Reinforcement Learning for Cost-Aware Markov Decision Processes

Literature Review: Finite-Time Results for Two-Timescale Stochastic Approximation

Our main contribution is to propose two new RL algorithms with theoretical guarantees for solving the ratio maximization problem. Though desirable, finite-time analysis of our algorithms is a challenging problem that we do not address. Existing finite-time works, such as (Gupta et al., 2019; Hong et al., 2020; Wu et al., 2020), do not apply, since our algorithms are either too structurally dissimilar from the algorithms these works consider or do not satisfy the necessary assumptions imposed in these works. We leave the important and challenging problem of providing finite-time analysis of our algorithms to future work. Nonetheless, we provide a review of the literature on finite-time results for two-timescale stochastic approximation for completeness.

Both of the algorithms proposed in this paper follow two-timescale stochastic approximation (2TSA) schemes, which are well-studied in the RL literature (Borkar and Konda, 1997; Konda and Tsitsiklis, 2000; Bhatnagar et al., 2009). The standard technique for analyzing 2TSA is via the ODE method, which can be used to prove asymptotic convergence and rate results (Borkar, 2008). Building off the importance of asymptotic results, there has been great recent interest in non-asymptotic, finite-time results for 2TSA. Contributions to the linear setting include the finite-sample analysis (Gupta et al., 2019), which treats constant stepsizes, as well as (Doan and Romberg, 2019; Doan, 2019), which provide refined methods for selecting stepsizes. Another work in a specialized setting is (Yang et al., 2019), which provides non-asymptotic global convergence for actor-critic applied to the linear-quadratic regulator. Recasting actor-critic as a bi-level optimization problem, (Qiu et al., 2019; Hong et al., 2020) provide non-asymptotic results for certain actor-critic variants. In addition, (Kumar et al., 2019) gives convergence rates for an actor-critic scheme that uses Monte Carlo rollouts to estimate the policy gradient. More recently, (Wu et al., 2020) provides finite-time convergence for fairly general forms of actor-critic with a linear critic, while (Khodadadian et al., 2021) gives global convergence rates of natural actor-critic in the tabular setting.

Though desirable, finite-time analysis of our algorithms is a challenging problem that we do not address. Existing finite-time works, such as (Gupta et al., 2019; Doan and Romberg, 2019; Doan, 2019; Hong et al., 2020; Wu et al., 2020), do not apply, since our algorithms are either too structurally dissimilar from the algorithms these works consider or do not satisfy the necessary assumptions imposed in these works. On the one hand, the limit point of the faster timescale of our RVI Q-learning-based algorithm may fail to satisfy even the basic conditions needed to prove asymptotic results (Borkar, 2008), let alone the more nuanced conditions needed to prove existing finite-time results, such as those in (Gupta et al., 2019; Doan and Romberg, 2019; Doan, 2019; Hong et al., 2020). The policy gradient for our actor-critic algorithm, on the other hand, is of such a form that existing results such as (Qiu et al., 2019; Kumar et al., 2019; Wu et al., 2020; Khodadadian et al., 2021) do not apply.

Maximizing the Omega Ratio of a Financial Portfolio as a CAMDP

One of the principal areas where cost-aware ideas are already well developed is in financial applications, where the notion of *risk-sensitivity* is widely studied and applied. For the purposes of this paper, we consider the multistage portfolio optimization problem without transaction costs; see (Mulvey and Shetty, 2004; Calafiore, 2008; Dantzig and Infanger, 1993) for an overview and historical perspective on this problem. There are K assets with given initial prices $S_0^{(k)}$ for each $k = 1, \ldots, K$, whose prices $S_n^{(k)}$ evolve according to the stochastic difference equations

$$S_{n+1}^{(k)} - S_n^{(k)} = \mu^{(k)} S_n^{(k)} \Delta t + \sigma S_n^{(k)} W_n^{(k)}, \tag{13}$$

where the $W_n^{(k)} \sim N(0, \Delta t)$ are iid normal random variables with mean 0 and variance Δt. From a principal portfolio value P_0 the manager selects, at discrete time periods $t = 0, 1, 2, \ldots$, weights $\omega_n \in [0, 1]^K$ such that $\sum_{k=1}^K \omega_n^{(k)} = 1$ to invest in the available assets. The asset prices evolve, the new portfolio value $P_{n+1} = P_n \sum_{k=1}^K \omega_n^{(k)} S_n^{(k)}$ is realized, and the process repeats. The return on the portfolio R_{n+1} is given by $R_{n+1} = P_{n+1}/P_n - 1$. The formulation of multistage portfolio optimization problems is naturally expressed in the language of MDPs, but a general question in this field concerns

the choice of total performance measure against which optimization is performed.

A pervasive trend in the financial literature is to formulate measures of total performance for multistage problems as ratios of some measures of risk and reward functions; see (Glen and Jorion, 1993; Chekhlov et al., 2004) for examples in the literature. Let $\tau \in \mathbb{R}$ be a given real number. One useful such example is the *Sharpe ratio*, defined by

$$\mathrm{Sh}(R;\tau) = \frac{\mathbb{E}[R-\tau]}{\sqrt{\mathrm{Var}(R-\tau)}}, \tag{14}$$

which is a classical case of a measure of risk-adjusted returns. Here, τ is a target return, and the ratio rewards returns that exceed the target but punishes high-variation returns. A potential weakness of the Sharpe ratio is its emphasis on the first and second moments of the distribution of portfolio returns. This is suitable when the distribution is roughly normal, but suffers against skewed or multi-modal distributions.

Letting the cumulative distribution function for the portfolio return R be denoted F_R, a useful performance measure that captures both risk and reward is the *Omega ratio* (Keating and Shadwick, 2002), given by

$$\Omega(R;\tau) = \frac{\int_{\tau}^{\infty}[1-F_R(r)]\,dr}{\int_{-\infty}^{\tau} F_R(r)\,dr}. \tag{15}$$

This expression indicates that the Omega ratio is the ratio of the expected excess (i.e., above threshold τ) returns to the expected shortfall (i.e., below threshold τ) returns of the portfolio. A distinct advantage of the Omega ratio over financial measures such as the Sharpe ratio is that the Omega ratio incorporates information about all moments of the the distribution of R.

Given a portfolio, a policy π, the returns distribution R of the portfolio induced by π, and a threshold τ, we can explicitly formulate the Omega ratio $\Omega(R;\tau)$ as an instance of (2). To see that this is true, assume (for simplicity) that $f_R(r) = F_R'(r)$ exists, and perform a simple integration by parts to obtain

$$\int_{\tau}^{\infty} r(-f(r))dr = r[1-F_R(r)]\Big|_{\tau}^{\infty} - \int_{\tau}^{\infty}[1-F_R(r)]dr$$

$$= -\tau[1-F_R(\tau)] - \int_{\tau}^{\infty}[1-F_R(r)]dr = \int_{\tau}^{\infty}\tau(-f(r))dr - \int_{\tau}^{\infty}[1-F_R(r)]dr,$$

which can be rearranged to get

$$\int_{\tau}^{\infty}[1-F_R(r)]dr = \int_{\tau}^{\infty}(r-\tau)f(r)dr = E[\max(0,R-\tau)].$$

Similar reasoning applies to obtain $\int_{-\infty}^{\tau} F_R(r)dr = E[\max(0,\tau-R)]$. Thus, assuming that each state s_n contains the information necessary to calculate the one-step return R_n of the portfolio over the previous state s_{n-1}, we can take $r(s_n,a_n) = \max(0,R_n-\tau)$ and $c(s_n,a_n) = \max(0,\tau-R_n)$. Using these simple reward and cost functions, we can perform online maximization of $\Omega(R;\tau)$ using the algorithms described in this paper.

Proofs of Theoretical Results

We repeat all the theoretical statements here and provide their proofs.

Lemma 1. *Given the optimal ratio ρ^* of the CAMDP $(\mathcal{S},\mathcal{A},p,r,c)$, any optimal policy for $(\mathcal{S},\mathcal{A},p,\eta^{\rho^*})$ is an optimal policy for $(\mathcal{S},\mathcal{A},p,r,c)$.*

Proof. Let ρ^* denote the optimal ratio and π^* an optimal policy for the original CAMDP. Since ρ^* is optimal, we know that $J_r(\pi)/J_c(\pi) \le \rho^*$ and thus $J_r(\pi) \le \rho^* J_c(\pi)$, for any policy π. Furthermore, $J_{\eta^{\rho^*}}(\pi) = J_c(\pi) - \rho^* J_c(\pi) \le 0$, for any π. The long-run average reward of any policy for the auxiliary MDP $(\mathcal{S},\mathcal{A},p,\eta^{\rho^*})$ is therefore no more than zero. Moreover, since $J_r(\pi^*)/J_c(\pi^*) = \rho^*$ and thus $J_r(\pi^*) - \rho^* J_c(\pi^*) = 0$, this upper bound is achieved by π^*, so π^* is optimal for $(\mathcal{S},\mathcal{A},p,\eta^{\rho^*})$. By the optimality of π^* for the auxiliary MDP and the fact that $J_{\eta^{\rho^*}}(\pi^*) = 0$, any alternative optimal policy π^{**} must satisfy $J_{\eta^{\rho^*}}(\pi^{**}) = J_r(\pi^{**}) - \rho^* J_c(\pi^{**}) = 0$. This implies that $J_r(\pi^{**})/J_c(\pi^{**}) = \rho^*$, and therefore any policy π^{**} that is optimal for $(\mathcal{S},\mathcal{A},p,\eta^{\rho^*})$ must also be optimal for the original CAMDP. □

Lemma 2. *Solving (3) yields the optimal ratio and an optimal policy for CAMDP* $(\mathcal{S}, \mathcal{A}, p, r, c)$.

Proof. Let ρ^* denote the optimal ratio for the CAMDP and let π^* be a policy satisfying $\rho^* = J_r(\pi^*)/J_c(\pi^*)$. We first show that (ρ^*, π^*) is feasible to (3). Notice that $J_r(\pi^*) - \rho^* J_c(\pi^*) = 0$. Assume that there exists a policy $\hat{\pi}$ such that $J_r(\hat{\pi}) - \rho^* J_c(\hat{\pi}) > J_r(\pi^*) - \rho^* J_c(\pi^*) = 0$. This implies that $\hat{\rho} := J_c(\hat{\pi})/J_c(\hat{\pi}) > \rho^*$, contradicting the optimality of ρ^* for the CAMDP. Thus $\pi^* \in \operatorname{argmax}_{\pi'}\{J_r(\pi') - \rho^* J_c(\pi')\}$, so (ρ^*, π^*) is feasible for (3).

Since (ρ^*, π^*) is feasible and $[J_r(\pi^*) - \rho^* J_c(\pi^*)]^2 = 0$, the optimal objective function value for (3) is 0. Let (ρ^{**}, π^{**}) be an optimal solution to (3). We show that $\rho^{**} = \rho^*$ and that π^{**} is optimal for the original CAMDP. By feasibility, $\pi^{**} \in \operatorname{argmax}_{\pi'}\{J_r(\pi') - \rho^{**} J_c(\pi')\}$. We also know that $[J_r(\pi^{**}) - \rho^{**} J_c(\pi^{**})]^2 = 0$ by optimality, which implies that $\rho^{**} = J_r(\pi^{**})/J_c(\pi^{**})$. Notice that $\rho^{**} > \rho^*$ contradicts the fact that ρ^* is optimal for the CAMDP, so $\rho^{**} \leq \rho^*$. If $\rho^{**} < \rho^*$, however, we have that $J_r(\pi^*) - \rho^{**} J_c(\pi^*) > J_r(\pi^*) - \rho^* J_c(\pi^*) = 0 = J_r(\pi^{**}) - \rho^{**} J_c(\pi^{**})$, since $J_c(\pi^*) > 0$ by §2.2. But this contradicts the fact that $\pi^{**} \in \operatorname{argmax}_{\pi'}\{J_r(\pi') - \rho^{**} J_c(\pi')\}$, so it must be true that $\rho^{**} = \rho^*$. Finally, $\rho^* = J_r(\pi^{**})/J_c(\pi^{**})$, so π^{**} is an optimal policy for $(\mathcal{S}, \mathcal{A}, p, r, c)$. □

Lemma 3. *If* $\rho > \rho^*$, *then* $J_r(\pi^\rho) - \rho J_c(\pi^\rho) < 0$. *If* $\rho < \rho^*$, *then* $J_r(\pi^\rho) - \rho J_c(\pi^\rho) > 0$. *If* $\rho = \rho^*$, *then* $J_r(\pi^\rho) - \rho J_c(\pi^\rho) = 0$.

Proof. Recall that $\pi^\rho \in \operatorname{argmax}_\pi\{J_r(\pi) - \rho J_c(\pi)\}$ and $J_r(\pi) \geq 0$ and $J_c(\pi) > 0$, for all π. Since $\rho^* = J_r(\pi^*)/J_c(\pi^*)$, we know that $\rho = \rho^*$ implies $J_r(\pi^\rho) - \rho J_c(\pi^\rho) = J_r(\pi^*) - \rho^* J_c(\pi^*) = 0$. If $\rho > \rho^*$, clearly $J_r(\pi^\rho) - \rho J_c(\pi^\rho) < J_r(\pi^\rho) - \rho^* J_c(\pi^\rho) \leq J_r(\pi^*) - \rho^* J_c(\pi^*) = 0$. If $\rho < \rho^*$, then $J_r(\pi^\rho) - \rho J_c(\pi^\rho) > J_r(\pi^\rho) - \rho^* J_c(\pi^\rho) \geq 0$, where the last inequality holds since $J_r(\pi^\rho) - \rho^* J_c(\pi^\rho) < 0$ implies $\rho^* < J_r(\pi^\rho)/J_c(\pi^\rho)$, contradicting the optimality of ρ^*. □

Lemma 4. *The function* $\hat{g}(\rho) := g(\lambda(\rho), \rho) = g(Q^\rho, \rho) = V^\rho(s_{\text{ref}})$ *is strictly decreasing and piecewise linear (and thus Lipschitz) in* ρ.

Proof. By Assumption 3 and Lemma A.1.1, $V^\rho(s_{\text{ref}}) = \kappa_\rho = \max_\pi[J_r(\pi) - \rho J_c(\pi)]$, the optimal long-run average reward for the auxiliary MDP $(\mathcal{S}, \mathcal{A}, p, \eta^\rho)$. Since RVI Q-learning generates only deterministic policies, we may assume without loss of generality that the maximum is taken over the set of all deterministic polices. There are only finitely many deterministic policies, so the functions $J_r(\pi)$ and $J_c(\pi)$ each take only finitely many values as functions of π. Since $J_c(\pi) > 0$ for all policies π, we thus have that $\hat{g}(\rho)$ is simply the maximum of finitely many strictly decreasing linear functions of ρ. □

Lemma 5. $\{\rho_n\}$ *is a.s. bounded.*

Proof. Consider Lemma A.5.1 in the appendix, where in our case

$$g_d(\rho) = \frac{\hat{g}(d\rho)}{d} = \frac{J_r(\lambda(d\rho))}{d} - \rho J_c(\lambda(d\rho)). \tag{16}$$

Since there are only finitely many policies, there exist $\rho^- < 0$ and $\rho^+ > 0$ such that

$$\pi_\rho = \begin{cases} \pi_{\rho^-} & \text{if } \rho \leq \rho^-, \\ \pi_{\rho^+} & \text{if } \rho \geq \rho^+. \end{cases}$$

If there exist multiple such policies, make an arbitrary choice so that the slight abuses of notation $J_c(\pi_{\rho^+})$ and $J_c(\pi_{\rho^-})$ are well-defined. Let

$$g_\infty(\rho) = \begin{cases} -\rho J_c(\pi_{\rho^-}) & \text{if } \rho \leq 0, \\ -\rho J_c(\pi_{\rho^+}) & \text{if } \rho > 0. \end{cases}$$

Notice that, since $\hat{g}(\rho)$ is Lipschitz in ρ, $g_d(\rho)$ is Lipschitz in ρ, for all $0 < d \leq \infty$. To see that $g_d \to g_\infty$ uniformly on compact subsets of $\mathbb{R}$ as $d \to \infty$, fix $\varepsilon > 0$. Let $M = \max_\rho J_c(\lambda(\rho)), m = \min_\rho J_c(\lambda(\rho))$. Fix $d_1 > 0$ such that $J_r(\lambda(d\rho))/d < \varepsilon/2$ for all $\rho \in \mathbb{R}$ and $d > d_1$. Now notice that

$$|g_d(\rho) - g_\infty(\rho)| \leq \frac{|J_r(\lambda(d\rho))|}{d} + |-\rho J_c(\lambda(d\rho)) - g_\infty(\rho)|. \tag{17}$$

Let $\delta = \varepsilon/2|M-m|$. Choose $d_2 > d_1$ such that $d_2\delta > \max\{|\rho^-|, \rho^+\}$. Then, for $\rho \geq \delta$ and $d > d_2$, $g_\infty(\rho) = -\rho J_c(\lambda(d\rho))$ and thus the right-hand side of (17) is less than $\varepsilon/2$. When $\rho < \delta$, on the other hand, we have

$$|-\rho J_c(\lambda(d\rho)) - g_\infty(\rho)| \leq |\rho| \cdot |M-m| < \delta|M-m| = \varepsilon/2,$$

and the right-hand side of (17) is less than or equal to ε, proving uniform convergence on compacts. Finally, the ODE

$$\dot{\rho}(t) = g_\infty(\rho(t)) \tag{18}$$

clearly has $\rho = 0$ as its unique globally asymptotically stable equilibrium point. Since $\hat{g}$ is Lipschitz, given Assumption 2, and since the noise $\{\epsilon_n\}$ is asymptotically negligible, we can apply Lemma A.5.1 in the appendix to obtain that $\sup_n |\rho_n| < \infty$ a.s. □

Theorem 1. $(Q_n, \rho_n) \to \{(Q^\rho, \rho) \mid \rho \in \mathbb{R}\}$ *a.s. as* $n \to \infty$.

Proof. Rewrite (5) as, for each $(s_n, a_n) \in \mathcal{S} \times \mathcal{A}$ as follows:

$$Q_{n+1}(s_n,a_n) = Q_n(s_n,a_n) + \alpha_n\Big[[T_{\rho_n}Q_n](s_n,a_n) - V_n(s_{\text{ref}})\mathbb{1} - Q_n(s_n,a_n) + M_{n+1}(s_n,a_n)\Big], \tag{19}$$

where

$$[T_\rho Q](s,a) = \sum_{s'} p(s'|s,a)[r(s,a) - \rho c(s,a) + \max_{a'} Q(s',a')],$$

$$M_{n+1}(s_n,a_n) = r(s_n,a_n) - \rho_n c(s_n,a_n) + \max_a Q_n(s_{n+1},a) - [T_{\rho_n}Q_n](s_n,a_n).$$

Consider the equation

$$Q(s,a) = \sum_{s'} p(s' \mid s,a)\left[r(s,a) - \rho c(s,a) + \max_{a'} Q(s',a')\right] - \kappa_\rho, \tag{20}$$

where, as in Lemma A.2.1, Q^ρ denotes the unique solution to (20) satisfying $\max_a Q(s_{\text{ref}},a) = \kappa_\rho$. Letting $h(Q,\rho) = T'_\rho Q - Q$, where $T'_\rho = T_\rho Q - V(s_{\text{ref}})\mathbb{1}$, we see (19) is just a rewriting of (8). Now h is clearly Lipschitz in both Q and ρ, so we can invoke Lemmas A.2.1 and A.2.2 to obtain that Q^ρ is the unique globally asymptotically stable equilibrium point of the ODE

$$\dot{Q}(t) = T'_\rho(Q(t)) - Q(t) \tag{21}$$

which is just a rewriting of (11). Finally, $\{M_{n+1}\}$ is clearly a $\mathcal{F}_n$-martingale difference sequence, and there exists $K > 0$ such that $E[\|M_{n+1}\|^2 \mid \mathcal{F}_n] \leq K(1 + \|Q_n\|^2 + |\rho_n|^2)$. To see that the latter holds, note that

$$\begin{aligned}
E[|M_{n+1}(s_n,a_n)|^2 \mid \mathcal{F}_n] &= E[|r(s_n,a_n) - \rho_n c(s_n,a_n) + \max_a Q_n(s_{n+1},a_n) \\
&\quad - \sum_{s'} p(s'|s_n,a_n)[r(s_n,a_n) - \rho_n c(s_n,a_n) + \max_{a'} Q_n(s',a')]|^2 \mid \mathcal{F}_n] \\
&= E[|\max_a Q_n(s_{n+1},a_n) - \sum_{s'} p(s'|s,a)\max_{a'} Q_n(s',a')|^2 \mid \mathcal{F}_n],
\end{aligned}$$

and that the latter term can be expanded and bounded by $K(1 + \|Q_n\|^2)$, for some $K > 0$. Since this holds for each entry, we can simply revise K so that $E[\|M_{n+1}\|^2 \mid \mathcal{F}_n] \leq K(1 + \|Q_n\|^2) \leq K(1 + \|Q_n\|^2 + |\rho_n|^2)$.

With Lemma 5 in hand, and viewing the sequence of ρ_n as quasi-static, the sequence $\{Q_n\}$ is a.s. bounded by Lemma A.2.3. We can now apply Lemma A.4.1 to complete the proof. This application relies on the additional facts that $\{\epsilon_{n+1}\}$ from (9) is a.s. bounded, which follows from the a.s. boundedness of $\{Q_n\}$, and that $g(Q_n,\rho_n)$ is a.s. bounded since $\{(Q_n,\rho_n)\}$ remains a.s. bounded. These facts combine with the rewriting of (6) as

$$\rho_{n+1} = \rho_n + \alpha_n\Big[\frac{\beta_n}{\alpha_n} g(Q_n,\rho_n) + \frac{\beta_n}{\alpha_n}\epsilon_{n+1}\Big]$$

to show that, at the faster timescale, the ρ_n updates asymptotically track (10), since $\frac{\beta_n}{\alpha_n} \to 0$ by Assumption 2. Thus, (5) and (6) can be viewed as asymptotically tracking (10) and (11), completing the proof. □

Corollary 1. $|g(Q_n, \rho_n) - g(Q^{\rho_n}, \rho_n)| = |V_n(s_{\text{ref}}) - V^{\rho_n}(s_{\text{ref}})| \to 0$ *a.s. as* $n \to \infty$.

Proof. The claim follows immediately from Theorem 1. □

Lemma 6. ρ^* *is the unique globally asymptotically stable equilibrium point of* (12).

Proof. Recall that $\rho^* = J_r(\pi^*)/J_c(\pi^*)$, where $\pi^* = \text{argmax}_\pi J_r(\pi)/J_c(\pi)$. Clearly $g(\lambda(\rho^*), \rho^*) = 0$, and we know that $g(\lambda(\rho), \rho)$ is continuous and strictly decreasing in ρ by Lemma 4. ρ^* is thus the unique globally asymptotically stable equilibrium of (12) via a straightforward Lyapunov function argument: notice that $V(\rho) = (\rho - \rho^*)^2$ satisfies

$$\dot{V}(\rho(t)) = \frac{\partial V}{\partial \rho}\frac{\partial \rho}{\partial t} = 2(\rho(t) - \rho^*)\dot{\rho}(t) = 2(\rho(t) - \rho^*)g(\lambda(\rho(t)), \rho(t)) \leq 0 \tag{22}$$

with equality only when $\rho(t) = \rho^*$. □

Theorem 2. $(Q_n, \rho_n) \to (\lambda(\rho^*), \rho^*)$ *a.s. as* $n \to \infty$.

Proof. Letting $\epsilon_{n+1} = V_{n+1}(s_{\text{ref}}) - g(Q_n, \rho_n) = V_{n+1}(s_{\text{ref}}) - V_n(s_{\text{ref}})$, we see that updates (6) and (9) are the same, i.e.,

$$\rho_{n+1} = \rho_n + \beta_n[g(Q_n, \rho_n) + \epsilon_{n+1}] = \rho_n + \beta_n V_{n+1}(s_{\text{ref}}).$$

Following a combination of the proof strategies in Theorem 1.1 in (Borkar, 1997) and Theorem 6.2 in (Borkar, 2008), we show that the function obtained by a suitable linear interpolation of the $\{\rho_n\}$ iterates generated in this way is a.s. a (T, δ)-perturbation of the ODE (12), and thus, given Lemma 6, $\rho_n \to \rho^*$ a.s. (see the appendix A.3 for the definition of (T, δ)-perturbation). Coupled with Theorem 1, this will imply that $(Q_n, \rho_n) \to (\lambda(\rho^*), \rho^*)$.

Let $s(0) = 0$ and $s(n) = \sum_{i=0}^{n-1} \beta_i$, $n \geq 1$. Define the continuous, piecewise linear function

$$\tilde{\rho}(t) = \begin{cases} \rho_n & \text{if } t = s(n), \\ \rho_n + (\rho_{n+1} - \rho_n)\frac{t - s(n)}{s(n+1) - s(n)} & \text{otherwise.} \end{cases}$$

Let $\psi_n = \sum_{m=0}^{n-1} \beta_m \epsilon_{m+1}$, for $n \geq 1$. Letting $[t] = \max\{s(n) \mid s(n) \leq t\}$, for $t \geq 0$, we have, for any $n, m \geq 0$,

$$\begin{aligned} \tilde{\rho}(s(n+m)) = \tilde{\rho}(s(n)) &+ \int_{s(n)}^{s(n+m)} g(\lambda(\tilde{\rho}(t)), \tilde{\rho}(t))dt \\ &+ \int_{s(n)}^{s(n+m)} [g(\lambda(\tilde{\rho}([t])), \tilde{\rho}([t])) - g(\lambda(\tilde{\rho}(t)), \tilde{\rho}(t))]dt \\ &+ \sum_{k=1}^{m-1} \beta_{n+k}(g(Q_{n+k}, \rho_{n+k}) - g(\lambda(\rho_{n+k}), \rho_{n+k})) \\ &+ (\psi_{n+m+1} - \psi_n). \end{aligned}$$

For $s \geq 0$, let $\rho^s(t)$, $t \geq 0$, denote the solution of (12) with initial condition $\rho^s(s) = \tilde{\rho}(s)$. Notice that we can write

$$\rho^{s(n)}(s(m+n)) = \tilde{\rho}(s(n)) + \int_{s(n)}^{s(n+m)} g(\lambda(\tilde{\rho}(t)), \tilde{\rho}(t))dt,$$

which means that

$$\begin{aligned} \tilde{\rho}(s(n+m)) - \rho^{(s(n))}(s(n+m)) = &\int_{s(n)}^{s(n+m)} (g(\lambda(\tilde{\rho}([t])), \tilde{\rho}([t])) - g(\lambda(\tilde{\rho}(t)), \tilde{\rho}(t)))dt \\ &+ \sum_{k=1}^{m-1} \beta_{n+k}(g(Q_{n+k}, \rho_{n+k}) - g(\lambda(\rho_{n+k}), \rho_{n+k})) \\ &+ (\psi_{n+m+1} - \psi_n). \end{aligned}$$

Exactly the same argument as in Lemma 2.1 of (Borkar, 2008) applies to allow us to bound the first term in this expression. The third term can be a.s. bounded since, for given $m \in \mathbb{N}$, $\{\psi_{n+m+1} - \psi_n\}$ is a.s. bounded and converges a.s. to 0 due to the facts that $\beta_n \downarrow 0$ by Assumption 2 and $\epsilon_n \downarrow 0$ by Theorem 1. Thus, for any $T > 0$ and $s(n+m) \in [s(n), s(n)+T]$, there exists a nonnegative sequence $\{K_{T,n}\}$ such that $K_{T,n} \to 0$ a.s. as $n \to \infty$ and

$$|\tilde{\rho}(s(n+m)) - \rho^{(s(n))}(s(n+m))| \leq K_{T,n} + \sum_{k=1}^{m-1} \beta_{n+k} |g(Q_{n+k}, \rho_{n+k}) - g(\lambda(\rho_{n+k}), \rho_{n+k})|, \tag{23}$$

for each $n \geq 0$. But we know by Corollary 1 that the remaining right-hand side term converges to 0 a.s. as $n \to \infty$. Following the same arguments as in the proof of Lemma 2.1 in (Borkar, 2008), we have that, for any $T > 0$,

$$\lim_{s \to \infty} \sup_{t \in [s, s+T]} |\tilde{\rho}(t) - \rho^s(t)| = 0 \text{ a.s.}$$

The same arguments as in the proof Theorem 6.2 in (Borkar, 2008) now apply to complete the proof. □

Lemma 7. *For a given policy π_θ and for both $i = r, c$, with $\{\mu_n^i\}$ and $\{\nu_n^i\}$ generated from the critic step in Algorithm 2, we have $\lim_{n\to\infty} \mu_n^i = J_i(\theta)$ and $\lim_{n\to\infty} \nu_n^i = \nu_\theta^i$ a.s., where ν_θ^r and ν_θ^c are the unique solutions to*

$$\Phi^\top D^\theta \left[r^\theta - J_r(\theta) \cdot \mathbb{1} + P^\theta(\Phi \nu_\theta^r) - \Phi \nu_\theta^r \right] = \mathbf{0},$$
$$\Phi^\top D^\theta \left[c^\theta - J_c(\theta) \cdot \mathbb{1} + P^\theta(\Phi \nu_\theta^c) - \Phi \nu_\theta^c \right] = \mathbf{0}.$$

Proof. The same arguments as in the proof of Lemma 5 in (Bhatnagar et al., 2009) apply to obtain the result. □

Now that we have access to v_θ^r, v_θ^c by the preceding lemma, we will use them to estimate the following policy gradient:

Lemma 8. *For any $\theta \in \Theta$, let*

$$\delta_n^{\theta,r} = r_n - J_r(\theta) + [\phi(s_{n+1})]^\top \nu_\theta^r - [\phi(s_n)]^\top \nu_\theta^r,$$
$$\delta_n^{\theta,c} = c_n - J_c(\theta) + [\phi(s_{n+1})]^\top \nu_\theta^c - [\phi(s_n)]^\top \nu_\theta^c,$$

denote the stationary estimates of the TD-errors, let

$$\bar{V}_r^\theta(s) = \mathbb{E}\Big\{ r(s,a) - J_r(\theta) + [\phi(s')]^\top \nu_\theta^r \Big\},$$
$$\bar{V}_c^\theta(s) = \mathbb{E}\Big\{ c(s,a) - J_c(\theta) + [\phi(s')]^\top \nu_\theta^c \Big\},$$

where the expectation is taken over $a \sim \pi_\theta(\cdot \,|\, s)$ and $s' \sim p(\cdot \,|\, s, a)$, and let $e_i^\theta = \sum_{s \in \mathcal{S}} d^{\pi_\theta}(s) \left[\nabla_\theta \bar{V}_i^\theta(s) - [\phi(s)]^\top \nabla_\theta \nu_\theta^i \right]$ for $i = r, c$. Then,

$$\mathbb{E}\left[\frac{J_r(\theta)}{J_c(\theta)} \cdot \nabla_\theta \log \pi_\theta(a_n \,|\, s_n) \cdot \left(\frac{\delta_n^{\theta,r}}{J_r(\theta)} - \frac{\delta_n^{\theta,c}}{J_c(\theta)} \right) \Big| \theta \right] = \nabla_\theta L(\theta) + \frac{J_r(\theta)}{J_c(\theta)} \cdot \left[\frac{e_r^\theta}{J_r(\theta)} - \frac{e_c^\theta}{J_c(\theta)} \right].$$

Proof. From Lemma 4 in (Bhatnagar et al., 2009), we have that for any $\theta \in \Theta$ and $i = r, c$,

$$\mathbb{E}[\delta_n^{\theta,i} \psi_\theta(s_n, a_n) \,|\, \theta] = \nabla_\theta J_i(\theta) + e_i^\theta,$$

where $e_i^\theta = \sum_{s \in \mathcal{S}} d^{\pi_\theta}(s) \big\{ \nabla_\theta \bar{V}_i^\theta(s) - [\phi(s)]^\top \nabla_\theta \nu_\theta^i \big\}$. Hence, given the limit points v_θ^r, v_θ^c obtained by Lemma 7, we further have

$$\begin{aligned}
\mathbb{E}\left[\frac{J_r(\theta)}{J_c(\theta)} \cdot \psi_\theta(s_n, a_n) \cdot \left(\frac{\delta_n^{\theta,r}}{J_r(\theta)} - \frac{\delta_n^{\theta,c}}{J_c(\theta)} \right) \Big| \theta \right] &= \frac{[\nabla_\theta J_r(\theta) + e_r^\theta]}{J_c(\theta)} - \frac{J_r(\theta)}{[J_c(\theta)]^2} \cdot [\nabla_\theta J_c(\theta) + e_c^\theta] \\
&= \nabla_\theta L(\theta) + \frac{J_r(\theta)}{J_c(\theta)} \cdot \Big[\frac{e_r^\theta}{J_r(\theta)} - \frac{e_c^\theta}{J_c(\theta)} \Big],
\end{aligned}$$

which completes the proof. □

Theorem 3. *Under Assumptions 2 and 4–6, given any $\epsilon > 0$, there exists $\delta > 0$ such that, for $\{\theta_n\}$ obtained from Algorithm 2, if $\sup_{\theta_n} \|e^{\theta_n}\| < \delta$, then $\theta_n \to \mathcal{Z}^\epsilon$ a.s. as $n \to \infty$.*

Proof. The proof proceeds along the same lines as that of Theorem 1 in (Bhatnagar et al., 2009). From Lemma 8, we obtain that $\delta_n^i \to \delta_n^{\theta_n,i}$ and $\mu_n^i \to J_i(\theta_n)$ for both $i = r, c$ as $t \to \infty$. Let

$$\mathbb{E}\left[\frac{J_r(\theta)\cdot\psi_\theta(s_n,a_n)}{J_c(\theta)}\left(\frac{\delta_n^{\theta,r}}{J_r(\theta)}-\frac{\delta_n^{\theta,c}}{J_c(\theta)}\right)\Big|\theta=\theta_n\right] = -\frac{1}{J_c(\theta_n)}\cdot h_1(\theta_n) - \frac{J_r(\theta_n)}{[J_c(\theta_n)]^2}\cdot h_2(\theta_n) =: -h(\theta),$$

where

$$h_1(\theta) = -\mathbb{E}\Big(\psi_\theta(s,a)\cdot\big\{r(s,a)-J_r(\theta)+[\phi(s')]^\top\nu_\theta^r\big\}\Big),$$
$$h_2(\theta) = \mathbb{E}\Big(\psi_\theta(s,a)\cdot\big\{c(s,a)-J_c(\theta)+[\phi(s')]^\top\nu_\theta^c\big\}\Big),$$

and the expectation is taken over $s' \sim p(\cdot\,|\,s,a)$, $s \sim d^{\pi_\theta}(\cdot)$, and $a \sim \pi_\theta(\cdot\,|\,s)$. It has been shown in the proof of Theorem 1 in (Bhatnagar et al., 2009) that both $h_1(\theta)$ and $h_2(\theta)$ are Lipschitz continuous, as are $J_c(\theta)$ and $J_r(\theta)$. Also, by Assumption 6, $1/J_c(\theta)$ is upper-bounded by $1/K_c$, which by definition shows that $1/J_c(\theta)$ is also Lipschitz continuous. Furthermore, $1/[J_c(\theta)]^2$ is Lipschitz continuous since $1/J_c(\theta)$ is Lipschitz continuous and bounded. On the other hand, since both $h_1(\theta)$ and $h_2(\theta)$ are continuous over the compact Θ, they are bounded. Therefore, both $-h_1(\theta_n)/J_c(\theta_n)$ and $J_r(\theta_n)\cdot h_2(\theta_n)/[J_c(\theta_n)]^2$ are Lipschitz continuous since they are products of Lipschitz continuous and bounded functions. This establishes the Lipschitz continuity of $-h(\theta)$. The rest of the proof is identical to that of Theorem 1 in (Bhatnagar et al., 2009), which completes the proof. □

Experimental Results

Rewards, Costs, and Hyperparameters for Synthetic CAMDP Experiments

$r(s,a)$	s^3	$s+a$	$(s\cdot a) \mod 2$	s^2+a^2
$c(s,a)$	$\max\{1, s\cdot a\}$	$\max\{1, s\cdot a\}^{-1}$	$\max\{1, s+a\}$	$\max\{1,(s-a)^2\}$

Table 1. Reward/cost function combinations.

		Problem 1			Problem 2			Problem 3			Problem 4		
$r(s,a)$		s^3			$s+a$			$(s\cdot a) \mod 2$			s^2+a^2		
$c(s,a)$		$\max\{1, s\cdot a\}$			$\max\{1, s\cdot a\}^{-1}$			$\max\{1, s+a\}$			$\max\{1,(s-a)^2\}$		
		CARVI Q-learning											
$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{A}\rvert$	α	β	ε	α	β	ε	α	β	ε	α	β	ε
5	5	0.01	0.005	0.001	0.1	0.1	0.01	0.01	0.001	0.001	0.0001	0.0001	0.005
10	10	0.01	0.005	0.001	0.1	0.1	0.01	0.001	0.0005	0.001	0.005	0.001	0.001
		CAAC											
$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{A}\rvert$	α	β	μ_{lr}	α	β	μ_{lr}	α	β	μ_{lr}	α	β	μ_{lr}
5	5	0.01	0.005	0.005	0.01	0.01	0.01	0.01	0.005	0.005	0.01	0.01	0.005
10	10	0.01	0.005	0.05	0.01	0.01	0.005	0.01	0.01	0.1	0.01	0.01	0.005

Table 2. Experiment hyperparameters. Columns correspond to reward/cost function combinations, while rows correspond to problem sizes.

Deep CARVI Q-learning Experiments

We implemented CARVI Q-learning using deep neural networks for the Q function approximators and tested it on a cost-aware modification of the classic mountain car control environment (Moore, 1990) provided in OpenAI's Gym RL testbed (OpenAI, 2021a; Brockman et al., 2016). The goal of the agent is to drive the car up the hill to reach the flag in Figure 2. The car does not have enough power to directly accelerate up the hill, however, so the agent must learn to build momentum driving up and down the sides of the valley. In the OpenAI implementation of this environment, the state s consists of two values, $s = (p, v)$: the car's position p on the interval [-1.2, 0.6], representing its horizontal position in Figure 2, and its current velocity v. The action space contains three actions: accelerate to the left, do not accelerate, and accelerate to the right. The reward is -1 for each timestep that the agent has not successfully reached the flag at position 0.5. Each episode lasts until either the agent reaches the flag or 500 timesteps have elapsed, whichever comes first. According to OpenAI's MountainCar-v0 leaderboard at the time of writing (OpenAI, 2021b), this problem is considered "solved" when the agent consistently achieves an episode reward of -110 or greater.

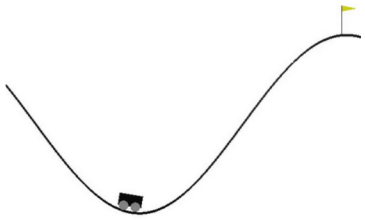


Figure 2. OpenAI Gym's MountainCar-v0

To test out our implementation of deep CARVI Q-learning, we augmented the original MountainCar-v0 environment by imposing a cost of the form

$$c(s, a) = c(p, v, a) = \min\{1 - 0.99 \cdot I_{[0.5,0.6]}(p),\ 1 - 0.9 \cdot I_{(0.2,0.6]}(p)\}$$

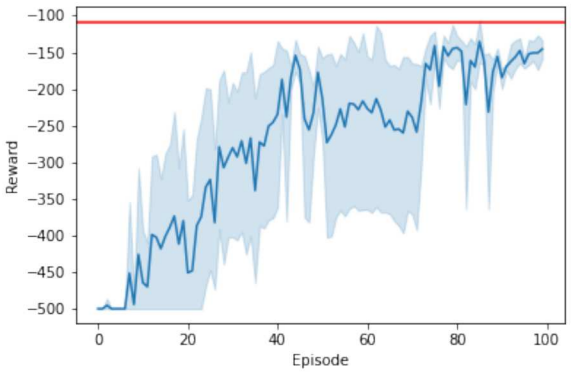
at each timestep, where I_A is the indicator function on the set A, i.e., $I_A(a) = 1$ if $a \in A$ and 0 if $a \notin A$. This simple cost function, which outputs 1 when $p \leq 0.2$, 0.1 when $0.2 < p < 0.5$, and 0.01 when $p \geq 0.5$, penalizes the agent less as it gets closer to the goal. To complete the conversion of MountainCar-v0 into a proper cost-aware environment, the original reward is multiplied by -1, so that the agent receives a reward of 1 at each timestep. The resulting goal of the agent in our cost-aware mountain car environment is to maximize the long-run average reward divided by the long-run average cost.

Our batch learning implementation in PyTorch (PyTorch, 2021) of Algorithm 1 uses a fully connected neural network with two hidden layers of 256 neurons each and ReLU activation functions for the Q network. We determined hyperparameters through trial and error, eventually settling on: a replay buffer of length $1,000,000$; batch sizes of 256 sampled with replacement from the replay buffer; a Q function learning rate of 0.0001; a ρ learning rate of 0.00001; an initial greedy $\varepsilon = 0.5$, with a per-timestep exponential decay rate of 0.999; and the Adam optimizer with gradient clipping of radius 1.

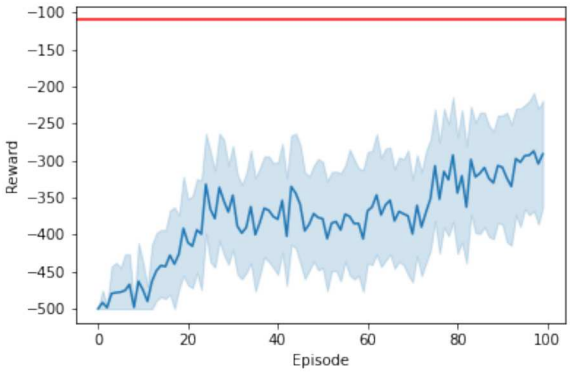
We ran 15 independent replications of our algorithm on the cost-aware mountain car environment for 100 episodes each. The results can be seen in Figures 3 and 4. In order to demonstrate how incorporating *a priori* knowledge to shape rewards can improve performance on the original problem, Figure 3 plots the agents' scores using the original reward function in OpenAI's MountainCar-v0. As can be seen from Figure 3a, the five best runs achieved solid performance, nearly solving the problem after only 100 episodes. Figure 3b, which represents rewards from all 15 runs, also demonstrates a clear upward trend. Figure 4 illustrates improvement in the ratio of sample average reward to sample average cost for each episode. It is interesting to note that the mean ratios across all 15 runs in Figure 4b appear to be somewhat better than those of the five agents who performed best in Figure 3a. This highlights the fact that, though reward shaping in this way can indeed be used to improve performance on the original MountainCar-v0, CARVI Q-learning is nonetheless solving a distinct ratio maximization problem.

A. Appendix

Some important definitions and results from the literature that we have used above are reproduced here for easy reference.

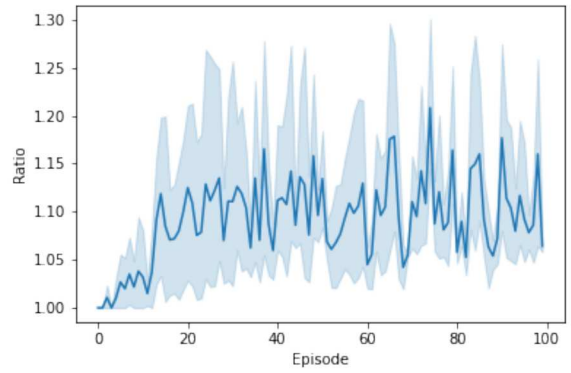


(a) Top five runs of CARVI Q-learning on original MountainCar.

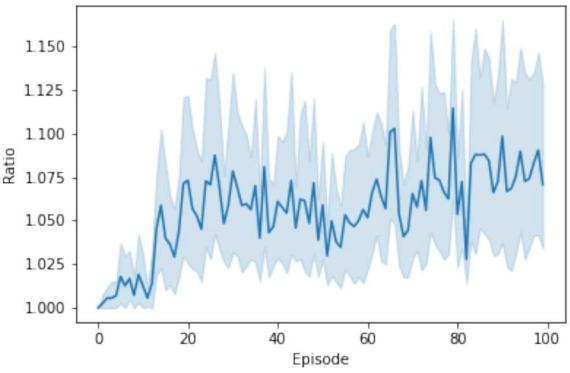


(b) All runs of CARVI Q-learning on original MountainCar.

Figure 3. Deep CARVI Q-learning rewards per episode using the original MountainCar-v0 reward function of -1 per timestep. The solid blue line represents the mean across agents, while the light blue shading gives a 95% confidence interval. The solid red line shows episode reward at which the original OpenAI Gym environment is considered "solved".



(a) Top five runs on cost-aware MountainCar.



(b) All runs on cost-aware MountainCar.

Figure 4. Deep CARVI Q-learning sample average reward over sample average cost per episode. The solid blue line represents the mean across agents, while the light blue shading gives a 95% confidence interval.

A.1. Relative Value Iteration

Fix an average reward MDP $(\mathcal{S}, \mathcal{A}, p, r)$ and a reference state s_{ref}. Consider the relative value iteration (RVI) update

$$V_{n+1}(s) = \max_a \left[\sum_{s'} p(s' \mid s, a)[r(s,a) + V_n(s')] - V_n(s_{\text{ref}}) \right] \tag{24}$$

adapted from §2.2 of (Abounadi et al., 2001). Under Assumption 3, we then have the following simplified version of Proposition 5.3.2 of (Bertsekas, 2012):

Lemma A.1.1. *The sequence $\{V_n\}$ converges to a vector V such that $V(s_{\text{ref}}) = \kappa$, where κ is the optimal long-run average reward for $(\mathcal{S}, \mathcal{A}, p, r)$.*

A.2. RVI Q-learning

Fix an average reward MDP $(\mathcal{S}, \mathcal{A}, p, r)$ and a reference state s_{ref}, and let κ denote the optimal average reward of the MDP. Consider the equation

$$Q(s,a) = \sum_{s'} p(s' \mid s,a) \left[r(s,a) + \max_{a'} Q(s',a') \right] - \kappa, \tag{25}$$

and the synchronous RVI Q-learning updates

$$Q_{n+1}(s_n,a_n) = \; Q_n(s_n,a_n) + \alpha_n \Big[r(s_n,a_n) + \max_{a} Q_n(s_{n+1},a) - V_n(s_{\text{ref}}) - Q_n(s_n,a_n) \Big], \tag{26}$$

where $V_n(s) := \max_a Q_n(s,a)$. Define $T'(Q) = T(Q) - \kappa \mathbb{1}$, where T is define entry-wise by

$$[T(Q)](s,a) = \sum_{s'} p(s' \mid s,a) \left[r(s,a) + \max_{a'} Q(s',a') \right],$$

and consider the ODE

$$\dot{Q}(t) = T'(Q(t)) - Q(t). \tag{27}$$

We have the following versions of Lemma 3.2, Theorem 3.4, and Lemma 3.7, respectively, from (Abounadi et al., 2001):

Lemma A.2.1. *Equation (27) has a unique equilibrium at Q^*, which is the unique solution to (25) such that $\max_a Q^*(s_{\text{ref}}, a) = \kappa$.*

Lemma A.2.2. *Q^* is the global asymptotically stable equilibrium point for (27).*

Lemma A.2.3. *The sequence $\{Q_n\}$ generated by (26) is a.s. bounded.*

A.3. (T, δ)-perturbations

Let $f : \mathbb{R}^N \to \mathbb{R}^N$ be Lipschitz and consider the ODE

$$\dot{z}(t) = f(z(t)) \tag{28}$$

with globally asymptotically stable attractor A.

Definition A.3.1. *For $T, \delta > 0$, a bounded, Borel-measurable function $w : \mathbb{R}^+ \to \mathbb{R}^N$ is a (T,δ)-perturbation of (28) if there exists a sequence $\{T_n\}$ in $[0,\infty)$ with $T_0 = 0$ such that $T_{n+1} - T_n \geq T$, for all $n \geq 1$, and solutions $z^j(t)$, $t \in [T_j, T_{j+1}]$ of (28) such that, for $j \geq 0$,*

$$\sup_{t \in [T_j, T_{j+1}]} \| z^j(t) - w(t) \| < \delta.$$

A.4. Faster Timescale Convergence

Consider the scheme

$$x_{n+1} = x_n + a_n [h(x_n, y_n) + M^{(1)}_{n+1}], \tag{29}$$

$$y_{n+1} = y_n + b_n [g(x_n, y_n) + M^{(2)}_{n+1} + \epsilon_{n+1}], \tag{30}$$

where $h : \mathbb{R}^{d+k} \to \mathbb{R}^d$, $g : \mathbb{R}^{d+k} \to \mathbb{R}^k$, $\{M^{(i)}_n\}$, $i = 1, 2$ are martingale difference sequences with respect to the σ-fields $\mathcal{F}_n = \sigma(x_m, y_m, M^{(1)}_m, M^{(2)}_m; m \leq n)$, $\{\epsilon_n\}$ is an error sequence, and the a_n, b_n form decreasing stepsize sequences. Consider the ODE

$$\dot{y}(t) = 0, \tag{31}$$

$$\dot{x}(t) = h(x(t), y(t)). \tag{32}$$

We make the following assumptions:

Assumption A.4.1. *For each $y \in \mathbb{R}^k$, the ODE $\dot{x}(t) = h(x(t), y)$ corresponding to (31), (32) has a unique globally asymptotically stable equilibrium $\lambda(y)$.*

Assumption A.4.2. $\sum a_n = \sum b_n = \infty, \sum(a_n^2 + b_n^2) < \infty$, *and* $b_n = o(a_n)$.

Assumption A.4.3. *There exists some $K > 0$ such that $E[\|M_{n+1}^{(i)}\|^2 \mid \mathcal{F}_n] \leq K(1 + \|x_n\|^2 + \|y_n\|^2)$, for $i = 1, 2$ and $n \geq 0$, and $\{\epsilon_n\}$ is asymptotically negligible, i.e., $\|\epsilon_n\| = o(1)$.*

Assumption A.4.4. *There exists a bounded set $Q \subset \mathbb{R}^d \times \mathbb{R}^k$ such that $\{(x_n, y_n)\} \subset Q$ with probability one.*

Assumption A.4.5. *h is Lipschitz, and g is bounded on Q.*

We have the following result, which appears as Lemma 6.1 in (Borkar, 2008):

Lemma A.4.1. $(x_n, y_n) \to \{(\lambda(y), y) \,|\, y \in \mathbb{R}^k\}$ *a.s. as* $n \to \infty$.

Note that the noise terms $\{\epsilon_n\}$ are accommodated by the third extension discussed in §2.2 of (Borkar, 2008). Also note that Lemma A.4.1 does not rely on $\lambda(y)$ being Lipschitz in y, which is important given that the limit point Q^ρ of the faster timescale of CARVI Q-learning is not guaranteed to be Lipschitz in ρ.

A.5. Stability Criterion

Consider the stochastic approximation scheme in $\mathbb{R}^N$ given by

$$z_{n+1} = z_n + a_n \left[g(z_n) + M_{n+1} + \epsilon_{n+1}\right], \tag{33}$$

with the following assumptions:

Assumption A.5.1. *$g : \mathbb{R}^N \to \mathbb{R}^N$ is Lipschitz.*

Assumption A.5.2. *The sequence $\{a_n\} \subset \mathbb{R}$ satisfies $\sum_n a_n = \infty$, $\sum_n a_n^2 < \infty$.*

Assumption A.5.3. *$\{M_n\}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_n = \sigma(z_m, M_m, m \leq n)$, there exists $K > 0$ such that $E[\|M_{n+1}\|^2 \mid \mathcal{F}_n] \leq K(1 + \|z_n\|^2)$ a.s., and $\{\epsilon_n\}$ is $o(1)$, i.e., is asymptotically negligible.*

Assumption A.5.4. *The functions $g_d(z) = g(dz)/d$, $d \geq 1$ satisfy $g_d(z) \to g_\infty(z)$ as $d \to \infty$ uniformly on compacts for some continuous function $g_\infty : \mathbb{R}^N \to \mathbb{R}^N$. In addition, the ODE*

$$\dot{z}(t) = g_\infty(z(t)) \tag{34}$$

has the origin as its globally asymptotically stable equilibrium.

We then have

Lemma A.5.1. $\sup_n \|z_n\| < \infty$ *a.s.*

See §2.2 and §3.2 in (Borkar, 2008) for the proof. Since the stability proofs in §3.2 of (Borkar, 2008) are path-wise, the comments at the end of §2.2 in (Borkar, 2008) regarding how to handle noise terms apply to accommodate the asymptotically negligible noise terms $\{\epsilon_n\}$ in (33).

References

Abounadi, J., Bertsekas, D., and Borkar, V. S. (2001). Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698.

Bertsekas, D. P. (2012). *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific Belmont, MA, 4 edition.

Bhatnagar, S., Sutton, R., Ghavamzadeh, M., and Lee, M. (2009). Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482.

Borkar, V. S. (1997). Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294.

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

Borkar, V. S. and Konda, V. R. (1997). The actor-critic algorithm as multi-time-scale stochastic approximation. *Sadhana*, 22(4):525–543.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). OpenAI Gym. *arXiv preprint arXiv:1606.01540*.

Calafiore, G. C. (2008). Multi-period portfolio optimization with linear control policies. *Automatica*, 44(10):2463–2473.

Chekhlov, A., Uryasev, S., and Zabarankin, M. (2004). Portfolio optimization with drawdown constraints. In *Supply chain and finance*, pages 209–228. World Scientific.

Dantzig, G. B. and Infanger, G. (1993). Multi-stage stochastic linear programs for portfolio optimization. *Annals of Operations Research*, 45(1):59–76.

Doan, T. T. (2019). Finite-time analysis and restarting scheme for linear two-time-scale stochastic approximation. *arXiv preprint arXiv:1912.10583*.

Doan, T. T. and Romberg, J. (2019). Linear two-time-scale stochastic approximation a finite-time analysis. In *57th Annual Allerton Conference on Communication, Control, and Computing*, pages 399–406. IEEE.

Glen, J. and Jorion, P. (1993). Currency hedging for international portfolios. *The Journal of Finance*, 48(5):1865–1886.

Gupta, H., Srikant, R., and Ying, L. (2019). Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4706–4715.

Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. (2020). A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*.

Keating, C. and Shadwick, W. F. (2002). A universal performance measure. *Journal of Performance Measurement*, 6(3):59–84.

Khodadadian, S., Doan, T. T., Maguluri, S. T., and Romberg, J. (2021). Finite sample analysis of two-time-scale natural actor-critic algorithm. *arXiv preprint arXiv:2101.10506*.

Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014.

Kumar, H., Koppel, A., and Ribeiro, A. (2019). On the sample complexity of actor-critic method for reinforcement learning with function approximation. *arXiv preprint arXiv:1910.08412*.

Moore, A. W. (1990). *Efficient Memory-Based Learning for Robot Control*. PhD thesis, University of Cambridge, Computer Laboratory.

Mulvey, J. M. and Shetty, B. (2004). Financial planning via multi-stage stochastic optimization. *Computers Operations Research*, 31(1):1 – 20.

OpenAI (2021a). OpenAI Gym MountainCar-v0 environment. https://gym.openai.com/envs/MountainCar-v0.

OpenAI (2021b). OpenAI Gym MountainCar-v0 environment leaderboard. https://github.com/openai/gym/wiki/Leaderboard#mountaincar-v0.

PyTorch (2021). PyTorch machine learning library. https://pytorch.org/.

Qiu, S., Yang, Z., Ye, J., and Wang, Z. (2019). On the finite-time convergence of actor-critic algorithm. In *Optimization Foundations for Reinforcement Learning Workshop at Advances in Neural Information Processing Systems (NeurIPS)*.

Wu, Y., Zhang, W., Xu, P., and Gu, Q. (2020). A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350*.

Yang, Z., Chen, Y., Hong, M., and Wang, Z. (2019). Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *Advances in Neural Information Processing Systems*, 32:8353–8365.