix Φ_M has rank m.*

We now discuss how the approximation parameters w_J and w_M are chosen. The idea behind TD methods is to fit the approximate $\tilde{J}$ and $\tilde{M}$ to obey Eq. (1) in some sense. Specifically, this is done by considering a *projection* of T onto the approximation subspaces S_J and S_M, and choosing $\tilde{J}$ and $\tilde{M}$ as the unique fixed point of this projected operator. As outlined earlier, our ultimate goal is to learn w_J and w_M from simulated trajectories of the MDP. Thus, it is constructive to consider projections onto S_J and S_M with respect to a norm that is weighted according to the state occupancy in these trajectories. We now define this projection.

For a trajectory $x_0, \dots, x_{\tau-1}$, where x_0 is drawn from a fixed distribution $\zeta_0(x)$, and the states evolve according to the MDP with policy π, define the state occupancy probabilities

$$q_t(x) = P(x_t = x), \quad x \in X, \quad t = 0, 1, \dots,$$

and let

$$q(x) = \sum_{t=0}^{\infty} q_t(x), \quad x \in X,$$
$$Q \triangleq diag(q).$$

We make the following assumption on the policy π and initial distribution ζ_0

Assumption 4 *Each state has a positive probability of being visited, namely, $q(x) > 0$ for all $x \in X$.*

Note that if ζ_0 may be controlled (for example, if we have access to a simulator), Assumption 4 may be trivially satisfied by choosing a positive ζ_0 for all states. Alternatively, if some state has a zero probability of being visited under π, then it is irrelevant for policy evaluation, and we can remove it from the state space. In this case, so long as Assumption 3 still holds (i.e., the linear features remain identifiable) all our subsequent derivations remain valid.

For vectors in $\mathbb{R}^n$, we recall the weighted Euclidean norm

$$\|y\|_q = \sqrt{\sum_{i=1}^{n} q(i) \left(y(i)\right)^2}, \quad y \in \mathbb{R}^n,$$

and we denote by Π_J and Π_M the projections from $\mathbb{R}^n$ onto the subspaces S_J and S_M, respectively, with respect to this norm. Note that the projection operators Π_J and Π_M are linear, and may be written explicitly as $\Pi_J = \Phi_J(\Phi_J^\top Q \Phi_J)^{-1}\Phi_J^\top Q$, and similarly for Π_M.

For some $z \in \mathbb{R}^{2n}$ we denote by Π the projection of z_J onto S_J and z_M onto S_M, namely

$$\Pi = \begin{pmatrix} \Pi_J & 0 \\ 0 & \Pi_M \end{pmatrix}. \tag{3}$$

We are now ready to fully describe our approximation scheme. We consider the *projected* fixed point equation

$$z = \Pi T z, \tag{4}$$

and, letting z^* denote its solution (which we will show to be unique), propose the approximate value function $\tilde{J} = z^*_J \in S_J$ and second moment function $\tilde{M} = z^*_M \in S_M$.

We shall now derive an important property of the projected operator ΠT, namely, that it is a contraction. This leads to the uniqueness of z^*, and to a simple bound on the approximation error. As in regular TD algorithms, this contraction property also underlies the convergence of several sampling-based algorithms, to be presented in the next section.

We begin by stating a well known result (Proposition 7.1.1 of Bertsekas, 2012) regarding the contraction properties of the *projected Bellman operator* $\Pi_J T_J$, where $T_J y = r + \gamma P y$.

Lemma 5 *(Proposition 7.1.1 of Bertsekas, 2012) Let Assumptions 1, 3, and 4 hold. The linear operator P and the projected linear operator $\Pi_J P$ are non-expansions in the $\|\cdot\|_q$ norm, and satisfy*

$$\|\Pi_J P y\|_q \leq \|P y\|_q \leq \|y\|_q \quad \forall y \in \mathbb{R}^n.$$

In addition, $\Pi_J P$ is a contraction in some norm, i.e., there exists some norm $\|\cdot\|_J$ and some $\beta_J < 1$ such that

$$\|\Pi_J P y\|_J \leq \beta_J \|y\|_J, \quad \forall y \in \mathbb{R}^n.$$

Lemma 5 immediately leads to the following result:

Lemma 6 *Let Assumptions 1, 3, and 4 hold. Then, there exists some norm $\|\cdot\|_J$ and some $\beta_J < 1$ such that*

$$\|\gamma \Pi_J P y\|_J \leq \beta_J \|y\|_J, \quad \forall y \in \mathbb{R}^n.$$

Similarly, there exists some norm $\|\cdot\|_M$ and some $\beta_M < 1$ such that

$$\|\gamma \Pi_M P y\|_M \leq \beta_M \|y\|_M, \quad \forall y \in \mathbb{R}^n.$$

Note that for $\gamma < 1$, Lemma 6 holds with the norm $\|\cdot\|_q$ and contraction modulus γ, by the non-expansiveness property of $\Pi_J P$ in Lemma 5. The more difficult case $\gamma = 1$, however, requires the expressions in Lemma 6.

Next, we define a weighted-norm on $\mathbb{R}^{2n}$, in which a parameter α balances between the weight of the J components' norm, and the weight of the M components' norm, as defined in Lemma 6. The intuition behind this weighted-norm, is that by carefully selecting the balance α, we shall show that the contraction properties in Lemma 6 guarantee a contraction property for the projected operator ΠT, in this norm.

Definition 7 *For a vector $z \in \mathbb{R}^{2n}$ and a scalar $0 < \alpha < 1$, the α-weighted norm is*

$$\|z\|_\alpha = \alpha \|z_J\|_J + (1-\alpha)\|z_M\|_M, \tag{5}$$

where $\|\cdot\|_J$ and $\|\cdot\|_M$ are defined in Lemma 6.

Our main result of this section is given in the following proposition, where we show that the projected operator ΠT is a contraction with respect to a suitable α-weighted norm.

Proposition 8 *Let Assumptions 1, 3, and 4 hold. Then, there exists some $0 < \alpha < 1$ and some $\beta < 1$ such that ΠT is a β-contraction with respect to the α-weighted norm, i.e.,*

$$\|\Pi T z_1 - \Pi T z_2\|_\alpha \leq \beta \|z_1 - z_2\|_\alpha, \quad \forall z_1, z_2 \in \mathbb{R}^{2n}.$$

Proof From the definition of ΠT in (2) and (3), we have that for any $z_1, z_2 \in \mathbb{R}^{2n}$ we have $\|\Pi T z_1 - \Pi T z_2\|_\alpha = \|\Pi \mathcal{P}(z_1 - z_2)\|_\alpha$, where

$$\Pi \mathcal{P} = \begin{pmatrix} \gamma \Pi_J P & 0 \\ 2\gamma \Pi_M R P & \gamma^2 \Pi_M P \end{pmatrix}.$$

Thus, it suffices to show that for all $z \in \mathbb{R}^{2n}$

$$\|\Pi \mathcal{P} z\|_\alpha \leq \beta \|z\|_\alpha.$$

We will now show that $\|\Pi \mathcal{P} z\|_\alpha$ may be separated into two terms which may be bounded by Lemma 6, and an additional cross term. By balancing α and β, this term may be contained to yield the required contraction.

We have

$$\begin{aligned} \|\Pi \mathcal{P} z\|_\alpha =& \alpha \|\gamma \Pi_J P z_J\|_J \\ &+ (1-\alpha)\|2\gamma \Pi_M R P z_J + \gamma^2 \Pi_M P z_M\|_M \\ \leq& \alpha \|\gamma \Pi_J P z_J\|_J + (1-\alpha)\|\gamma^2 \Pi_M P z_M\|_M \\ &+ (1-\alpha)\|2\gamma \Pi_M R P z_J\|_M \\ \leq& \alpha \beta_J \|z_J\|_J + (1-\alpha)\gamma \beta_M \|z_M\|_M \\ &+ (1-\alpha)\|2\gamma \Pi_M R P z_J\|_M, \end{aligned} \tag{6}$$

where the equality is by definition of the α weighted norm (5), the first inequality is from the triangle inequality, and the second inequality is by Lemma 6. Now, we claim that there exists some finite C such that

$$\|2\gamma \Pi_M R P y\|_M \leq C \|y\|_J, \quad \forall y \in \mathbb{R}^n. \tag{7}$$

To see this, note that since $\mathbb{R}^n$ is a finite dimensional real vector space, all vector norms are equivalent (Horn and Johnson, 2012, Corollary 5.4.5) therefore there exist finite C_1 and C_2 such that for all $y \in \mathbb{R}^n$

$$C_1 \|2\gamma \Pi_M R P y\|_2 \leq \|2\gamma \Pi_M R P y\|_M \leq C_2 \|2\gamma \Pi_M R P y\|_2,$$

where $\|\cdot\|_2$ denotes the Euclidean norm. Let λ denote the spectral norm of the matrix $2\gamma \Pi_M R P$, which is finite since all the matrix elements are finite. We have that

$$\|2\gamma \Pi_M R P y\|_2 \leq \lambda \|y\|_2, \quad \forall y \in \mathbb{R}^n.$$

Using again the fact that all vector norms are equivalent, there exists a finite C_3 such that

$$\|y\|_2 \leq C_3 \|y\|_J, \quad \forall y \in \mathbb{R}^n.$$

Setting $C = C_2 \lambda C_3$ we get the desired bound. Let $\tilde{\beta} = \max\{\beta_J, \gamma \beta_M\} < 1$, and choose $\epsilon > 0$ such that

$$\tilde{\beta} + \epsilon < 1.$$

Now, choose α such that $\alpha = \frac{C}{\epsilon + C}$. We have that

$$(1-\alpha)C = \alpha\epsilon,$$

and plugging this into (7) yields

$$(1-\alpha)\|2\gamma\Pi_M RPy\|_M \leq \alpha\epsilon\|y\|_J. \tag{8}$$

We now return to (6), where we have

$$\begin{aligned}
&\alpha\beta_J\|z_J\|_J + (1-\alpha)\gamma\beta_M\|z_M\|_M + (1-\alpha)\|2\gamma\Pi_M RPz_J\|_M \\
\leq&\alpha\beta_J\|z_J\|_J + (1-\alpha)\gamma\beta_M\|z_M\|_M + \alpha\epsilon\|z_J\|_J \\
\leq&(\tilde{\beta}+\epsilon)\left(\alpha\|z_J\|_J + (1-\alpha)\|z_M\|_M\right),
\end{aligned}$$

where the first inequality is by (8), and the second is by the definition of $\tilde{\beta}$. We have thus shown that

$$\|\Pi\mathcal{P}z\|_\alpha \leq (\tilde{\beta}+\epsilon)\|z\|_\alpha.$$

Finally, choose $\beta = \tilde{\beta} + \epsilon$. ■

Proposition 8 guarantees that the projected operator ΠT has a unique fixed point. Let us denote this fixed point by z^*, and let w_J^*, w_M^* denote the corresponding weights, which are unique due to Assumption 3

$$\begin{aligned}
\Pi T z^* &= z^*, \\
z_J^* &= \Phi_J w_J^*, \\
z_M^* &= \Phi_M w_M^*.
\end{aligned} \tag{9}$$

In the next proposition, using a standard result of Bertsekas and Tsitsiklis (1996), we provide a bound on the approximation error.

Proposition 9 *Let Assumptions 1, 3, and 4 hold. Denote by $z_{true} \in \mathbb{R}^{2n}$ the true value and second moment functions, i.e., $[z_{true}]_J = J$, and $[z_{true}]_M = M$. Then,*

$$\|z_{true} - z^*\|_\alpha \leq \frac{1}{1-\beta}\|z_{true} - \Pi z_{true}\|_\alpha,$$

with α and β defined in Proposition 8.

Proof This result is similar to Lemma 6.9 in Bertsekas and Tsitsiklis (1996). We have

$$\begin{aligned}
\|z_{true} - z^*\|_\alpha \leq& \|z_{true} - \Pi z_{true}\|_\alpha + \|\Pi z_{true} - z^*\|_\alpha \\
=& \|z_{true} - \Pi z_{true}\|_\alpha + \|\Pi T z_{true} - \Pi T z^*\|_\alpha \\
\leq& \|z_{true} - \Pi z_{true}\|_\alpha + \beta\|z_{true} - z^*\|_\alpha.
\end{aligned}$$

Rearranging gives the stated result. ■

Note that by definition, Πz_{true} is the best approximation we can hope for (in terms of the α-weighted squared error) in our approximation subspace. Thus, the approximation error $\|z_{true} - z^*\|_\alpha$ is ultimately bounded by the choice of features, which in practice should be chosen wisely.

At this point, the reader may question the usefulness of the projected fixed-point approximation over simpler approximation schemes, such as the direct projection Πz_{true}. As we show in the next section, the projected fixed-point architecture supports a family of sampling-based TD estimation algorithms, with efficient batch and online implementations. Furthermore, as we show empirically in Section 6, these TD algorithms perform well in practice, especially in the regime of a small sample size. For conventional TD algorithms, these benefits are well-established (Bertsekas, 2012), and gave rise to their popularity. Here we extend this to the variance of the reward-to-go.

4. Simulation Based Estimation Algorithms

In this section we propose algorithms that estimate $\tilde{J}$ and $\tilde{M}$ from sampled trajectories of the MDP, based on the approximation architecture of the previous section.

We begin by writing the projected equation (9) in matrix form. First, let us write the equation explicitly as

$$\begin{aligned} \Pi_J \left(r + \gamma P \Phi_J w_J^*\right) &= \Phi_J w_J^*, \\ \Pi_M \left(Rr + 2\gamma R P \Phi_J w_J^* + \gamma^2 P \Phi_M w_M^*\right) &= \Phi_M w_M^*. \end{aligned} \tag{10}$$

Recalling the definition of Q, projecting a vector y onto Φw satisfies the following orthogonality condition

$$\Phi^\top Q(y - \Phi w) = 0.$$

We therefore have

$$\begin{aligned} \Phi_J^\top Q \left(\Phi_J w_J^* - \left(r + \gamma P \Phi_J w_J^*\right)\right) &= 0, \\ \Phi_M^\top Q \left(\Phi_M w_M^* - \left(Rr + 2\gamma R P \Phi_J w_J^* + \gamma^2 P \Phi_M w_M^*\right)\right) &= 0, \end{aligned}$$

which can be written as

$$A w_J^* = b, \quad C w_M^* = d, \tag{11}$$

with

$$\begin{aligned} A &= \Phi_J^\top Q \left(I - \gamma P\right) \Phi_J, \quad b = \Phi_J^\top Q r, \\ C &= \Phi_M^\top Q \left(I - \gamma^2 P\right) \Phi_M, \quad d = \Phi_M^\top Q R \left(r + 2\gamma P \Phi_J A^{-1} b\right), \end{aligned} \tag{12}$$

and the matrices A and C are invertible since Proposition 8 guarantees a unique solution to (9) and Assumption 3 guarantees the unique weights of its projection.

Let us now outline our proposed algorithms. The first algorithm is a variant of the Least Squares Temporal Difference algorithm (LSTD; Boyan 2002), and aims to solve Eq. (11) directly, by forming sample based estimates of the terms A, b, C, and d. This is a batch algorithm that is known to make efficient use of data in its nominal version, and as we show empirically, demonstrates efficient performance in our case as well. The second algorithm is a variant of online TD(0) (Sutton and Barto, 1998). In its nominal form, TD(0) has been successfully used as the critic in actor-critic algorithms (Konda and Tsitsiklis, 2003). Our extended TD(0) variant may be used similarly in a *risk-adjusted* actor-critic algorithm (Tamar and Mannor, 2013; Prashanth and Ghavamzadeh, 2013). The third algorithm is a variant of LSTD(λ), in which, similarly to standard LSTD(λ), Eq. (11) is extended to its multi-step counterpart. The fourth algorithm is not based on the TD equation (11), but uses least squares regression to estimate the direct projection Πz_{true}. We compare this algorithm with the LSTD variants in Section 6.

4.1 A Least Squares TD Algorithm

Our first simulation-based algorithm is an extension of the LSTD algorithm (Boyan, 2002). We simulate N trajectories of the MDP with the policy π and initial state distribution ζ_0. Let $x_0^k, x_1^k, \ldots, x_{\tau^k-1}^k$ and τ^k, where $k = 0, 1, \ldots, N$, denote the state sequence and visit times to the terminal state within these trajectories, respectively. We now use these trajectories to form the following estimates of the terms in (12)

$$\begin{aligned}
A_N &= \mathbb{E}_N\left[\sum_{t=0}^{\tau-1}\phi_J(x_t)(\phi_J(x_t)-\gamma\phi_J(x_{t+1}))^\top\right],\\
b_N &= \mathbb{E}_N\left[\sum_{t=0}^{\tau-1}\phi_J(x_t)r(x_t)\right],\\
C_N &= \mathbb{E}_N\left[\sum_{t=0}^{\tau-1}\phi_M(x_t)(\phi_M(x_t)-\gamma^2\phi_M(x_{t+1}))^\top\right],\\
d_N &= \mathbb{E}_N\left[\sum_{t=0}^{\tau-1}\phi_M(x_t)r(x_t)\left(r(x_t)+2\gamma\phi_J(x_{t+1})^\top A_N^{-1}b_N\right)\right],
\end{aligned} \tag{13}$$

where $\mathbb{E}_N$ denotes an empirical average over trajectories, i.e., $\mathbb{E}_N\left[f(x,\tau)\right] = \frac{1}{N}\sum_{k=1}^N f(x^k,\tau^k)$. The LSTD approximation is given by

$$\hat{w}_J^* = A_N^{-1}b_N, \quad \hat{w}_M^* = C_N^{-1}d_N.$$

The next theorem shows that LSTD converges.

Theorem 10 *Let Assumptions 1, 3, and 4 hold. Then $\hat{w}_J^* \to w_J^*$ and $\hat{w}_M^* \to w_M^*$ as $N \to \infty$ with probability 1.*

The proof involves a straightforward application of the law of large numbers and is described in Appendix B. For regular LSTD, $\mathcal{O}(1/\sqrt{n})$ convergence rates were derived under certain mixing conditions of the MDP by Konda (2002, based on a central limit theorem argument) and Lazaric et al. (2010, based on a finite time analysis), and may be extended to the algorithm presented here.

4.2 An Online TD(0) Algorithm

Our second estimation algorithm is an extension of the well known TD(0) algorithm (Sutton and Barto, 1998). Again, we simulate trajectories of the MDP corresponding to the policy

π and initial state distribution ζ_0, and we iteratively update our estimates at every visit to the terminal state. An extension to an algorithm that updates at every state transition is also possible, but we do not pursue such here.

For some $0 \leq t < \tau^k$ and weights w_J, w_M, we introduce the TD terms

$$\begin{aligned}\delta_J^k(t, w_J, w_M) =& r(x_t^k) + \left(\gamma\phi_J(x_{t+1}^k)^\top - \phi_J(x_t^k)^\top\right) w_J,\\ \delta_M^k(t, w_J, w_M) =& r^2(x_t^k) + 2\gamma r(x_t^k)\phi_J(x_{t+1}^k)^\top w_J\\ &+ \left(\gamma^2\phi_M(x_{t+1}^k)^\top - \phi_M(x_t^k)^\top\right) w_M.\end{aligned}$$

Note that δ_J^k is the standard TD error (Sutton and Barto, 1998). For the intuition behind δ_M^k, observe that M in (1) is equivalent to the value function of an MDP with stochastic reward $r(x)^2 + 2\gamma r(x)J(x')$, where $x' \sim P(x'|x)$. The TD term δ_M^k is the equivalent TD error, with $\phi_J(x')^\top w_J$ substituting $J(x')$. The TD(0) algorithm is given by

$$\begin{aligned}\hat{w}_{J;k+1} &= \hat{w}_{J;k} + \xi_k \sum_{t=0}^{\tau^k-1} \phi_J(x_t)\delta_J^k(t, \hat{w}_{J;k}, \hat{w}_{M;k}),\\ \hat{w}_{M;k+1} &= \hat{w}_{M;k} + \xi_k \sum_{t=0}^{\tau^k-1} \phi_M(x_t)\delta_M^k(t, \hat{w}_{J;k}, \hat{w}_{M;k}),\end{aligned}$$

where $\{\xi_k\}$ are positive step sizes.

The next theorem shows that TD(0) converges.

Theorem 11 *Let Assumptions 1, 3, and 4 hold, and let the step sizes satisfy*

$$\sum_{k=0}^{\infty} \xi_k = \infty, \quad \sum_{k=0}^{\infty} \xi_k^2 < \infty.$$

Then $\hat{w}_{J;k} \to w_J^*$ and $\hat{w}_{M;k} \to w_M^*$ as $k \to \infty$ with probability 1.

Proof The proof is based on representing the algorithm as a stochastic approximation, and uses a result of Borkar (2008) to show that the iterates asymptotically track a certain ordinary differential equation (ODE). This ODE will then be shown to have a unique asymptotically stable equilibrium exactly at w_J^*, w_M^*.

A straightforward expectation calculation (see (22) and (23) in Appendix B for the derivation) shows that for all k we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{\tau^k-1} \phi_J(x_t)\delta_J^k(t, w_J, w_M)\right] &= \Phi_J^\top Q r - \Phi_J^\top Q\left(I - \gamma P\right)\Phi_J w_J,\\ \mathbb{E}\left[\sum_{t=0}^{\tau^k-1} \phi_M(x_t)\delta_M^k(t, w_J, w_M)\right] &= \Phi_M^\top Q R\left(r + 2\gamma P\Phi_J w_J\right) - \Phi_M^\top Q\left(I - \gamma^2 P\right)\Phi_M w_M.\end{aligned}$$

Letting $\hat{w}_k = (\hat{w}_{J;k}, \hat{w}_{M;k})$ denote a concatenated weight vector in the joint space $\mathbb{R}^l \times \mathbb{R}^m$ we can write the TD algorithm in a stochastic approximation form as

$$\hat{w}_{k+1} = \hat{w}_k + \xi_k\left(z + M\hat{w}_k + \delta M_{k+1}\right), \tag{14}$$

where

$$M = \begin{pmatrix} \Phi_J^\top Q\left(\gamma P - I\right)\Phi_J & 0 \\ 2\gamma\Phi_M^\top QRP\Phi_J & \Phi_M^\top Q\left(\gamma^2 P - I\right)\Phi_M \end{pmatrix},$$

$$z = \begin{pmatrix} \Phi_J^\top Qr \\ \Phi_M^\top QRr \end{pmatrix},$$

and the noise terms δM_{k+1} satisfy

$$\mathbb{E}\left[\delta M_{k+1}|F_n\right] = 0,$$

where F_n is the filtration $F_n = \sigma(\hat{w}_m, \delta M_m, m \le n)$, since different trajectories are independent.

We first claim that the eigenvalues of M have a negative real part. To see this, observe that M is block triangular, and its eigenvalues are just the eigenvalues of $M_1 \triangleq \Phi_J^\top Q\left(\gamma P - I\right)\Phi_J$ and $M_2 \triangleq \Phi_M^\top Q\left(\gamma^2 P - I\right)\Phi_M$. Lemma 6.10 of Bertsekas and Tsitsiklis (1996), shows that under Assumptions 1 and 4, the matrix $Q(\gamma P - I)$ is negative definite in the sense that $x^\top(\gamma P - I)x < 0 \quad \forall x \neq 0$ (Lemma 6.10 of Bertsekas and Tsitsiklis, 1996 is stated for the case $\gamma = 1$, but an extension to the simpler discounted case is trivial). By Assumption 3, this implies that the matrices M_1 and M_2 are negative definite in the sense that $x^\top M_1 x < 0 \quad \forall x \neq 0$, and $x^\top M_2 x < 0 \quad \forall x \neq 0$. Example 6.6 of Bertsekas, 2012 shows that the eigenvalues of a negative definite matrix have a negative real part. It therefore follows that the eigenvalues of M_1 and M_2 have a negative real part. Thus, the eigenvalues of M have a negative real part.

Next, let $h(w) = Mw + z$, and observe that the following conditions hold.

Condition 1 *The map h is Lipschitz.*

Condition 2 *The step sizes satisfy*

$$\sum_{k=0}^{\infty}\xi_k = \infty, \quad \sum_{k=0}^{\infty}\xi_k^2 < \infty.$$

Condition 3 *$\{\delta M_n\}$ is a martingale difference sequence, i.e., $\mathbb{E}\left[\delta M_{n+1}|F_n\right] = 0$.*

The next condition also holds

Condition 4 *The functions $h_c(w) \triangleq h(cw)/c, c \ge 1$ satisfy $h_c(w) \to h_\infty(w)$ as $c \to \infty$, uniformly on compacts, and $h_\infty(w)$ is continuous. Furthermore, the ODE*

$$\dot{w}(t) = h_\infty(w(t))$$

has the origin as its unique globally asymptotically stable equilibrium.

This is easily verified by noting that $h(cw)/c = Mw + c^{-1}z$, and since z is finite, $h_c(w)$ converges uniformly as $c \to \infty$ to $h_\infty(w) = Mw$. The stability of the origin is guaranteed since the eigenvalues of M have a negative real part (Khalil and Grizzle, 1996).

Theorem 7 in Chapter 3 of Borkar (2008) states that if Conditions 1-4 hold, the following condition holds

Condition 5 *The iterates of* (14) *remain bounded almost surely, i.e.,* $\sup_k \|\hat{w}_k\| < \infty$, *a.s.*

Finally, we use a standard stochastic approximation result that, given that the above conditions hold, relates the convergence of the iterates of (14) with the asymptotic behavior of the ODE

$$\dot{w}(t) = h(w(t)). \tag{15}$$

Since the eigenvalues of M have a negative real part, (15) has a unique globally asymptotically stable equilibrium point (Khalil and Grizzle, 1996), which by (11) is exactly $\hat{w}* = (\hat{w}_J^*, \hat{w}_M^*)$. Formally, by Theorem 2 in Chapter 2 of Borkar (2008) we have that if Conditions 1, 2, 3 and 5 hold, then $\hat{w}_k \to \hat{w}*$ as $k \to \infty$ with probability 1. ■

It is interesting to note that despite the fact that the update of w_M depends on w_J, the algorithm converges using a single time scale, i.e., the same step-size schedule, for both w_J and w_M. This is in contrast with, for example, actor critic algorithms, that also have dependent updates but require multiple time-scales for convergence (Konda and Tsitsiklis, 2003). An intuitive reason for this is that the update for w_J is independent of w_M, therefore w_J will converge regardless, and w_M will 'track' it until convergence. Asymptotic convergence rates for TD(0) may also be derived along the lines of Konda (2002).

4.3 Multistep LSTD(λ) Algorithms

A common method in value function approximation (Bertsekas, 2012) is to replace the single-step mapping T_J with a multistep version, that takes into account multi-step transitions. For some $0 < \lambda < 1$, the multistep Bellman operator $T_J^{(\lambda)}$ is given by

$$T_J^{(\lambda)}(y) \triangleq (1-\lambda)\sum_{l=0}^{\infty} \lambda^l T_J^{l+1}(y) = (I - \lambda\gamma P)^{-1} r + \gamma P^{(\lambda,\gamma)} y,$$

where $P^{(\lambda,\gamma)} = (1-\lambda)\sum_{l=0}^{\infty} \lambda^l \gamma^l P^{l+1}$. The projected equation (10) then becomes

$$\Pi_J T_J^{(\lambda)}\left(\Phi_J w_J^{*(\lambda)}\right) = \Phi_J w_J^{*(\lambda)}.$$

Similarly, we may write a multistep equation for M

$$\Pi_M T_M^{(\lambda)}\left(\Phi_M w_M^{*(\lambda)}\right) = \Phi_M w_M^{*(\lambda)}, \tag{16}$$

where

$$T_M^{(\lambda)} \triangleq (1-\lambda)\sum_{l=0}^{\infty} \lambda^l T_{M^*}^{l+1},$$

and

$$T_{M^*}(y) \triangleq Rr + 2\gamma R P \Phi_J w_J^{*(\lambda)} + \gamma^2 P y.$$

Note the difference between T_{M^*} and $[T]_M$ defined earlier: we are no longer working on the joint space of J and M but instead we have an independent equation for approximating J, and its solution $w_J^{*(\lambda)}$ is part of Equation (16) for approximating M. We can also write $T_M^{(\lambda)}$ explicitly as:

$$T_M^{(\lambda)}(y) = (I - \lambda\gamma^2 P)^{-1}\left(Rr + 2\gamma RP\Phi_J w_J^{*(\lambda)}\right) + \gamma^2 P^{(\lambda,\gamma^2)} y,$$

where $P^{(\lambda,\gamma^2)} = (1-\lambda)\sum_{l=0}^{\infty}\lambda^l \gamma^{2l} P^{l+1}$.

Proposition 7.1.1 of Bertsekas (2012) shows that for any $0 < \lambda < 1$ and $0 < \gamma \leq 1$ the projected operator $\Pi_J P^{(\lambda,\gamma)}$ is a contraction in the $\|\cdot\|_q$ norm. Therefore, both $\Pi_J T_J^{(\lambda)}$ and $\Pi_M T_M^{(\lambda)}$ are contractions with respect to the $\|\cdot\|_q$ norm, and both multistep projected equations have a unique solution. In a similar manner to the single step version, the projected equations may be written in matrix form

$$A^{(\lambda)} w_J^{*(\lambda)} = b^{(\lambda)}, \quad C^{(\lambda)} w_M^{*(\lambda)} = d^{(\lambda)}, \tag{17}$$

where

$$\begin{aligned}
A^{(\lambda)} &= \Phi_J^\top Q\left(I - \gamma P^{(\lambda,\gamma)}\right)\Phi_J, \quad b^{(\lambda)} = \Phi_J^\top Q (I - \lambda\gamma P)^{-1} r,\\
C^{(\lambda)} &= \Phi_M^\top Q\left(I - \gamma^2 P^{(\lambda,\gamma^2)}\right)\Phi_M,\\
d^{(\lambda)} &= \Phi_M^\top Q (I - \lambda\gamma^2 P)^{-1} R\left(r + 2\gamma P\Phi_J w_J^{*(\lambda)}\right).
\end{aligned}$$

Simulation based estimates $A_N^{(\lambda)}$ and $b_N^{(\lambda)}$ of the expressions above may be obtained by using eligibility traces, as described in Section 6.3.6 of Bertsekas (2012), and the LSTD(λ) approximation is then given by $\hat{w}_J^{*(\lambda)} = (A_N^{(\lambda)})^{-1} b_N^{(\lambda)}$. By substituting $w_J^{*(\lambda)}$ with $\hat{w}_J^{*(\lambda)}$ in the expression for $d^{(\lambda)}$, a similar procedure may be used to derive estimates $C_N^{(\lambda)}$ and $d_N^{(\lambda)}$, and to obtain the LSTD(λ) approximation $\hat{w}_M^{*(\lambda)} = (C_N^{(\lambda)})^{-1} d_N^{(\lambda)}$. A convergence result similar to Theorem 10 may also be obtained. Due to the similarity to the LSTD procedure in (13), the details are omitted. Finally, we note that a straightforward modification of the TD(0) algorithm to a multistep TD(λ) variant is also possible, using eligibility traces and following the procedure described in Section 6.3.6 of Bertsekas (2012).

4.4 A Direct Least Squares Regression Algorithm

We conclude this section with a simple regression style algorithm, which is not based on the TD approximation architecture of Section 3, but to our knowledge has not been proposed before.

As before, we let $x_0^k, x_1^k, \ldots, x_{\tau^k-1}^k$ denote the state sequence of the k'th simulated trajectory, and define the regression targets as

$$\hat{B}_t^k = \sum_{i=t}^{\tau^k-1} \gamma^{i-t} r(x_t^k).$$

Our approximation weights are now given by the solutions to the least squares problems

$$\hat{w}_J^* = \arg\min_{w_J} \sum_{k=1}^{N}\sum_{t=0}^{\tau^k-1}\left(\phi_J(x_t^k)^\top w_J - \hat{B}_t^k\right)^2,$$

and

$$\hat{w}_M^* = \arg\min_{w_M} \sum_{k=1}^{N} \sum_{t=0}^{\tau^k - 1} \left(\phi_M(x_t^k)^\top w_M - \left(\hat{B}_t^k\right)^2 \right)^2 .$$

It may easily be verified that the approximate value $\tilde{J}$ and second moment $\tilde{M}$ of such a procedure converge, as $N \to \infty$, to the *direct* approximations $\Pi_J J$ and $\Pi_M M$, respectively. We further explore this algorithm and its relation to TD based algorithms in the empirical evaluation of Section 6.

5. Non Negative Approximate Variance

The TD algorithms of the preceding section approximate J and M by the solution to the fixed point equation (9). While Proposition 9 shows that the approximation errors of $\tilde{J}$ and $\tilde{M}$ are bounded, it does not guarantee that the approximated variance $\tilde{V}$, given by $\tilde{M} - \tilde{J}^2$, is non-negative for all states. A trivial remedy is to set all negative values of $\tilde{V}$ to zero; however, by such we lose information in these states. In this section we propose two alternative approaches to this problem. The first is through the choice of features, where we show that for the direct approximation $\Pi_J J$ and $\Pi_M M$, we can choose features that guarantee non-negative variance.

The second approach is based on the observation that non-negativeness of the variance may be written as a linear constraint in the weights for M. By adding such constraints to the projection in the fixed point equation (9), we obtain a different approximation architecture, in which non-negative variance is inherent. We show that this approximation scheme may be computed efficiently.

5.1 A Suitable Features Approach

For this section consider the direct approximation of J and M, as in Section 4.4, where we have $\tilde{J} = \Pi_J J$ and $\tilde{M} = \Pi_M M$. We investigate conditions under which $\tilde{M}(x) - \tilde{J}(x)^2 \geq 0$ for all $x \in X$.

Consider the following assumptions on the features:

Assumption 12 *The same features are used for J and M, i.e., $\Phi_J = \Phi_M$.*

Assumption 13 *The features are able to exactly represent a constant function, i.e., there exists w such that $\phi_J(x)^\top w = 1$ for all $x \in X$.*

We claim that Assumptions 12 and 13 suffice for guaranteeing non-negative approximate variance.

Proposition 14 *Let Assumptions 12 and 13 hold. Then $\tilde{M}(x) - \tilde{J}(x)^2 \geq 0$ for all $x \in X$.*

Proof First, by definition we have

$$V(x) = M(x) - J(x)^2 \geq 0. \tag{18}$$

Next, observe that Assumption 12 implies $\Pi_J = \Pi_M$.

Let $x \in X$, and recall that the projection operator is linear, thus we can write

$$\tilde{J}(x) = \sum_{i \in X} J(i)\omega_x(i), \quad \tilde{M}(x) = \sum_{i \in X} M(i)\omega_x(i), \tag{19}$$

where $\omega_x(i)$ are the projection weights for state x. Let $\bar{\omega}_x = \sum_{i \in X} \omega_x(i)$. We have

$$\tilde{J}(x)^2 = \bar{\omega}_x^2 \left(\sum_{i \in X} J(i) \frac{\omega_x(i)}{\bar{\omega}_x} \right)^2 \leq \bar{\omega}_x^2 \sum_{i \in X} J(i)^2 \frac{\omega_x(i)}{\bar{\omega}_x} \leq \bar{\omega}_x \sum_{i \in X} M(i)\omega_x(i) = \bar{\omega}_x \tilde{M}(x),$$

where the first inequality is by Jensen's inequality, the second inequality is by (18), and the equalities are by (19). Thus, $\bar{\omega}_x \leq 1$ guarantees $\tilde{V}(x) \geq 0$. We now claim that Assumption 13 guarantees $\bar{\omega}_x = 1$ for all x. To see this, consider a constant value function $J = 1$ for all states; clearly we have $\tilde{J} = 1$, as the weighted Euclidean error for this approximation is zero. Plugging in (19) gives $\sum_{i \in X} \omega_x(i) = 1$ for all x. ■

Proposition 14 concerns the approximation architecture itself, and not the estimation procedure. Therefore, it applies to the algorithms discussed above only asymptotically.

Many popular linear function approximation features such as grid tiles and CMAC's (Sutton and Barto, 1998) are able to represent a constant function. For these schemes, $\tilde{V}(x) \geq 0$ is guaranteed. For other schemes, we can guarantee $\tilde{V}(x) \geq 0$ by simply adding a constant feature to the feature set. Thus, at least for the direct approximation, it appears that a non-negative approximate variance is easily obtained. Whether a similar procedure may be applied to the fixed-point approximation is currently not known. However, Proposition 9 suggests that at least when the contraction modulus is small, the fixed-point approximation should behave similarly to the direct approximation. In the next section we propose a different approach, which *modifies* the fixed-point approximation to guarantee non-negative variance, regardless of the choice of features.

5.2 A Linearly Constrained Projection Approach

In this section we show that by adding linear constraints to the projected fixed point equation, we can guarantee a non-negative approximate variance. This modified approximation architecture admits a computationally efficient solution by a modification of the LSTD algorithm of Section 4.

First, let us write the equation for the second moment weights (10) with the projection operator as an explicit minimization

$$w_M^* = \arg\min_w \|\Phi_M w - \left(Rr + 2\gamma R P \Phi_J w_J^* + \gamma^2 P \Phi_M w_M^* \right)\|_q.$$

Observe that a non-negative approximate variance in some state x may be written as a *linear* inequality in w_M^* (but non-linear in w_J^*)

$$\phi_M(x)^\top w_M^* - (\phi_J(x)^\top w_J^*)^2 \geq 0.$$

We propose to add such inequality constraints to the projection operator. Let $\{x_1, \ldots, x_s\}$ denote a set of states in which we demand that the variance be non-negative. Let $H \in \mathbb{R}^{s \times m}$

denote a matrix with the features $-\phi_M^\top(x_i)$ as its rows, and let $g \in \mathbb{R}^s$ denote a vector with elements $-(\phi_J(x_i)^\top w_J^*)^2$. We write the non-negative-variance projected equation for the second moment as

$$w_M^+ = \begin{cases} \arg\min_w & \|\Phi_M w - \left(Rr + 2\gamma RP\Phi_J w_J^* + \gamma^2 P\Phi_M w_M^+\right)\|_q \\ \text{s.t.} & Hw \leq g \end{cases}. \tag{20}$$

Here, w_M^+ denotes the weights of $\tilde{M}$ in the *modified* approximation architecture. We now discuss whether a solution to (20) exists, and how it may be obtained.

Let us assume that the constraints in (20) admit a feasible solution:

Assumption 15 *There exists w such that $Hw < g$.*

Note that a trivial way to satisfy Assumption 15 is to have some feature vector that is positive for all states. To see this, let i^+ denote the index of the positive feature vector, and choose w to be all zeros, except for the i^+ element, which should satisfy $w_{i^+} < -\left(\max_{1\leq i\leq s}|g_i|\right) / \left(\max_{1\leq i\leq s} H_{i,i^+}\right)$.

Equation (20) is a form of projected equation studied by Bertsekas (2011), the solution of which exists, and may be obtained by the following iterative procedure

$$w_{k+1} = \Pi_{\Xi,\hat{W}_M}[w_k - \eta\Xi^{-1}(Cw_k - d)], \tag{21}$$

where C and d are defined in (12), Ξ is an arbitrary positive definite symmetric matrix, $\eta \in \mathbb{R}$ is a positive step size, and $\Pi_{\Xi,\hat{W}_M}$ denotes a projection onto the convex set $\hat{W}_M = \{w|Hw \leq g\}$ with respect to the Ξ weighted Euclidean norm.

The following lemma, which is based on a convergence result of Bertsekas (2011), guarantees that for $\gamma < 1$, the iteration (21) converges. For the non-discounted setting a similar result may be obtained by using the multi-step approach with $\lambda > 0$, as detailed in Tamar et al. (2013).

Lemma 16 *Assume $\gamma < 1$, and let Assumptions 1, 3, 4, and 15 hold. Then* (20) *admits a unique solution w_M^+, and there exists $\bar{\eta} > 0$ such that $\forall \eta \in (0, \bar{\eta})$ and $\forall w_0 \in \mathbb{R}^m$ the iteration* (21) *converges at a linear rate to w_M^+ (i.e., $\|w_k - w_M^+\|$ converges to 0 at least as fast as a geometric progression).*

Proof Bertsekas (2011) shows that projected fixed-point equations of the form

$$w^* = \begin{cases} \arg\min_w & \|\Phi w - T_{\text{lin}}(\Phi w^*)\|_q \\ \text{s.t.} & w \in \Omega \end{cases},$$

where $T_{\text{lin}}(y) = A_{\text{lin}}y + b_{\text{lin}}$ is a contracting linear operator, and Ω is a polyhedral set, may be solved iteratively by

$$w_{k+1} = \Pi_{\Xi,\Omega}[w_k - \eta\Xi^{-1}(C_{\text{lin}}w_k - d_{\text{lin}})],$$

where $\Pi_{\Xi,\Omega}$ projects onto Ω w.r.t. the norm $\|y\|_\Xi = \sqrt{y^\top \Xi y}$ for an arbitrary symmetric and positive-definite matrix Ξ, $C_{\text{lin}} = \Phi^\top Q(I - A_{\text{lin}})\Phi$ and $d_{\text{lin}} = \Phi' Q b_{\text{lin}}$. Specifically, the convergence result of Bertsekas (2011) shows that when T_{lin} is a contraction in the $\|\cdot\|_q$

norm, Ω is polyhedral, and Φ is full rank, there exists $\bar{\eta} > 0$ such that for all $\eta \in (0, \bar{\eta})$, and for all $w_0 \in \mathbb{R}^m$, the preceding iteration converges at a linear rate to the unique solution of the projected fixed point equation described above.

Substituting $T_{\text{lin}}(y)$ with $T_M(y) = Rr + 2\gamma RP\Phi_J w_J^* + \gamma^2 Py$, and Ω with the set defined by $Hw \leq g$, we obtain the projected fixed point equation (20), and the corresponding iteration (21). To apply the convergence result, the full-rank of Φ_M is guaranteed by Assumption 3, and the contraction of T_M in the $\| \cdot \|_q$ norm is guaranteed by Lemma 5, since P is a non-expansion and $\gamma < 1$. ■

Generally, C, d, and w_J^* are not known in advance, and should be replaced in (21) with their simulation based estimates, C_N, d_N, and $\hat{w}_J^*$, proposed in the previous section. The convergence of these estimates, together with the result of Lemma 16, lead to the following result; the proof is detailed in Appendix C.

Theorem 17 *Consider the algorithm in* (21) *with C, d, and w_J^* replaced by C_N, d_N, and $\hat{w}_J^*$, respectively, and with $k(N)$ replacing k for a specific N. Also, let the assumptions in Lemma 16 hold, and let $\eta \in (0, \bar{\eta})$, with $\bar{\eta}$ defined in Lemma 16. Then $w_{k(N)} \to w_M^+$ as $N \to \infty$ and $k(N) \to \infty$ almost surely. Namely, for any $\bar{\epsilon} > 0$ w.p. 1 there is a $N(\bar{\epsilon})$ such that for any $N > N(\bar{\epsilon})$ there is a $k(N, \bar{\epsilon})$, such that for all $k > k(N, \bar{\epsilon})$ we have $\|w_{k;N} - w_M^+\| \leq \bar{\epsilon}$.*

We remark that we do not know how to quantify how the linear constraints affect the approximation error. While intuitively our constraints add prior information that is 'correct' in some sense (since we know that the true variance is positive), it is not hard to construct examples where the constraints actually increase the error. In the following, we provide an illustration of the linearly constrained projection approach on a toy problem. We qualitatively show that the method effectively produces a non-negative solution, without significantly affecting the approximation error.

Consider the Markov chain depicted in Figure 1, which consists of n states with reward -1 and a terminal state x^* with zero reward. Assume no discounting, i.e., $\gamma = 1$. The transitions from each state is either to a subsequent state (with probability p) or to a preceding state (with probability $1-p$), with the exception of the first state which transitions to itself instead. We chose to approximate J and M with polynomials of degree 1 and 2, respectively, i.e., $\Phi_J(x) = [1, x]^\top$ and $\Phi_M(x) = [1, x, x^2]^\top$. For such a small problem, the fixed point equation (17) may be solved exactly, yielding the approximation depicted in Figure 2 (dotted line), for $p = 0.7$, $N = 30$, and $\lambda = 0.95$. Note that the variance, in Figure 2C, is negative for the last two states. Using algorithm (21) we obtained a positive variance constrained approximation, which is depicted in Figure 2 (dashed line). Note how the approximate variance has been adjusted to be positive for all states.

6. Experiments

In this section we present numerical simulations of policy evaluation for an option pricing domain. We show that in terms of sample efficiency, our LSTD(λ) algorithm significantly outperforms the current state-of-the-art. We begin by describing the domain and its modeling as an MDP, and then present our policy evaluation results. We emphasize that our

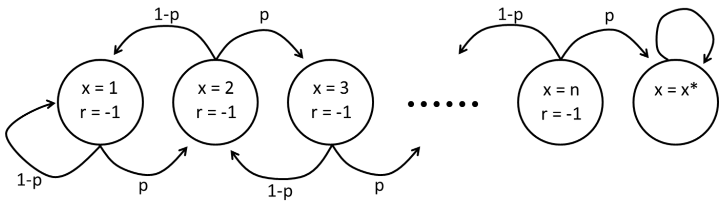

Figure 1: An example Markov chain.

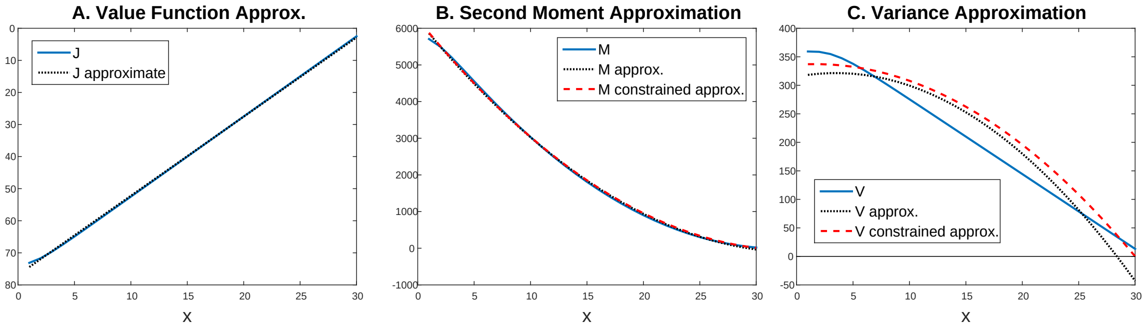

Figure 2: Value, second moment and variance approximation.

results only concern policy evaluation, and not policy optimization. The following MDP description is given for the purpose of presentation completeness.

6.1 Option Pricing

An American-style put option (Hull, 2006) is a contract which gives the owner the right, but not the obligation, to sell an asset at a specified strike price K on or before some maturity time t^*. Letting x_t denote the price (state) of the asset at time $t \leq t^*$, the immediate payoff of executing the option at that time is therefore $\max(0, K - x_t)$. Assuming Markov state transitions, an optimal execution policy may be found by solving a finite horizon MDP, and the expected profit under that policy is termed the 'fair' price of the option. Since the state space is typically continuous, an exact solution is infeasible, calling for approximate, sampling based techniques (Longstaff and Schwartz, 2001; Tsitsiklis and Van Roy, 2001; Li et al., 2009).

The option pricing problem may be formulated as an MDP as follows. To account for the finite horizon, we include time explicitly in the state, thus, the state at time t is $\{x_t; t\}$. The action set is binary, where 1 stands for executing the option and 0 for continuing to hold it. Once an option is executed, or when $t = t^*$, a transition to a terminal state takes place. Otherwise, the state transitions to $\{x_{t+1}; t + 1\}$ where x_{t+1} is determined by a stochastic kernel $P(x_{t+1}|x_t, t)$. In our experiments we used a Bernoulli price fluctuation model (Cox et al., 1979),

$$x_{t+1} = \begin{cases} f_u x_t, & \text{w.p. } p \\ f_d x_t, & \text{w.p. } 1-p \end{cases},$$

where the up and down factors, f_u and f_d, are constant. The reward for executing $u = 1$ at state x is $r(x) \triangleq \max(0, K - x)$ and zero otherwise. Note that by definition, for any state x in which the policy decides to execute, the reward-to-go is deterministic and equal to $r(x)$. Thus, we only need to estimate J and V for states in which the policy decides to hold. We focus on 'in-the-money' options, in which K is equal to the initial price x_0, and set $T = 20$.

A policy π was obtained using the LSPI algorithm (Lagoudakis and Parr, 2003; Li et al., 2009) with 2-dimensional (for x and t) radial basis function (RBF) features, as detailed in Tamar et al. (2014). It is well-known (Duffie, 2010), and intuitive, that the optimal policy (in terms of expected return) for the put option has a threshold structure—the policy executes if the price is below some boundary $\bar{x}_t$, and holds otherwise. It is also known, that $\bar{x}_t$ is monotonically increasing in t. Our policy π has such a structure as well. We emphasize, however, that the specific method of generating the policy π is not the focus of this work, as we are only interested in *evaluating* π. Thus, any policy generation method could have been used, and LSPI was chosen for convenience. In the following, we evaluate the value functions J and V for π.

6.2 Results

We now present our policy evaluation results for the put option domain. MATLAB® code for reproducing these results is available on the web-site `https://sites.google.com/site/variancetdcode/`.

We first calculate the 'true' value function J and standard deviation of reward-to-go $\sqrt{V}$, as shown in Figure 3. These plots were obtained using Monte Carlo (MC), by taking the empirical average and standard deviation of the reward of $10,000$ trajectories starting from 323 equally spaced points in the state space for which the policy π decides to hold, a total of $N = 3,230,000$ trajectories. To our knowledge, an MC approach is the current state-of-the-art for obtaining an estimate of V.

Note the exercise boundary $\bar{x}_t$, emphasized with a dashed line in the value function plot. For x smaller than $\bar{x}_t$, the policy decides to exercise, therefore the value is linear in x and the variance is zero. Also note the discontinuous ridges on the J and $\sqrt{V}$ plots. These ridges are due to the discrete transition model, and occur when a transition to the next state (or the state following the next state) crosses the exercise boundary. To the risk-sensitive decision maker, these ridges are important, as they separate states with roughly the same expected return but with very different variance.

In Figure 4 we show the RMS error of the approximation $\sqrt{\tilde{V}}$ (compared to the 'true' $\sqrt{V}$) computed using the LSTD(0) algorithm of Section 4, for different budgets of sample trajectories N. We tested two popular feature sets: RBF features with 77 equally spaced centers, and tile features with 600 uniform non-overlapping tiles. In both cases the same features were used for both J and M. The sample trajectories were simulated independently, starting from uniformly distributed initial states. We compare our results to MC estimates obtained with the same trajectories.

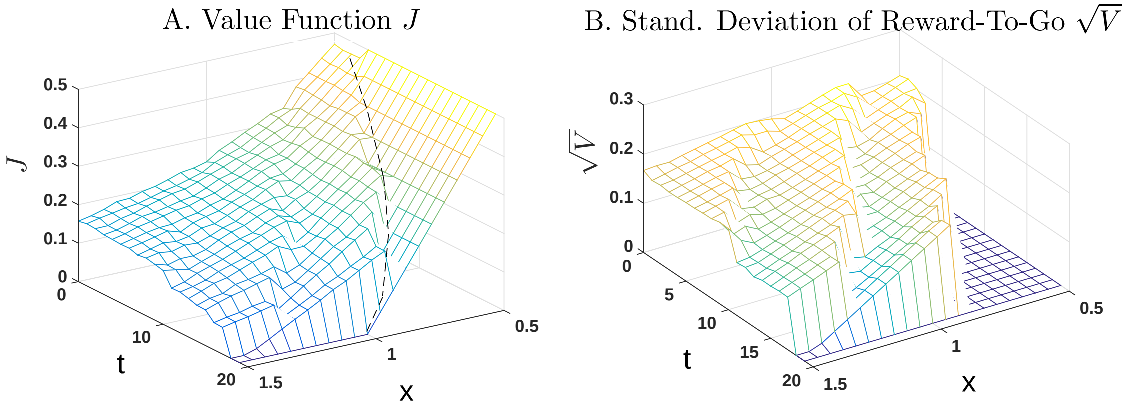

Figure 3: True value function J and standard deviation of the reward-to-go $\sqrt{V}$.

As can be seen, by exploiting relations *between* states and using the generalization capabilities of the function approximation, LSTD is able to fully exploit the data, and performs significantly better than MC for relatively small sample sizes. On the other hand, LSTD is limited by the expressiveness of its function approximation, and its error is therefore bounded.

Note that for $N \leq 323$ the MC estimate is meaningless, as the empirical standard deviation cannot be calculated from only one sample. LSTD however, is able to provide a reasonable result. Also note that the LSTD estimate is defined over the whole state-space, whereas the MC estimate is only defined for the discrete set of evaluation points.

To further appreciate the advantage of function approximation, we provide a visual comparison of the approximated standard deviation of reward-to-go $\sqrt{\tilde{V}}$. In Figure 5 we plot $\sqrt{\tilde{V}}$ obtained using a budget of $N = 2000$ sample trajectories starting from uniformly distributed states. In the left plot, we show the results of LSTD(λ) with RBF features (with 77 equally spaced centers in x and t). The variance in states where the policy decides to execute was set to zero manually, as there is no need to estimate it. In comparison, on the right plot we present the results of a Monte Carlo algorithm, with the same amount of data trajectories $N = 2000$. Clearly, LSTD(λ) makes better use of the limited data, with a plot that is much more similar to the true standard deviation (Figure 3; right). More importantly, the relevant structure in $\sqrt{V}$ outlined above is clear in the LSTD(λ) result (up to a smoothness limitation of the RBFs), yielding important information for the decision maker.

In Figure 6 we consider the LSTD(λ) algorithm with the tile features discussed above, and explore the effect of λ on the RMS error in $\sqrt{\tilde{V}}$. As in regular LSTD, λ can be seen to trade off estimation bias and variance (Bertsekas, 2012). In addition, we compare LSTD(λ) to the direct least squares algorithm of Section 4.4. For the case of the value function J, it is well-known (Bertsekas, 2012) that the direct approximation is equivalent to the limit $\lambda \to 1$. Our results suggest that a similar relation holds for the variance V as well. Furthermore, these results highlight the superior performance of the TD approach in the small sample size regime.

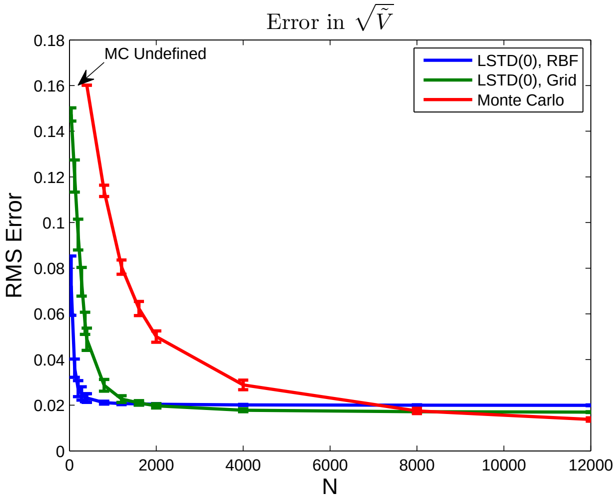

Figure 4: LSTD(0) vs. Monte Carlo. The RMS error of $\sqrt{\tilde{V}}$ on a set of evaluation points (see text) is shown vs. the budget of sample trajectories N, for LSTD(0) with two types of features and for Monte Carlo. Standard deviation error-bars from 20 runs are shown.

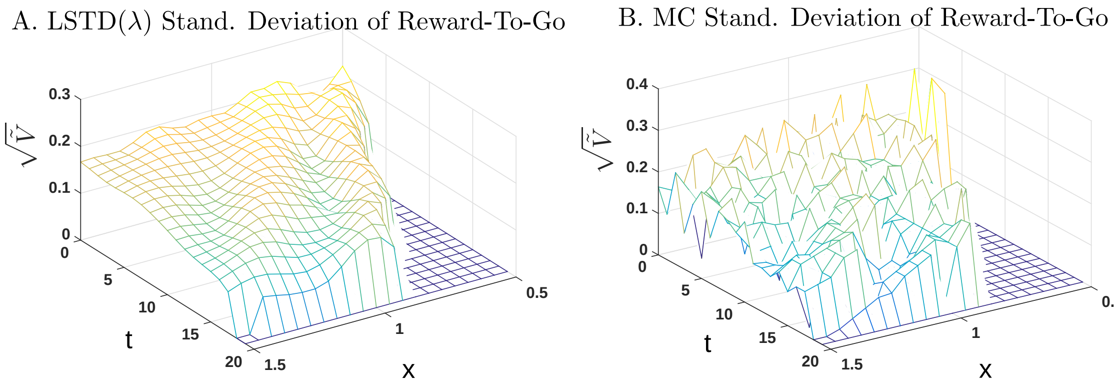

Figure 5: Approximate standard deviation of the reward-to-go $\sqrt{\tilde{V}}$. Left plot was obtained by LSTD(λ) with RBF features, using 2000 trajectories and $\lambda = 0.3$. Right plot was obtained using Monte Carlo, also with 2000 trajectories.

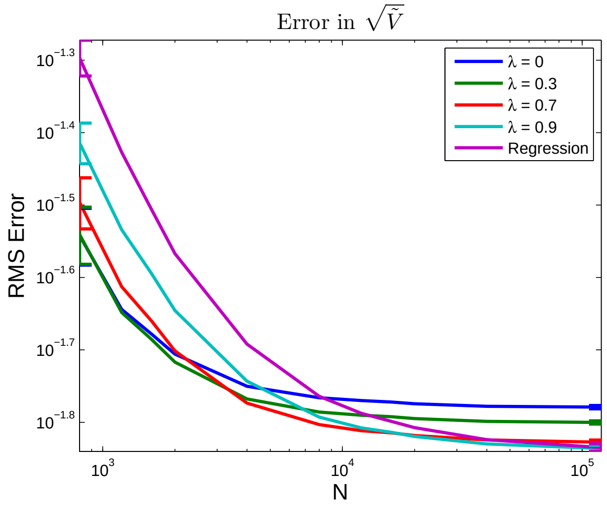

Figure 6: LSTD(λ) Results. The RMS error of $\sqrt{\tilde{V}}$ on a set of evaluation points (see text) is shown vs. the budget of sample trajectories N for different λ, and also for the least squares regression algorithm. For clarity, a log scale is used, and the error-bars (standard deviation from 20 runs) are shown only for marginal points.

7. Conclusion

We presented an extension of the TD framework for policy evaluation in MDPs with respect to the variance of the reward-to-go. Our framework deals with the curse of dimensionality by using function approximation, and uses a bootstrapping technique, based on an extension of the Bellman equation to the second moment, to achieve good performance even for a small sample size. We presented both formal guarantees and empirical evidence that this approach is useful in problems with a large state space, and limited sample budget.

A natural extension of this work is to consider higher moments, and statistical properties such as skewness and kurtosis of the reward-to-go. An extension of Bellman's equation to higher moments was proposed by Sobel (1982), and it may be used to derive TD equations similarly to the work presented here. This may also be useful for optimizing the expectation of a general function f of the accumulated reward $\mathbb{E}\left[f\left(B\right)\right]$, by looking at the first few terms in the Taylor expansion of f. It would be interesting to see whether a TD approach may be developed for other risk measures such as the value at risk or semi-deviation.

Another interesting direction is to use the variance of the reward-to-go to *guide feature selection*, or feature modification. For example, consider tile features. A large variance-to-go for states that belong to a particular tile may indicate that the value function in that tile varies greatly, and therefore it may be beneficial to split the tile into smaller segments. Of course, another explanation for the variance may be the inherent stochasticity of the system. Thus, a thoughtful feature-selection method should take that also into account. In a related topic, the variance of the reward-to-go may also be used to *guide exploration*, since intuitively, states with higher variance should be allocated more exploration resources, to potentially decrease the variance, if possible.

We conclude with a discussion on policy optimization with respect to a mean-variance tradeoff. While a naive variance-penalized policy iteration algorithm may be easily conceived, its usefulness should be questioned, as it was shown to be problematic for the standard deviation adjusted reward (Sobel, 1982) and the variance constrained reward (Mannor and Tsitsiklis, 2013). An alternative approach is to pursue *locally* optimal policies by using a gradient based method. Tamar et al. (2012) proposed policy gradient algorithms for a class of variance related criteria, and showed their convergence to local optima. These algorithms may be extended to use the variance function in an actor-critic type scheme (Sato et al., 2001), and recent work has extended these ideas to large-scale MDPs by employing function approximation, and using the TD policy evaluation algorithms presented here (Tamar and Mannor, 2013; Prashanth and Ghavamzadeh, 2013).

Acknowledgments

We would like to thank our anonymous reviewers for helpful comments and suggestions. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Program (FP/2007-2013) / ERC Grant Agreement n. 306638. Aviv Tamar is also partially funded by the Viterbi Scholarship, Technion.

Appendix A. Proof of Proposition 2

Proof The equation for $J(x)$ is well-known, and its proof is given here only for completeness. Choose $x \in X$. Then,

$$\begin{aligned}
J(x) &= \mathbb{E}\left[B | x_0 = x\right] \\
&= \mathbb{E}\left[\left.\sum_{k=0}^{\tau-1} \gamma^k r(x_k)\right| x_0 = x\right] \\
&= r(x) + \mathbb{E}\left[\left.\sum_{k=1}^{\tau-1} \gamma^k r(x_k)\right| x_0 = x\right] \\
&= r(x) + \gamma \mathbb{E}\left[\mathbb{E}\left[\left.\sum_{k=1}^{\tau-1} \gamma^{k-1} r(x_k)\right| x_0 = x, x_1 = x'\right]\right] \\
&= r(x) + \gamma \sum_{x' \in X} P(x'|x) J(x'),
\end{aligned}$$

where we excluded the terminal state from the last sum since reaching it ends the trajectory.

Similarly,

$$\begin{aligned}
M(x) &= \mathbb{E}\left[B^2 | x_0 = x\right] \\
&= \mathbb{E}\left[\left(\sum_{k=0}^{\tau-1} \gamma^k r(x_k)\right)^2 \middle| x_0 = x\right] \\
&= \mathbb{E}\left[\left(r(x_0) + \sum_{k=1}^{\tau-1} \gamma^k r(x_k)\right)^2 \middle| x_0 = x\right] \\
&= r(x)^2 + 2r(x)\mathbb{E}\left[\sum_{k=1}^{\tau-1} \gamma^k r(x_k) \middle| x_0 = x\right] + \mathbb{E}\left[\left(\sum_{k=1}^{\tau-1} \gamma^k r(x_k)\right)^2 \middle| x_0 = x\right] \\
&= r(x)^2 + 2\gamma r(x) \sum_{x' \in X} P(x'|x) J(x') + \gamma^2 \sum_{x' \in X} P(x'|x) M(x').
\end{aligned}$$

The uniqueness of the value function J for a proper policy is well known, cf. Proposition 3.2.1 in Bertsekas (2012). The uniqueness of M follows by observing that in the equation for M, M may be seen as the value function of an MDP with the same transitions but with reward $r(x)^2 + 2\gamma r(x) \sum_{x' \in X} P(x'|x) J(x')$. Since only the rewards change, the policy remains proper and Proposition 3.2.1 in Bertsekas (2012) applies. ■

Appendix B. Proof of Theorem 10

Proof Let $\phi_1(x)$, $\phi_2(x)$ be some vector functions of the state. We claim that

$$\mathbb{E}\left[\sum_{t=0}^{\tau-1} \phi_1(x_t)\phi_2(x_t)^\top\right] = \sum_x q(x)\phi_1(x)\phi_2(x)^\top \equiv \Phi_1^\top Q \Phi_2, \tag{22}$$

where Φ_1 and Φ_2 are matrices with rows $\phi_1(x)$ and $\phi_2(x)$, respectively. To see this, let $\mathbb{1}(\cdot)$ denote the indicator function and write

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{\tau-1} \phi_1(x_t)\phi_2(x_t)^\top\right] &= \mathbb{E}\left[\sum_{t=0}^{\tau-1} \sum_x \phi_1(x)\phi_2(x)^\top \mathbb{1}(x_t = x)\right] \\
&= \mathbb{E}\left[\sum_x \phi_1(x)\phi_2(x)^\top \sum_{t=0}^{\tau-1} \mathbb{1}(x_t = x)\right] \\
&= \sum_x \phi_1(x)\phi_2(x)^\top \mathbb{E}\left[\sum_{t=0}^{\tau-1} \mathbb{1}(x_t = x)\right].
\end{aligned}$$

Now, note that the last term on the right hand side is an expectation (over all possible trajectories) of the number of visits to a state x until reaching the terminal state, which is exactly $q(x)$ since

$$\begin{aligned} q(x) &= \sum_{t=0}^{\infty} P(x_t = x) \\ &= \sum_{t=0}^{\infty} \mathbb{E}[\mathbb{1}(x_t = x)] \\ &= \mathbb{E}\left[\sum_{t=0}^{\infty} \mathbb{1}(x_t = x)\right] \\ &= \mathbb{E}\left[\sum_{t=0}^{\tau-1} \mathbb{1}(x_t = x)\right], \end{aligned}$$

where the third equality is by the dominated convergence theorem (Grimmett and Stirzaker, 2001, Sec. 5.6), and last equality follows from the absorbing property of the terminal state. Similarly, we have

$$\mathbb{E}\left[\sum_{t=0}^{\tau-1} \phi_1(x_t)\phi_2(x_{t+1})^\top\right] = \sum_x \sum_{x'} q(x)P(x'|x)\phi_1(x)\phi_2(x')^\top \equiv \Phi_1^\top Q P \Phi_2, \tag{23}$$

since

$$\begin{aligned} \mathbb{E}\left[\sum_{t=0}^{\tau-1} \phi_1(x_t)\phi_2(x_{t+1})^\top\right] &= \mathbb{E}\left[\sum_{t=0}^{\tau-1} \sum_x \sum_{x'} \phi_1(x)\phi_2(x')^\top \mathbb{1}(x_t = x, x_{t+1} = x')\right] \\ &= \mathbb{E}\left[\sum_x \sum_{x'} \phi_1(x)\phi_2(x')^\top \sum_{t=0}^{\tau-1} \mathbb{1}(x_t = x, x_{t+1} = x')\right] \\ &= \sum_x \sum_{x'} \phi_1(x)\phi_2(x')^\top \mathbb{E}\left[\sum_{t=0}^{\tau-1} \mathbb{1}(x_t = x, x_{t+1} = x')\right], \end{aligned}$$

and

$$\begin{aligned} q(x)P(x'|x) &= \sum_{t=0}^{\infty} P(x_t = x)P(x'|x) \\ &= \sum_{t=0}^{\infty} P(x_t = x, x_{t+1} = x') \\ &= \sum_{t=0}^{\infty} \mathbb{E}[\mathbb{1}(x_t = x, x_{t+1} = x')] \\ &= \mathbb{E}\left[\sum_{t=0}^{\infty} \mathbb{1}(x_t = x, x_{t+1} = x')\right] \\ &= \mathbb{E}\left[\sum_{t=0}^{\tau-1} \mathbb{1}(x_t = x, x_{t+1} = x')\right]. \end{aligned}$$

Since trajectories between visits to the recurrent state are statistically independent, the law of large numbers together with the expressions in (22) and (23) suggest that the approximate expressions in (13) converge to their expected values with probability 1, therefore we have

$$\begin{aligned} A_N &\rightarrow A, \quad b_N \rightarrow b, \\ C_N &\rightarrow C, \quad d_N \rightarrow D, \end{aligned}$$

and

$$\begin{aligned} \hat{w}^*_{J;N} &= A_N^{-1} b_N \rightarrow A^{-1} b = w^*_J, \\ \hat{w}^*_{M;N} &= C_N^{-1} d_N \rightarrow C^{-1} d = w^*_M. \end{aligned}$$

■

Appendix C. Proof of Theorem 17

To show the convergence of the simulation-based version of (21) to a solution of (20), we need to bound the effect of simulation noise on the fixed point of (21). The difficulty, is that simulation noise affects both the terms in the update, C and d, and terms in the projection step—the set $\hat{W}_M$ onto which we project. In addition, the noise in C and d effectively adds noise to the weights q of the norm in (20), which should also be bounded.

We begin by proving several intermediate results. The first concerns the continuity of fixed points of contraction operators.

Lemma 18 *Let T_1 be a γ-contraction in the q_1 norm, and T_2 be a γ-contraction in the q_2 norm. Assume that there exists some δ' such that*

$$\|T_1 x - T_2 x\|_{q_1} \leq \delta + \delta' \|x\|_{q_1}, \quad \forall x.$$

Let x_1^* and x_2^* denote the fixed points of T_1 and T_2, respectively. Then the following holds:

$$\|x_2^* - x_1^*\|_{q_1} \leq \frac{\delta + \delta' \|x_2^*\|_{q_1}}{1-\gamma}.$$

Proof We have

$$\begin{aligned} \|x_2^* - x_1^*\|_{q_1} &= \|T_2 x_2^* - x_1^*\|_{q_1} \\ &= \|T_2 x_2^* + T_1 x_2^* - T_1 x_2^* - x_1^*\|_{q_1} \\ &\leq \|T_1 x_2^* - x_1^*\|_{q_1} + \|T_2 x_2^* - T_1 x_2^*\|_{q_1} \\ &\leq \|T_1 x_2^* - T_1 x_1^*\|_{q_1} + \delta + \delta' \|x_2^*\|_{q_1} \\ &\leq \gamma \|x_2^* - x_1^*\|_{q_1} + \delta + \delta' \|x_2^*\|_{q_1}. \end{aligned}$$

Rearranging, gives:

$$\|x_2^* - x_1^*\|_{q_1} \leq \frac{\delta + \delta' \|x_2^*\|_{q_1}}{1-\gamma}.$$

■

The following results concerns the sensitivity of weighted Euclidean-norm projections.

Lemma 19 *Let $\|\cdot\|_q$ and $\|\cdot\|_{q'}$ denote weighted Euclidean-norms on $\mathbb{R}^n$ with weights $q > 0$ and $q' > 0$, respectively. Let Π and Π' denote projections onto a closed and convex set $S \subset \mathbb{R}^n$, w.r.t. the norms $\|\cdot\|_q$ and $\|\cdot\|_{q'}$, respectively. For any $x \in \mathbb{R}^n$ we have:*

$$\|\Pi x - \Pi' x\|_q^2 \leq 2\|q - q'\|_\infty \left(\|\Pi x - x\|_2^2 + \|\Pi' x - x\|_2^2\right).$$

Proof If $x \in S$ the result is trivial. We assume in the following $x \notin S$. Let $Q = diag(q)$ and $Q' = diag(q')$. For any $x, y \in \mathbb{R}^n$ we have

$$\begin{aligned}\left|\|x-y\|_q^2 - \|x-y\|_{q'}^2\right| &= \left|(x-y)^\top Q(x-y) - (x-y)^\top Q'(x-y)\right| \\ &= \left|(x-y)^\top (Q-Q')(x-y)\right| \\ &\leq \sum_{i=1}^n |q_i - q_i'|(x_i - y_i)^2 \\ &\leq \|q-q'\|_\infty \|x-y\|_2^2.\end{aligned} \tag{24}$$

Therefore, we have that

$$\|\Pi' x - x\|_{q'}^2 \geq \|\Pi' x - x\|_q^2 - \|q - q'\|_\infty \|\Pi' x - x\|_2^2. \tag{25}$$

Now, let H denote the hyper-plane that is orthogonal to the projection error $\Pi x - x$, and passes through Πx:

$$H \doteq \left\{ y \in \mathbb{R}^n : (y - \Pi x)^\top Q(x - \Pi x) = 0 \right\},$$

and let L denote a line that passes through x and $\Pi' x$:

$$L \doteq \left\{ y \in \mathbb{R}^n : y = x + z(\Pi' x - x), \quad z \in \mathbb{R} \right\}.$$

By properties of the projection Πx (Hiriart-Urruty and Lemaréchal, 2013) we have $(y - \Pi x)^\top Q(x - \Pi x) \leq 0 \quad \forall y \in S$. Since $(x - \Pi x)^\top Q(x - \Pi x) > 0$, it follows that H separates x from S. Since $\Pi' x \in S$, H also separates x from $\Pi' x$. Let p^* denote the intersection of L and H. By the previous arguments, p^* exists, and

$$\Pi' x - x = \alpha(p^* - x), \tag{26}$$

with $\alpha \geq 1$. Now, we have

$$\begin{aligned}\|\Pi' x - x\|_q^2 &= \|\alpha(p^* - x)\|_q^2 \\ &= \alpha^2 \|p^* - x\|_q^2 \\ &= \alpha^2 \|p^* - \Pi x\|_q^2 + \alpha^2 \|\Pi x - x\|_q^2 \\ &\geq \alpha^2 \|p^* - \Pi x\|_q^2 + \|\Pi x - x\|_{q'}^2 - \|q - q'\|_\infty \|\Pi x - x\|_2^2,\end{aligned}$$

where the last equality is by the Pythagorean theorem, which holds due to the orthogonality of H to the error $\Pi x - x$, and the inequality is since $\alpha \geq 1$, and (24). Plugging in (25), we obtain:

$$\|\Pi' x - x\|_{q'}^2 - \|\Pi x - x\|_{q'}^2 \geq \alpha^2 \|p^* - \Pi x\|_q^2 - \|q - q'\|_\infty \left(\|\Pi x - x\|_2^2 + \|\Pi' x - x\|_2^2\right). \tag{27}$$

However, by definition of the projection $\Pi' x$, we must have $\|\Pi' x - x\|_{q'}^2 - \|\Pi x - x\|_{q'}^2 \leq 0$, therefore rearranging (27) leads to:

$$\alpha^2 \|p^* - \Pi x\|_q^2 \leq \|q - q'\|_\infty \left(\|\Pi x - x\|_2^2 + \|\Pi' x - x\|_2^2 \right). \tag{28}$$

Now, let H' denote a parallel hyper-plane to H that passes through $\Pi' x$:

$$H' \doteq \left\{ y \in \mathbb{R}^n : (y - \Pi' x)^\top Q(x - \Pi x) = 0 \right\}.$$

Also, let L' denote the line between x and Πx:

$$L' \doteq \{ y \in \mathbb{R}^n : y = x + z(\Pi x - x), \quad z \in \mathbb{R} \}.$$

By definition, H' is orthogonal to L'; denote by p^{**} their intersection. By triangle similarity (the triangles $\{x, \Pi x, p^*\}$ and $\{x, p^{**}, \Pi' x\}$), and (26) we have

$$\frac{\|\Pi' x - p^{**}\|_q^2}{\|\Pi x - p^*\|_q^2} = \frac{\|\Pi' x - x\|_q^2}{\|p^* - x\|_q^2} = \alpha^2, \tag{29}$$

therefore, using (28)

$$\|\Pi' x - p^{**}\|_q^2 = \alpha^2 \|p^* - \Pi x\|_q^2 \leq \|q - q'\|_\infty \left(\|\Pi x - x\|_2^2 + \|\Pi' x - x\|_2^2 \right). \tag{30}$$

From the Pythagorean theorem (by the orthogonality of H' to L') we have:

$$\|\Pi' x - \Pi x\|_q^2 = \|\Pi x - p^{**}\|_q^2 + \|\Pi' x - p^{**}\|_q^2, \tag{31}$$

and

$$\|\Pi' x - x\|_q^2 = \|\Pi' x - p^{**}\|_q^2 + \|p^{**} - x\|_q^2,$$

and from the last equation we also have

$$\|\Pi' x - x\|_q^2 \geq \|p^{**} - x\|_q^2.$$

Now, from the last inequality:

$$\begin{aligned} \|\Pi' x - x\|_q^2 &\geq \|p^{**} - x\|_q^2 \\ &\geq \|p^{**} - \Pi x\|_q^2 + \|\Pi x - x\|_q^2 \\ &\geq \|p^{**} - \Pi x\|_q^2 + \|\Pi x - x\|_{q'}^2 - \|q - q'\|_\infty \|\Pi x - x\|_2^2, \end{aligned}$$

where the second inequality is since x, Πx, and p^{**} are on L', therefore $\|p^{**} - x\|_q = \|p^{**} - \Pi x\|_q + \|\Pi x - x\|_q$, and the last inequality is by (24). Proceeding similarly as in (27), we plug in (25) to obtain:

$$\|\Pi' x - x\|_{q'}^2 - \|\Pi x - x\|_{q'}^2 \geq \|p^{**} - \Pi x\|_q^2 - \|q - q'\|_\infty \left(\|\Pi x - x\|_2^2 + \|\Pi' x - x\|_2^2 \right), \tag{32}$$

and similarly to (28), by definition of the projection $\Pi' x$, we must have $\|\Pi' x - x\|_{q'}^2 - \|\Pi x - x\|_{q'}^2 \leq 0$, therefore rearranging (32) leads to:

$$\|p^{**} - \Pi x\|_q^2 \leq \|q - q'\|_\infty \left(\|\Pi x - x\|_2^2 + \|\Pi' x - x\|_2^2 \right). \tag{33}$$

Finally, plugging in (30), and (33) in (31) we obtain

$$\|\Pi' x - \Pi x\|_q^2 \leq 2\|q - q'\|_\infty \left(\|\Pi x - x\|_2^2 + \|\Pi' x - x\|_2^2\right).$$

■

We now proceed with the proof of Theorem 17. To simplify the presentation, we break the proof into several parts.
In part 1, we show show that the sampled version of algorithm (21) with N samples corresponds to solving (20) with P_N, a sampled version of the transition matrix, replacing P.
In part 2, we show that for each N, algorithm (21) would converge (in k) by Lemma 16 to a fixed point of the *sampled* projected equation.
In part 3, we show that the solution of the sampled projected equation converges (in N) to the the solution of the original projected equation. We do this by showing a continuity of the solution w.r.t. P and its derived quantities, q and w_J^*, from which convergence then follows by the law of large numbers.
In part 4, we collect our convergence results in k and N and complete the proof.

C.1 A Sampled Version of Eq. (21)

Let $S^+ = \{\Phi_M w | w \in \mathbb{R}^m, Hw \leq g\}$ denote the set onto which we project in the modified projection (20), and let Π_q^+ denote a projection onto S^+ w.r.t. the q-weighted Euclidean norm. Note that S^+ is a convex set, therefore Π_q^+ is a non-expansion in the $\|\cdot\|_q$ norm (Hiriart-Urruty and Lemaréchal, 2013). Furthermore, we can write Eq. (20) as follows:

$$w_M^+ = \Pi_q^+ T^+ w_M^+,$$

where $T^+(w) = Rr + 2\gamma R P \Phi_J w_J^* + \gamma^2 P \Phi_M w$.

After we have observed N trajectories, let P_N denote the corresponding empirical transition matrix, given by:

$$P_N(x'|x) = \left(\frac{1}{\sum_{k=1}^N \tau_k}\right) \sum_{k=1}^N \sum_{t=0}^{\tau_k - 1} \mathbb{1}(x_t^k = x, x_{t+1}^k = x'),$$

and let $\zeta_{0;N}$ denote the empirical initial state distribution, i.e.,

$$\zeta_{0;N}(x) = \left(\frac{1}{N}\right) \sum_{k=1}^N \mathbb{1}(x_0^k = x).$$

Also, let q_N denote the state occupancy probabilities in an MDP with P and ζ_0 replaced by P_N and $\zeta_{0;N}$ (cf. the definition of q in Section 3). For large enough N, q_N satisfies Assumption 4.

Let $\hat{w}_J^* = A_N^{-1} b_N$, with A_N and b_N defined in (13); for large enough N, $\hat{w}_J^*$ is well defined (Boyan, 2002).

Furthermore, let g_N denote a vector with elements $-(\phi_J(x_i)^\top \hat{w}_J^*)^2$.

We define the set $S_N^+ = \{\Phi_M w | w \in \mathbb{R}^m, Hw \leq g_N\}$, and denote by $\Pi_{q_N}^+$ a projection onto S_N^+ w.r.t. the q_N-weighted Euclidean norm. We also define the operator

$$T_N^+(w) \doteq Rr + 2\gamma R P_N \Phi_J \hat{w}_J^* + \gamma^2 P_N \Phi_M w,$$

which is the sampled version of T^+. Note that T_N^+ is a γ^2-contraction.

Consider now the following projected fixed point equation:

$$w_{M;N}^+ = \Pi_{q_N}^+ T_N^+ w_{M;N}^+, \tag{34}$$

and the iterative procedure

$$w_{k+1;N} = \Pi_{\Xi, \hat{W}_{M;N}}[w_{k;N} - \eta \Xi^{-1}(C_N w_{k;N} - d_N)], \tag{35}$$

where C_N and d_N are defined in (13), Ξ is an arbitrary positive definite matrix, $\eta \in \mathbb{R}$ is a positive step size, and $\Pi_{\Xi, \hat{W}_{M;N}}$ denotes a projection onto the convex set $\hat{W}_{M;N} = \{w | Hw \leq g_N\}$ with respect to the Ξ weighted Euclidean norm.

C.2 Convergence in k

By definition, the sampled C_N, d_N, q_N and $\hat{w}_J^*$ correspond to their non-sampled counterparts C, d, q and w_J^*, respectively, on an MDP with the empirical probabilities P_N and $\zeta_{0;N}$ replacing P and ζ_0. As a result, applying Lemma 16 to Eq. 35, we have that $w_{k;N}$ converges to $w_{M;N}^+$. Therefore, for each N and $\delta > 0$ there exists some $k(N, \delta)$ such that for all $k > k(N, \delta)$

$$\|w_{k;N} - w_{M;N}^+\| \leq \delta. \tag{36}$$

C.3 Convergence in N

We will now show that as $N \to \infty$, $w_{M;N}^+ \to w_M^+$.

Let $\epsilon, \tilde{\epsilon} > 0$. We claim that w.p. 1, there exists $N(\epsilon, \tilde{\epsilon})$, such that for all $N > N(\epsilon, \tilde{\epsilon})$ we have

$$\|\Pi_q^+ T^+ w - \Pi_{q_N}^+ T_N^+ w\|_q \leq \epsilon + \tilde{\epsilon}\|w\|_q, \quad \forall w. \tag{37}$$

We now prove (37). First, we have:

$$\begin{aligned} \|\Pi_q^+ T^+ w - \Pi_{q_N}^+ T_N^+ w\|_q &= \|\Pi_q^+ T^+ w + \Pi_q^+ T_N^+ w - \Pi_q^+ T_N^+ w - \Pi_{q_N}^+ T_N^+ w\|_q \\ &\leq \underbrace{\|\Pi_q^+ T^+ w - \Pi_q^+ T_N^+ w\|_q}_{A} + \underbrace{\|\Pi_q^+ T_N^+ w - \Pi_{q_N}^+ T_N^+ w\|_q}_{B}. \end{aligned} \tag{38}$$

C.3.1 A Bound on (A):

We have:

$$\begin{aligned} \|\Pi_q^+ T^+ w - \Pi_q^+ T_N^+ w\|_q &\leq \|T^+ w - T_N^+ w\|_q \\ &= \|2\gamma R P \Phi_J w_J^* - 2\gamma R P_N \Phi_J \hat{w}_J^* + \gamma^2 (P - P_N)\Phi_M w\|_q \\ &\leq \|2\gamma R P \Phi_J w_J^* - 2\gamma R P_N \Phi_J \hat{w}_J^*\|_q + \|\gamma^2 (P - P_N)\Phi_M w\|_q \\ &\triangleq \eta_1(N) + \|\gamma^2 (P - P_N)\Phi_M w\|_q \\ &\leq \eta_1(N) + \eta_2(N)\|w\|_q, \end{aligned} \tag{39}$$

where the first inequality is by the non-expansion property of the projection, and the third inequality is by defining $\eta_2(N)$ as the $\|\cdot\|_q$ induced matrix norm of $\gamma^2(P - P_N)\Phi_M$ (Horn and Johnson, 2012, Definition 5.6.1).

C.3.2 A Bound on (B):

Denote by $\hat{\Pi}_q^+$ a projection onto S_N^+ w.r.t. the q-weighted Euclidean norm. We have

$$\begin{aligned}\|\Pi_q^+ T_N^+ w - \Pi_{q_N}^+ T_N^+ w\|_q &= \|\Pi_q^+ T_N^+ w + \hat{\Pi}_q^+ T_N^+ w - \hat{\Pi}_q^+ T_N^+ w - \Pi_{q_N}^+ T_N^+ w\|_q \\ &\le \underbrace{\|\Pi_q^+ T_N^+ w - \hat{\Pi}_q^+ T_N^+ w\|_q}_{B_1} + \underbrace{\|\hat{\Pi}_q^+ T_N^+ w - \Pi_{q_N}^+ T_N^+ w\|_q}_{B_2}.\end{aligned}$$

C.3.3 A Bound on (B_1):

We bound B_1 using a result of Yen (1995), which gives a general Lipschitz bound for perturbations of projections onto convex polyhedra (S_N^+ by definition is a convex polyhedron). By theorem 2.1 of Yen (1995), for all w, there exists a constant K such that

$$\|\Pi_q^+ T_N^+ w - \hat{\Pi}_q^+ T_N^+ w\|_q \le K\|g - g_N\|_2 \triangleq \eta_3(N). \tag{40}$$

C.3.4 A Bound on (B_2):

We bound B_2 using Lemma 19, which yields:

$$\|\hat{\Pi}_q^+ T_N^+ w - \Pi_{q_N}^+ T_N^+ w\|_q^2 \le 2\|q - q_N\|_\infty \left(\|\hat{\Pi}_q^+ T_N^+ w - T_N^+ w\|_2^2 + \|\Pi_{q_N}^+ T_N^+ w - T_N^+ w\|_2^2\right).$$

By norm equivalence on finite-dimensional spaces, there exists λ such that $\|x\|_2 \le \lambda\|x\|_q$ and $\|x\|_2 \le \lambda\|x\|_{q_N}$ for all x. Therefore

$$\|\hat{\Pi}_q^+ T_N^+ w - \Pi_{q_N}^+ T_N^+ w\|_q^2 \le 2\|q - q_N\|_\infty \lambda^2 \left(\|\hat{\Pi}_q^+ T_N^+ w - T_N^+ w\|_q^2 + \|\Pi_{q_N}^+ T_N^+ w - T_N^+ w\|_{q_N}^2\right).$$

For any $\hat{s} \in S_N^+$ we now have, by definition of the projections $\hat{\Pi}_q^+$ and $\Pi_{q_N}^+$:

$$\|\hat{\Pi}_q^+ T_N^+ w - \Pi_{q_N}^+ T_N^+ w\|_q^2 \le 2\|q - q_N\|_\infty \lambda^2 \left(\|\hat{s} - T_N^+ w\|_q^2 + \|\hat{s} - T_N^+ w\|_{q_N}^2\right).$$

As before, by norm equivalence on finite-dimensional spaces, there exists $\tilde{\lambda}$ such that $\|x\|_{q_N} \le \lambda\|x\|_q$ for all x, therefore

$$\|\hat{\Pi}_q^+ T_N^+ w - \Pi_{q_N}^+ T_N^+ w\|_q^2 \le 2\|q - q_N\|_\infty \lambda^2 (1 + \tilde{\lambda}^2)\|\hat{s} - T_N^+ w\|_q^2,$$

and setting $\bar{\lambda} = \sqrt{\lambda^2(1 + \tilde{\lambda}^2)}$ we have

$$\begin{aligned}\|\hat{\Pi}_q^+ T_N^+ w - \Pi_{q_N}^+ T_N^+ w\|_q &\le \sqrt{2\|q - q_N\|_\infty}\bar{\lambda}\|\hat{s} - T_N^+ w\|_q \\ &\le \sqrt{2\|q - q_N\|_\infty}\bar{\lambda}\left(\|\hat{s}\|_q + \|T_N^+ w\|_q\right) \\ &\le \sqrt{2\|q - q_N\|_\infty}\bar{\lambda}\left(\|\hat{s}\|_q + C + \|w\|_q\right),\end{aligned}$$

where the constant C exists since T_N^+ is linear and a contraction. Therefore, setting $\eta_4(N) = \sqrt{2\|q - q_N\|_\infty}\bar{\lambda}\left(\|\hat{s}\|_q + C\right)$ and $\eta_5(N) = \sqrt{2\|q - q_N\|_\infty}\bar{\lambda}$ we have

$$\|\hat{\Pi}_q^+ T_N^+ w - \Pi_{q_N}^+ T_N^+ w\|_q \le \eta_4(N) + \eta_5(N)\|w\|_q. \tag{41}$$

C.3.5 Proof of (37):

We now return to (38), where, using (39), (40), and (41) we have

$$\|\Pi_q^+ T^+ w - \Pi_{q_N}^+ T_N^+ w\|_q \leq \eta_1(N) + \eta_2(N)\|w\|_q + \eta_3(N) + \eta_4(N) + \eta_5(N)\|w\|_q.$$

The uniform convergence of empirical distributions (Van der Vaart, 2000, Theorem 19.1) guarantees that P_N and $\zeta_{0;N}$ uniformly converge to P and ζ_0 w.p. 1, respectively, and therefore $q_N \to q$ and $\hat{w}_J^* \to w_J^*$ w.p. 1. Therefore, for every $\epsilon, \tilde{\epsilon} > 0$, w.p. 1 there is some $N(\epsilon, \tilde{\epsilon})$ such that for $N > N(\epsilon, \tilde{\epsilon})$ we have $\eta_1(N)+\eta_3(N)+\eta_4(N) \leq \epsilon$, and $\eta_2(N)+\eta_5(N) \leq \tilde{\epsilon}$, therefore Eq. (37) holds.

Using Lemma 18 and Eq. (37) we have that for $N > N(\epsilon, \tilde{\epsilon})$

$$\|w_{M;N}^+ - w_M^+\|_q \leq \frac{\epsilon + \tilde{\epsilon}\|w_M^+\|_q}{1-\gamma}. \tag{42}$$

C.4 Convergence in k and N

Finally, using (42) and (36) we have that for any $\bar{\epsilon} > 0$, w.p. 1 there is a $N(\bar{\epsilon})$ such that for any $N > N(\bar{\epsilon})$ there is a $k(N, \bar{\epsilon})$, such that for all $k > k(N, \bar{\epsilon})$

$$\|w_{k;N} - w_M^+\| \leq \bar{\epsilon}.$$

References

D. P. Bertsekas. Temporal difference methods for general projected equations. *IEEE Transactions on Automatic Control*, 56(9):2128–2139, 2011.

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

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

V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint.* Cambridge University Press, 2008.

J. A. Boyan. Technical update: least-squares temporal difference learning. *Machine Learning*, 49(2):233–246, 2002.

J. C. Cox, S. A. Ross, and M. Rubinstein. Option pricing: A simplified approach. *Journal of Financial Economics*, 7(3):229–263, 1979.

D. Duffie. *Dynamic Asset Pricing Theory.* Princeton University Press, 2010.

Y. Engel, S. Mannor, and R. Meir. Reinforcement learning with Gaussian processes. In *International Conference on Machine Learning*, 2005.

J. A. Filar, L. C. M. Kallenberg, and H. M. Lee. Variance-penalized Markov decision processes. *Mathematics of Operations Research*, 14(1):pp. 147–161, 1989.

J. A. Filar, D. Krass, and K. W. Ross. Percentile performance criteria for limiting average Markov decision processes. *IEEE Transaction on Automatic Control*, 40(1):2–10, 1995.

P. Geibel and F. Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24(1):81–108, 2005.

G. Grimmett and D. Stirzaker. *Probability and Random Processes.* Oxford university press, 2001.

J. B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I: Fundamentals.* Springer science & business media, 2013.

R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, second edition, 2012.

J. C. Hull. *Options, Futures, and Other Derivatives (6th edition).* Prentice Hall, 2006.

H. K. Khalil and J. W. Grizzle. *Nonlinear Systems.* Prentice hall New Jersey, 1996.

V. Konda. *Actor-Critic Algorithms.* PhD thesis, Department of Computer Science and Electrical Engineering, MIT, Cambridge, MA, 2002.

V. R. Konda and John N Tsitsiklis. On actor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003.

M. G. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003.

A. Lazaric, M. Ghavamzadeh, and R. Munos. Finite-sample analysis of LSTD. In *International Conference on Machine Learning*, 2010.

Y. Li, C. Szepesvari, and D. Schuurmans. Learning exercise policies for American options. In *International Conference on Artificial Intelligence and Statistics, JMLR: W&CP*, volume 5, pages 352–359, 2009.

F. A. Longstaff and E. S. Schwartz. Valuing American options by simulation: a simple least-squares approach. *Review of Financial Studies*, 14(1):113–147, 2001.

S. Mannor and J. N. Tsitsiklis. Algorithmic aspects of mean-variance optimization in Markov decision processes. *European Journal of Operational Research*, 231(3):645 – 653, 2013. ISSN 0377-2217.

O. Mihatsch and R. Neuneier. Risk-sensitive reinforcement learning. *Machine Learning*, 49 (2):267–290, 2002.

J. Moody and M. Saffell. Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4):875–889, 2001.

T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Parametric return density estimation for reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, 2010.

W. B. Powell. *Approximate Dynamic Programming.* John Wiley and Sons, 2011.

L. A. Prashanth and M. Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. In *Advances in Neural Information Processing Systems*, 2013.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., 1994.

A. Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, 125(2):235–261, 2010.

M. Sato, H. Kimura, and S. Kobayashi. TD algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence*, 16:353–362, 2001.

W. F. Sharpe. Mutual fund performance. *The Journal of Business*, 39(1):119–138, 1966.

S. M. Shortreed, E. Laber, D. J. Lizotte, T. S. Stroup, J. Pineau, and S. A. Murphy. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine learning*, 84(1):109–136, 2011.

M. J. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, pages 794–802, 1982.

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

R. S. Sutton and A. G. Barto. *Reinforcement Learning.* MIT Press, 1998.

A. Tamar and S. Mannor. Variance adjusted actor critic algorithms. *arXiv preprint arXiv:1310.3697, http://arxiv.org/abs/1310.3697*, 2013.

A. Tamar, D. Di Castro, and S. Mannor. Policy gradients with variance related risk criteria. In *International Conference on Machine Learning*, 2012.

A. Tamar, D. Di Castro, and S. Mannor. Temporal difference methods for the variance of the reward to go. In *International Conference on Machine Learning*, 2013.

A. Tamar, S. Mannor, and H. Xu. Scaling up robust MDPs using function approximation. In *International Conference on Machine Learning*, 2014.

G. Tesauro. Temporal difference learning and TD-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

J. N. Tsitsiklis and B. Van Roy. Regression methods for pricing complex American-style options. *IEEE Transactions on Neural Networks*, 12(4):694–703, 2001.

A. W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000.

N. D. Yen. Lipschitz continuity of solutions of variational inequalities with a parametric polyhedral constraint. *Mathematics of Operations Research*, 20(3):pp. 695–708, 1995.