Supplementary Material

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,

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

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

Similarly,

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

The uniqueness of the value function J for a proper policy is well known, c.f. 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 $r(x)^2 + 2r(x)\sum_{x' \in X} P(x'|x)J(x')$. Since only the rewards change, the policy remains proper and proposition 3.2.1 in (Bertsekas, 2012) applies. □

B. Proof of Proposition 8

This result is similar to Lemma 6.9 in (Bertsekas & Tsitsiklis, 1996).

Proof. 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. □

C. Proof of Theorem 9

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

$$\mathbb{E}\left[\sum_{t=0}^{\tau-1}\phi_1(x_t)\phi_2(x_t)^\top\right] = \sum_x q(x)\phi_1(x)\phi_2(x)^\top. \tag{18}$$

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, and last equality follows from the absorbing property of the terminal state. Similarly, we have

$$\mathbb{E}\left[\sum_{t=0}^{\tau-1}\phi_1(x_t)\phi_2(x_{t+1})^\top\right] = \sum_x\sum_{x'} q(x)P(x'|x)\phi_1(x)\phi_2(x')^\top, \tag{19}$$

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 (18) and (19) suggest that the approximate expressions in (13) converge to their expected values with probability 1, therefore we have

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

and

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

□

D. Proof of Theorem 10

Proof. Using (18) and (19) we have for all k

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

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

$$\hat{w}_{k+1} = \hat{w}_k + \xi_k \left(z + M\hat{w}_k + \delta M_{k+1}\right), \tag{20}$$

where

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

$$z = \begin{pmatrix} \Phi_J^\top Q r \\ \Phi_M^\top Q R r \end{pmatrix},$$

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 $F_n = \sigma(\hat{w}_m, \delta M_m, m \leq n)$, 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 $\Phi_J^\top Q\left(P-I\right)\Phi_J$ and $\Phi_M^\top Q\left(P-I\right)\Phi_M$. By Lemma 6.10 in (Bertsekas & Tsitsiklis, 1996) these matrices are negative definite. It therefore follows (see Bertsekas, 2012 example 6.6) that their eigenvalues 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.

A 1. *The map h is Lipschitz.*

A 2. *The step sizes satisfy*

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

A 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

A 4. *The functions $h_c(w) \triangleq h(cw)/c, c \geq 1$ satisfy $h_c(w) \to h_\infty(w)$ as $c \to \infty$, uniformly on compacts, and $h_\infty(w)$ is continuous. Furthermore, the Ordinary Differential Equation (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.

Theorem 7 in Chapter 3 of (Borkar, 2008) states that if A1 - A4 hold, the following condition holds

A 5. *The iterates of (20) remain bounded almost surely, i.e., $\sup_k \|\hat{w}_k\| < \infty$, a.s.*

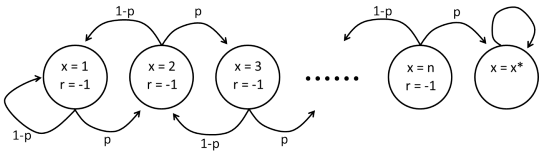
Finally, we use a standard stochastic approximation result that, given that the above conditions hold, relates the convergence of the iterates of (20) with the asymptotic behavior of the ODE

$$\dot{w}(t) = h(w(t)). \tag{21}$$

Since the eigenvalues of M have a negative real part, (21) has a unique globally asymptotically stable equilibrium point, 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 A1 - A3 and A5 hold, then $\hat{w}_k \to \hat{w}*$ as $k \to \infty$ with probability 1. □

E. Illustration of the Positive Variance Constraint

We illustrate the effect of the positive variance constraint in a simple example. Consider the following Markov chain



which consists of N states with reward -1 and a terminal state x^* with zero reward. 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. For such a small problem the fixed point equation (15) may be solved exactly, yielding the approximation depicted in Figure 2 (dotted line), for $p = 0.7$, $N = 30$, and $\lambda = 0.95$.

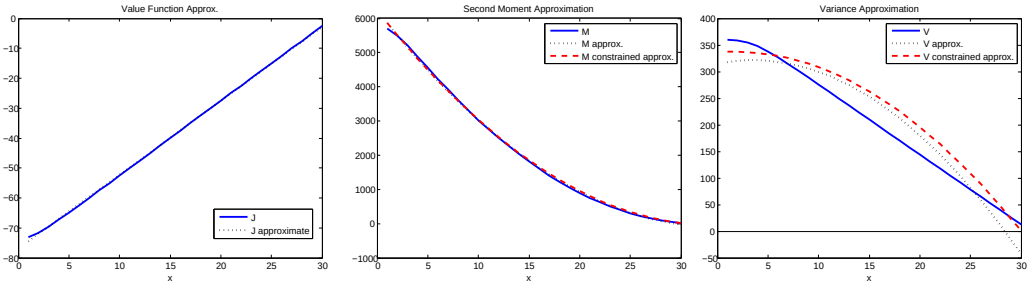


Figure 2. Value, second moment and variance approximation

Note that the variance is negative for the last two states. Using algorithm (17) we obtained a positive variance constrained approximation, which is depicted in figure 2 (dashed line). Note that the variance is now positive for all states (as was required by the constraints).