Reinforcement Learning with Quasi-Hyperbolic Discounting

S.R. Eshwar [1] **Nibedita Roy** [1] **Gugan Thoppe** [1 2]

Abstract

Reinforcement learning has traditionally been studied with exponential discounting or the average reward setup, mainly due to their mathematical tractability. However, such frameworks fall short of accurately capturing human behavior, which has a bias towards immediate gratification. Quasi-Hyperbolic (QH) discounting is a simple alternative for modeling this bias. Unlike in traditional discounting, though, the optimal QH-policy, starting from some time t_1, can be different to the one starting from t_2. Hence, the future self of an agent, if it is naive or impatient, can deviate from the policy that is optimal at the start, leading to sub-optimal overall returns. To prevent this behavior, an alternative is to work with a policy anchored in a Markov Perfect Equilibrium (MPE). In this work, we propose the first model-free algorithm for finding an MPE. Using a brief two-timescale analysis, we provide evidence that our algorithm converges to invariant sets of a suitable Differential Inclusion (DI). We also show that the QH Q-value function of any MPE would be an invariant set of our identified DI. Finally, we validate our claims numerically for the standard inventory system with stochastic demands. We believe our work significantly advances the practical application of reinforcement learning.

1. Introduction

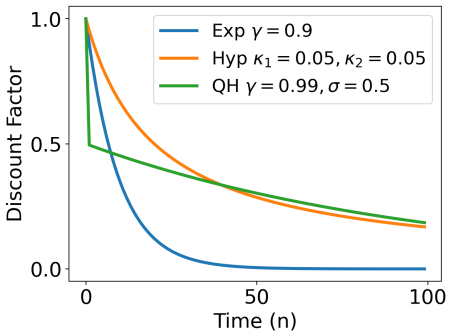
Reinforcement Learning (RL) (Sutton & Barto, 2018; Bertsekas, 2019) looks at identifying a policy/strategy for an agent to optimally complete a task with sequential decisions. So far, a strategy $\bar{\pi}$'s optimality has been decided based on either the expected exponentially discounted sum or the long-term average of the sequence of rewards received under that strategy. That is, based on either $\sum_{n=0}^{\infty} \gamma^n r_n$ or $\lim_{T\to\infty} \frac{1}{T} \sum_{n=0}^{T-1} r_n$, where r_n is the expected reward under policy $\bar{\pi}$ at time n and $\gamma \in [0, 1)$. Exponential discounting is preferred when the agent has impatience, i.e., immediate gains have emphasis over future gains, with the emphasis level decided by the γ value. In contrast, the average of the rewards is preferred when the present and future rewards are to be treated equally. However, evidence is now growing that these discounting ideas fail to model human behaviors accurately (Dhami, 2016).

Humans are known to be impatient over shorter horizons, but not so much over longer horizons. That is, we have a bias towards instant gratification. This can be understood from the famous example by Richard Thaler (Thaler, 1981), who said, "Most people would prefer one apple today to two apples tomorrow, but they prefer two apples in 51 days to one in 50 days." Observe that there is a reversal of preferences when the time frame shifts. This phenomenon is known as the *common difference effect* (Dhami, 2016). Such preference reversals cannot happen under a policy that is optimal with respect to either of the two traditional discounting models. This is because of their time-consistent nature (Sutton & Barto, 2018), i.e, this optimal policy remains optimal even when reconsidered from some later time as well. This demonstrates the limitations of these discounting models in explaining human behaviors.

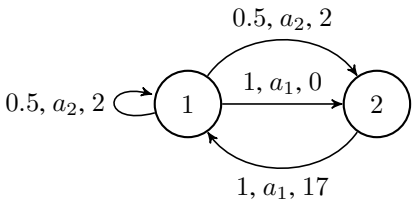
Hyperbolic discounting (Loewenstein & Prelec, 1992) is a leading candidate (Ainslie, 1975; Cropper et al., 1992; Frederick et al., 2002) for explaining the common difference effect. The value of a strategy $\bar{\pi}$ under this discounting model is $\sum_{n=0}^{\infty} b_n r_n$, where r_n is defined as before and $b_n = (1 + \kappa_1 n)^{-\kappa_2/\kappa_1}$ for some $\kappa_1, \kappa_2 > 0$. However, this form of discounting is quite complicated, making its study hard. This brings forth Quasi-Hyperbolic (QH) discounting (Phelps & Pollak, 1968; Laibson, 1997), which is a simpler and more tractable alternative. In QH discounting, $b_0 = 1$ and $b_n = \sigma\gamma^n$, $n \geq 1$, for some $\sigma \in [0, 1]$ and $\gamma \in [0, 1)$. The symbol σ is the short-term discount factor, while γ is the long-term discount factor. Clearly, for $\sigma = 1$, QH discounting matches exponential discounting. A comparison of the discount factors under exponential, hyperbolic, and quasi-hyperbolic discounting is given in Figure 1a. Unlike exponential discounting, note that there is a sharp decrease

[1]Department of Computer Science and Automation, Indian Institute of Science, Bengaluru, India. [2]Robert Bosch Centre for Data Science and Artificial Intelligence, IIT Madras, Chennai, India. Correspondence to: S.R. Eshwar <eshwarsr@iisc.ac.in>.

Workshop on Foundations of Reinforcement Learning and Control at the 41st International Conference on Machine Learning, Vienna, Austria.



(a) Comparison of discount factors



(b) Two-state MDP example

	$\bar{f}$	$\bar{g}$	$\bar{h}$
$(1, a_1)$	18.88	16.83	18.00
$(1, a_2)$	19.00	16.77	18.00
$(2, a_1)$	32.11	29.56	31.00

(c) Q-values under QH-discounting.

Figure 1: (1a) Comparison of discount factors under exponential, hyperbolic, and quasi-hyperbolic discounting models. (1b) A two-state MDP example where the action set of state 1 is $\{a_1, a_2\}$, while that of state 2 is $\{a_1\}$. For each tuple on the arrow, the first element is the probability of the transition, the second is the action taken, and the third is the instantaneous reward. (1c) Q-values under QH-discounting for the policies $\bar{f} \equiv (f, f, \ldots), \bar{g} \equiv (g, g, \ldots)$, and $\bar{h} \equiv (h, h, \ldots)$, where $f(1) = f(2) = g(2) = h(2) = a_1$, while $g(1) = a_2$ and $h(a_1|1) = h(a_2|1) = 0.5$. Each row in the table refer to the (s, a) pairs, while the columns represent the corresponding policies.

in hyperbolic and QH discount factors initially, after which they decrease more gradually. In this work, we initiate the study of RL with QH discounting.

Under exponential discounting, the optimal policy $\bar{\pi}_*$ is deterministic and stationary, and has a greedy relationship with its Q-value function (Sutton & Barto, 2018). However, such properties do not always hold under QH-discounting (Jaśkiewicz & Nowak, 2021). This can lead to a complicated agent behavior, as we now illustrate.

Consider the two-state Markov Decision Process (MDP) setup of Figure 1b, which is taken from (Jaśkiewicz & Nowak, 2021). Clearly, there are only two deterministic stationary policies here: $\bar{f} \equiv (f, f, \cdots)$ and $\bar{g} \equiv (g, g, \ldots)$, where f maps state 1 to action a_1, and g maps state 1 to action a_2, and both map state 2 to action a_1. For $\sigma = 0.5$ and $\gamma = 0.8$, their Q-value functions under QH-discounting are given in Table 1c. For a policy $\bar{\pi}$, its QH Q-value function, denoted by $Q_{\bar{\pi}}^{\sigma,\gamma}$, is defined in the same way as in the exponential discounting case, but with discount factors $1, \gamma, \gamma^2, \ldots$ replaced by $1, \sigma\gamma, \sigma\gamma^2, \ldots$ In Table 1c, notice that $\bar{f}$ yields the highest returns from state 2. However, at state 1, neither $\bar{f}$ nor $\bar{g}$ shares a greedy relationship with its QH Q-value function. This fact implies that $g\bar{f}$ (acting as g at $n = 0$ and f for $n \geq 1$) is the policy that yields the highest returns, starting from state 1.

The optimal policies in the above example have three interesting dissimilarities compared to their counterparts in RL with exponential discounting or simple averaging. Firstly, the optimal policy varies depending on the initial state of the process. Secondly, $g\bar{f}$ is non-stationary. Thirdly, and significantly, both $\bar{f}$ and $g\bar{f}$ display time inconsistency. To elaborate the last point further, note that both $g\bar{f}$ and $\bar{f}$ advocate following f at any $n \geq 1$. Now suppose, at time $n = 1$, the MDP is in state 1 and the agent re-evaluates the optimal policy from that time onward. Then, from Table 1c, the agent would again discover $g\bar{f}$ to be optimal, i.e., act as per g at $n = 1$ and revert to f thereafter. This behavior contradicts the one that is optimal from $n = 0$, highlighting the time inconsistency.

We now describe a complex agent behavior for the above setup, primarily stemming from the time-inconsistent nature of the optimal policies. Consider the agent as a sequence of selves, each corresponding to a different time step n. Suppose each future self is naive, i.e., unaware of the time inconsistency in the optimal policy. Alternatively, suppose they all have self-control issues and a possibility to act contrary to their own interests. In both scenarios, the following situation could unfold. Each time the MDP visits state 1, the naive selves recalculate the optimal strategy from that point on and decide to act as per g at that time step. Similarly, each self with control issues could also decide to act as per g because (i.) it is aware that if the subsequent selves act as per f, then it would receive higher returns, and (ii.) it presumes that the subsequent selves will act as per f. However, if all selves end up acting as per g for all $n \geq 0$, then Table 1c shows that the expected overall returns would be substantially lower (only 16.77).

To safeguard against the above pitfall, it is desirable to have a stationary (possibly stochastic) policy $\bar{\pi} \equiv (\pi, \pi, \ldots)$ from which there is no incentive for deviation. For such a policy $\bar{\pi}$, it would then be true that

$$Q_{\bar{\pi}}^{\sigma,\gamma}(s, \pi) = \sup_{\nu: \mathcal{S} \to \Delta(\mathcal{A})} Q_{\bar{\pi}}^{\sigma,\gamma}(s, \nu), \qquad s \in \mathcal{S}, \quad (1)$$

where $\mathcal{S}$ (resp. $\mathcal{A}$) is the MDP state space (resp. ac-

tion space), $\Delta(\mathcal{A})$ is the set of distributions on $\mathcal{A}$, and $Q^{\sigma,\gamma}_{\bar{\pi}}(s,\nu) = \sum_{a\in\mathcal{A}(s)} \nu(a|s) Q^{\sigma,\gamma}_{\bar{\pi}}(s,a)$ is the average of $\bar{\pi}$'s QH Q-values for a starting state distribution $\nu : \mathcal{S} \to \Delta(\mathcal{A})$. Any stationary policy $\bar{\pi}^*$ which satisfies (1) is referred to as a Markov Perfect Equilibrium (MPE) (Jaśkiewicz & Nowak, 2021). For our two-state MDP example, Table 1c shows that the stationary policy $\bar{h}$ is an MPE. For a general MDP, an MPE is neither guaranteed to exist nor be unique; so far, they have been found only using analytical techniques, and that too only for simple MDPs (Jaśkiewicz & Nowak, 2021).

Our goal in this work is to develop a model-free RL algorithm to identify an MPE in a finite state and action MDP. An MPE's existence here is guaranteed by the conditions outlined in (Jaśkiewicz & Nowak, 2021). However, finding it poses significant challenges unlike finding an optimal policy in RL with a traditional discounting. Firstly, there exists no known Bellman-type contraction mapping for which an MPE's Q-value function serves as a fixed point. Consequently, the traditional fixed-point-type methods cannot be used to find this value function. Moreover, even if this function were identified somehow, determining the MPE itself remains challenging as it lacks a straightforward relationship with its value function. Secondly, MPEs often are stochastic. This means the search space for an MPE encompasses all stochastic policies, which is infinitely large even for finite state and action MDPs. In contrast, under exponential discounting or average reward, the optimal policy search is confined to deterministic policies, which is a finite set (albeit growing combinatorially).

Our main contributions are as follows. We propose the first model-free RL algorithm for finding an MPE. This algorithm is a two-timescale stochastic approximation and is inspired by the recently proposed critic-actor method for classical RL (Bhatnagar et al., 2023). Unlike the actor update in the latter, which follows a stochastic estimate of the value function's gradient, our method updates along the advantage function[1], enabling it to find an MPE. Secondly, by building upon (Ramaswamy & Bhatnagar, 2016), (Gopalan & Thoppe, 2023), and (Bhatnagar et al., 2023), we conjecture that the iterates from our critic update converge to an invariant set of a suitable Differential Inclusion[2] (DI). We provide evidence supporting this claim in the context of the two-state MDP from Figure 1b. Thirdly, using the MPE's definition from (1), we show that any MPE's Q-value function must be an invariant set of the limiting DI for our critic update. Additionally, for cases where our critic-actor iterates converge to an isolated point (W,π), we show that π must be an MPE and W its Q-value function. Finally, we provide numerical experiments in an inventory control setup, demonstrating our algorithm's success in identifying various MPEs.

[1]The QH advantage function is the difference between the QH Q-value function and the state-value function. In RL with exponential discounting, the advantage function and the value function's gradient are aligned, but it is not so under QH discounting; see (5).

[2]A DI is a set-valued generalization of a differential equation. It has the form $\dot{x}(t) \in h(x(t))$, where h is set valued.

2. Setup, Goal, Algorithm, and Main Results

In this section, we describe our problem setup, our goal, and our key contributions: the first algorithm for finding an MPE and conjectures that describe its asymptotic convergence.

2.1. Setup and Goal

Let $\Delta(\mathcal{U})$ denote the set of distributions over a set $\mathcal{U}$. Our setup consists of an MDP $M \equiv (\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{R},\sigma,\gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and finite action spaces, respectively, $\mathcal{P} : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the transition matrix, and $r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is the instantaneous reward function. Further, $\sigma,\gamma \in [0,1)$ are the parameters of QH discounting. Within the above setup, our goal is to find an MPE, i.e., a stationary policy $\bar{\pi} \equiv (\pi,\pi,\ldots)$ (henceforth denoted only by π) that satisfies the MPE relation given in (1).

2.2. MPE-learning Algorithm

For a stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, let

$$Q^{\sigma,\gamma}_{\pi}(s,a) := r(s,a) + \mathbb{E}\left[\sum_{n=1}^{\infty} \sigma\gamma^n r(s_n,a_n) \middle| \begin{matrix} s_0 = s, \\ a_0 = a \end{matrix}\right], \quad (2)$$

where $s_{n+1} \sim \mathcal{P}(\cdot|s_n,a_n)$ and $a_{n+1} \sim \pi(\cdot|s_{n+1})$ for $n \geq 0$. This function is the stationary policy π's Q-value function under QH discounting. Separately, for $\theta \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$, let $\pi_\theta(\cdot|s) = \text{softmax}(\theta(s,\cdot))$ for $s \in \mathcal{S}$.

Our novel approach for finding an MPE is given in Algorithm 1. The symbol θ_n parameterizes the policy representing our MPE estimate at time $n \geq 0$, while W_n is this policy's Q-value function estimate. Hence, we refer to the θ_n update (Step 11) as the actor update, and to the W_n update (Step 10) as the critic update. Throughout this work, we focus on the scenario where the stepsizes α_n and β_n, used in the critic and actor updates, respectively, satisfy the relation $\lim_{n\to\infty} \alpha_n/\beta_n = 0$. This ensures the critic updates are on a slower timescale compared to the actor. Consequently, Algorithm 1 falls under the category of critic-actor algorithms (Bhatnagar et al., 2023) (instead of actor-critic). We give a principled motivation for our algorithm in Section 3.

2.3. Main Conjecture and Other Results

We first state our assumptions.

$\mathcal{A}_1$. **Stepsizes:** $(\alpha_n)_{n\geq 0}$ and $(\beta_n)_{n\geq 0}$ are two sequences

Algorithm 1 Synchronous MPE-learning

1: **Input:** Stepsizes (α_n), (β_n), and discount factors σ, γ
2: **Initialize:** $\theta_0, W_0 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$
3: **for** $n = 0, 1, 2, ...$ **do**
4: **for** $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
5: Observe $s' \sim \mathcal{P}(\cdot|s, a)$
6: Sample $a' \sim \pi_{\theta_n}(\cdot|s')$
7: $r'_n(s, a) \leftarrow r(s', a'), \quad W'_n(s, a) \leftarrow W_n(s', a')$
8: $\hat{W}_n^{\theta_n}(s, a) \leftarrow \langle \pi_{\theta_n}(\cdot|s),\ W_n(s, \cdot)\rangle$
9: **end for**
10: $W_{n+1} \leftarrow W_n + \alpha_n \left[r - (1 - \sigma)\gamma r'_n + \gamma W'_n - W_n\right]$
11: $\theta_{n+1} \leftarrow \theta_n + \beta_n \left[W_n - \hat{W}_n^{\theta_n}\right]$
12: **end for**

Steps 10 and 11 define the algorithm's update rules, while Steps 4 to 9 setup the necessary vectors for these updates.

of monotonically decreasing positive real numbers such that (i) $\alpha_0 \leq 1, \beta_0 \leq 1$; (ii) $\sum_{n=0}^{\infty} \alpha_n = \sum_{n=0}^{\infty} \beta_n = \infty$, but $\sum_{n=0}^{\infty}(\alpha_n^2 + \beta_n^2) < \infty$; and (iii) $\lim_{n\to\infty}(\alpha_n/\beta_n) = 0$.

$\mathcal{A}_2$. **Bounded reward:** There exists $r_{\max} > 0$ such that $|r(s, a)| < r_{\max}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Next, we define two set-valued maps. For $W \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, let $\lambda(W)$ be the (convex) set of stochastic policies given by

$$\lambda(W) := \Big\{\mu : \mathcal{S} \to \Delta(\mathcal{A}) : \sum_{a\in\mathcal{A}} \mu(a|s) = 1 \text{ and } \text{supp}(\mu(\cdot|s)) \subseteq \arg\max W(s, \cdot)\} \ \forall s \in \mathcal{S}\Big\}.$$

Further, let $T_\lambda(W) := \{T^\mu(W) : \mu \in \lambda(W)\}$, where $T^g : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is the QH Bellman operator for the policy μ. That is,

$$T^\mu(W)(s, a) = r(s, a) + \gamma \sum_{s',a'} \mathcal{P}(s'|s, a)\mu(a'|s') \times \left[-(1 - \sigma)r(s', a') + W(s', a')\right]. \quad (3)$$

We now use DI theory (Aubin & Cellina, 2012) to explain the limiting dynamics of the (W_n) and (θ_n) iterates from Algorithm 1. Our key reason for relying on DIs is that, asymptotically, W_n's update function has discontinuities. This discontinuity arises because, as we discuss in Section 3, (W_n)'s asymptotic dynamics must be governed by the $T^\mu(W) - W$ operator for some $\mu \in \lambda(W)$ and the T^μ operator discontinuously changes. A DI helps in handing these discontinuities by allowing multiple update directions at those points (cf. (Gopalan & Thoppe, 2023)). This set-valued nature, though, implies that a DI can have multiple solutions for the same initial point.

With respect to the DI $\dot{x} \in h(x(t))$, we will say a set $\Gamma \subseteq \mathbb{R}^d$ is *invariant* if, for every $x_0 \in \Gamma$, there is *some* solution trajectory $(x(t))_{t\in(-\infty,\infty)}$ of the DI with $x(0) = x_0$ that lies entirely in Γ. An invariant set Γ is additionally *internally chain transitive* if it is compact and connected in a certain way: for $y, y' \in \Gamma$, $\nu > 0$, and $T > 0$, there exist $m \geq 1$ and points $z_0 = y, z_1, \ldots, z_{m-1}, z_m = y'$ in Γ such that a solution trajectory of the DI initiated at z_i meets the ν-neighborhood of z_{i+1} for $0 \leq i < m$ after a time that is equal or larger than T.

Our main conjecture can now be stated as follows. Let $\|\cdot\|$ be the standard ℓ_∞ norm.

Conjecture 2.1. *Suppose $\mathcal{A}_1$ and $\mathcal{A}_2$ are true. Then the following statements hold for the iterates (W_n) and (θ_n) obtained from Algorithm 1:*

(i) *(W_n) is stable, i.e., $\sup_n \|W_n\| < \infty$ a.s.;*

(ii) *$(W_n, \pi_{\theta_n}) \to \{(W, \lambda(W)) : W \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}\}$; and*

(iii) *(W_n) converges to a compact connected internally chain transitive invariant set of the DI*

$$\dot{W}(t) \in T_\lambda(W(t)) - W(t). \quad (4)$$

Remark 2.2. Statement (ii) shows that π_{θ_n} is asymptotically greedy with respect to W_n, while Statement (iii) shows that W_n converges to a suitable invariant set of the DI in (4).

Our next conjecture discusses an MPE's relation to (4).

Conjecture 2.3. *For any MPE π^*, we have $\pi_* \in \lambda(Q^{\sigma,\gamma}_{\pi^*})$ which, in turn, implies that $0 \in T_\lambda(Q^{\sigma,\gamma}_{\pi^*}) - Q^{\sigma,\gamma}_{\pi^*}$.*

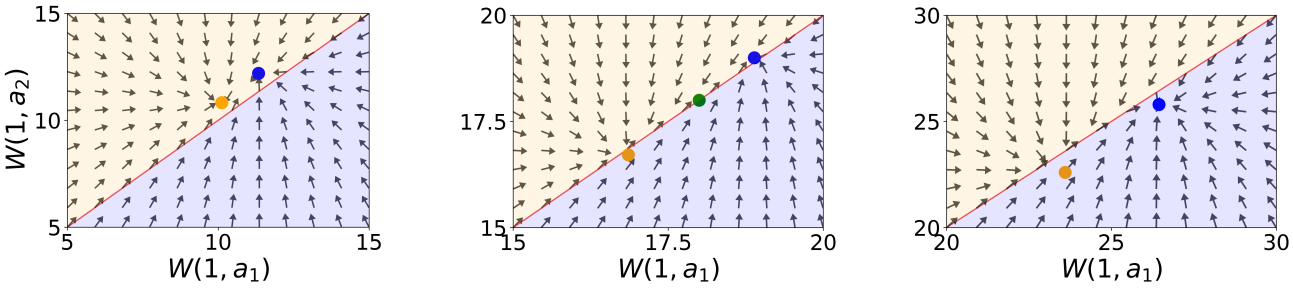
In general, our algorithm can converge to a set. However, our next conjecture shows that our algorithm's convergence to a singleton implies that the latter must define an MPE.

Conjecture 2.4. *If Algorithm 1 converges to a point, i.e., if $(W_n, \pi_{\theta_n}) \to (W^*, \pi^*)$, then π^* must be an MPE and W^* must be the QH Q-value function of π^*.*

3. Our Algorithm Design

Here we motivate the critic (Step 10) and actor (Step 11) update rules of our proposed algorithm and explain how they enable MPE estimation.

Our critic or the W_n update step is based on the temporal difference idea for minimizing the QH Bellman error at time n, i.e., $\|T^{\pi_{\theta_n}}(W) - W\|$. Hence, W_n can be seen as an estimate of the QH Q-value function of the policy π_{θ_n}.



(a) $\sigma = 0.3$ (b) $\sigma = 0.5$ (c) $\sigma = 0.7$

Figure 2: Vector fields for the DI in (4) for the MDP given in Figure 1b with $\gamma = 0.8$ and σ values of $0.3, 0.5$, and 0.7. Here, the orange dot is the vector $Q^{\sigma,\gamma}_{\bar{g}}$, blue is $Q^{\sigma,\gamma}_{\bar{f}}$, while green is $Q^{\sigma,\gamma}_{\bar{h}}$, where $\bar{f}, \bar{g}$ and $\bar{h}$ are as in Figure 1's caption.

Our actor or the θ_n update is approximately along the QH advantage function $A^{\sigma,\gamma}_{\pi_{\theta_n}}$ of π_{θ_n}, where $A^{\sigma,\gamma}_{\pi}(s,a) = Q^{\sigma,\gamma}_{\pi}(s,a) - \langle \pi(\cdot|s), Q^{\sigma,\gamma}_{\pi}(s,\cdot)\rangle$ for any π, s, and a. We say approximately because W_n is only an estimate of $Q^{\sigma,\gamma}_{\pi_{\theta_n}}$. Our main motivation to use the advantage function for updating θ_n is to ensure that π_{θ_n} is asymptotically greedy with respect to W_n; see Conjecture 2.1.(ii). Even in RL with exponential discounting, the actor updates are along the corresponding advantage function estimate of the current policy (Sutton et al., 1999). However, the advantage function there aligns with the gradient of the state value function and enables discovery of the optimal policy. In QH discounting, though, this alignment does not hold, as we show next.

Let $A^{\gamma}_{\pi}(s,a) = Q^{\gamma}_{\pi}(s,a) - \langle \pi(\cdot|s),\ Q^{\gamma}_{\pi}(s,\cdot)\rangle$ be the policy π's advantage function under exponential discounting with γ discount factor. Similarly, let $A^{0}_{\pi}(s,a)$ be the analogous $\gamma = 0$ expression. Now, if $\eta^{\sigma,\gamma}_{\pi_\theta}(\mu) := \mathbb{E}_{s\sim\mu, a\sim\pi_\theta}[Q^{\sigma,\gamma}_{\pi_\theta}(s,a)]$ for $\theta \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$, where μ is some fixed initial state distribution, then we have that

$$\frac{\partial \eta^{\sigma,\gamma}_{\pi_\theta}(\mu)}{\partial \theta(s,a)} = (1-\sigma)\mu(s)\pi_\theta(a|s)A^{0}_{\pi_\theta}(s,a) + \frac{\sigma}{1-\gamma} d^{\pi_\theta}_{\mu}(s)\pi_\theta(a|s)A^{\gamma}_{\pi_\theta}(s,a). \tag{5}$$

The RHS is not $A^{\sigma,\gamma}_{\pi_\theta}$, as claimed above, since

$$A^{\sigma,\gamma}_{\pi_\theta}(s,a) = (1-\sigma)A^{0}_{\pi_\theta}(s,a) + \sigma A^{\gamma}_{\pi_\theta}(s,a),$$

which itself holds since

$$Q^{\sigma,\gamma}_{\pi_\theta}(s,a) = \mathbb{E}\left[(1-\sigma)r(s_0,a_0) + \sum_{n=0}^{\infty}\sigma\gamma^n r(s_n,a_n)\middle|\begin{matrix} s_0 = s, \\ a_0 = a\end{matrix}\right].$$

Our actor update not being aligned with the gradient also explains why our algorithm does not track the optimal policy unlike the classical critic-actor method.

Now, because of two-timescale nature of our algorithm, it can be shown that (W_n) tracks the DI in (4). The main advantage of this DI is that an MPE's QH Q-value function is a zero of this DI, as shown in Proposition 2.3.

Finally, we explain our reasons for relying on the critic-actor family for developing our MPE learning algorithm, instead of extending value-function-based methods such as Q-learning and SARSA. In classical RL, these latter methods leverage the fact that the optimal policy can be derived from its Q-value function through a simple greedy relationship, making it sufficient to estimate only this Q-value function. However, under QH discounting, no relation exists to infer an MPE from its QH Q-value function. This is why we directly use the critic-actor family: it enables simultaneous estimation of an MPE and its Q-value function.

4. Proof Sketches for our Conjectures

We provide a brief overview of our planned approaches to prove our various conjectures.

4.1. Conjecture 2.1's Proof Sketch and its Utility

Using $\mathcal{A}_2$ and the fact that $\gamma \in [0,1)$, one should be able to inductively show that $\|W_n\| \leq C, n \geq 0$, for some constant $C \geq 0$. This would establish Statement (i).

Statements (ii) and (iii) should follow by building upon the two-timescale stochastic approximation analyses presented in (Borkar, 2009), (Ramaswamy & Bhatnagar, 2016), and (Yaji & Bhatnagar, 2020). All these prior analyses assume the iterates in both timescales to be bounded a.s. However, in our case, the (θ_n) iterates must diverge to infinity for π_{θ_n} to become asymptotically greedy with respect to W_n and transform into an MPE. While this divergence may force non-trivial modifications to the proof used in the above papers, we believe the core ideas and the main conclusions should still carry over mutatis mutandis.

We now illustrate the utility of the limiting DI in (4) in finding an MPE for the two-state MDP given in Figure 1b. In this case, the vector field associated with the above

DI for $\gamma = 0.8$ and three different values of σ ($0.3, 0.5$ and 0.7) is given in Figure 2. We first explain the three plots. In all of them, there is a blue and an orange colored region. These are the greedy regions associated with the policies $\bar{f}$ and $\bar{g}$, respectively (see Figure 1's caption for $\bar{f}$ and $\bar{g}$'s definition). That is, for any vector W in the blue region $W(1, f(1)) = W(1, a_1) \geq W(1, a_2) = W(1, g(1))$ and the reverse holds for any vector in the orange region. Hence, for any W in the interior of the blue (resp. orange) region, $\lambda(W)$ consists[3] of only f (resp. g) and the driving function is $T^f(W) - W$ (resp. $T^g(W) - W$), where T^f and T^g are defined as in (3). Because the greedy policy is different in the two colored regions, the local dynamics discontinuously changes across the $x = y$ boundary line. As pointed out before, the DI in (4) handles this discontinuity by allowing both the update directions (and also their convex combinations) at the boundary.

In Figure 2(a), any solution trajectory starting in the blue region is driven towards the blue dot, which represents $Q^{\sigma,\gamma}_{\bar{f}}$. This is not surprising since we use QH temporal difference learning for updating (W_n). However, once the trajectory crosses over to the orange region, the driving function changes and the trajectory is now driven towards the orange dot, which represents $Q^{\sigma,\gamma}_{\bar{f}}$. For $\sigma = 0.3$, it can be shown that $\bar{g}$ is the only MPE and tracking the solution trajectory of our DI helps in finding this MPE's QH Q-value. Figure 2(c) can be interpreted similarly. In Figure 2(b), i.e., for the case $\sigma = 0.5$, it can be shown that neither $\bar{g}$ nor $\bar{f}$ is an MPE. Instead, the stochastic policy $\bar{h}$ (see Figure 1's caption) is an MPE and its Q-value sits on the boundary. In this case, the driving function in either region pushes the solution trajectory towards the other which eventually forces it to converge to the green dot, which is $Q^{\sigma,\gamma}_{\bar{h}}$. Thus, tracking the solution trajectories of our DI again helps.

4.2. Conjecture 2.3's Proof Sketch

Our planned proof strategy is the following. We plan to use the definition of an MPE to show that $\pi_* \in \lambda(Q^{\sigma,\gamma}_{\pi_*})$. The fact that $0 \in T_\lambda(Q^{\sigma,\gamma}_{\pi^*}) - Q^{\sigma,\gamma}_{\pi^*}$ then follows immediately.

4.3. Conjecture 2.4's Proof Sketch

If Conjecture 2.1 is true, then it would follow from its Statement (ii) that, on every sample path where $(W_n, \pi_{\theta_n}) \to (W^*, \pi_*)$, we would have that $\pi_* \in \lambda(W^*)$. Using this fact, we then plan to show that $W^* = Q^{\sigma,\lambda}_{\pi_*}$. The fact that π_* is an MPE would then follow by verifying the MPE definition given in (1).

[3]We mean $\lambda(W)$ contains the stochastic representation of f.

5. Experiments

We now numerically illustrate the utility of our algorithm for the stochastic inventory control problem.

The inventory control problem involves managing stock levels, e.g., cars in a showroom, to meet the daily uncertain demand while maximizing overall profits. For our illustration, we consider an inventory system with a maximum capacity of $M = 2$. We suppose that the procurement (resp. holding) cost per item is $c = 500$ (resp. $h = 50$), while the selling price is $p = 900$. Further, we suppose that the daily demand is a random variable taking values of $0, 1,$ or 2 with probabilities 0.3, 0.2, and 0.5, respectively. At the start of day n, the inventory manager gets to see the current stock level s_n and then decide on the number of new items a_n to (immediately) procure to meet the (uncertain) demand d_n for that day; the capacity constraint implies that $s_n + a_n$ can be at most 2. Hence, the reward obtained for day n equals $r_n(s_n, a_n) = 900 \times \min\{s_n + a_n, d_n\} - 500 \times a_n - 50 \times \max\{s_n + a_n - d_n, 0\}$. Consequently, the expected infinite horizon QH-discounted cost, starting with an initial stock of s and initial procurement of a equals $\mathbb{E}[r_0(s_0, a_0) + \sum_{n=1}^{\infty} \sigma\gamma^n r_n(s_n, a_n) | s_0 = s, a_0 = a]$. For our illustration, we suppose $\sigma = 0.3$ and $\gamma = 0.9$.

We ran our proposed algorithm multiple times and it identified three different MPEs. Due to space limitation, we give details of only one of these MPEs. The MPE is

$$\pi^*_{MPE_1} = \begin{bmatrix} 0.00 & 0.53 & 0.47 \\ 0.53 & 0.47 & - \\ 1.00 & - & - \end{bmatrix},$$

while its QH Q-value function is as given in Table 1. See Appendix A for the details of other MPEs. Separately, we also ran the variant of the classical critic-actor method for QH discounting, the one where the actor update is along the gradient of the state value function as described in (5) (with the initial state distribution being uniform). The output of this algorithm was the policy

$$\pi^*_{naive} = \begin{bmatrix} 0.0 & 0.0 & 1.0 \\ 0.0 & 1.0 & - \\ 1.0 & - & - \end{bmatrix},$$

whose QH Q-value function is as given in Table 3. Clearly, under $\pi^*_{MPE_1}$, the inventory manager has no incentive to deviate. In contrast, π^*_{naive} provides an incentive to deviate: maintaining a stock of 1 now and keeping it at 2, thereafter, is better than keeping it at 2 always, i.e., following π^*_{naive} always. Hence, the naive agent can end up maintaining the stock at 1 always, i.e., it may end up following

$$\pi_{naive} = \begin{bmatrix} 0.0 & 1.0 & 0.0 \\ 1.0 & 0.0 & - \\ 1.0 & - & - \end{bmatrix}$$

Table 1: $Q^{\sigma,\gamma}_{\pi^*_{MPE_1}}$ values

s \ a	0	1	2
0	897.5	**1053**	**1053**
1	**1553**	**1553**	-
2	**2053**	-	-

Table 2: $Q^{\sigma,\gamma}_{\pi_{naive}}$ values

s \ a	0	1	2
0	675	**830.5**	839.6
1	**1330.5**	1339.6	-
2	**1839.6**	-	-

Table 3: $Q^{\sigma,\gamma}_{\pi^*_{naive}}$ value

s \ a	0	1	2
0	1080	<u>1235.5</u>	**1228**
1	<u>1735.5</u>	**1728**	-
2	**2228**	-	-

The numbers in bold represent the $Q^{\sigma,\gamma}$-values for actions recommended by the respective policies. In MPE $\pi^*_{MPE_1}$, no other actions have higher values than those in bold, so the agent has no incentive to deviate from $\pi^*_{MPE_1}$. In contrast, under π^*_{naive} from the Vanilla QH Policy Gradient Algorithm, the underlined actions have higher values, hence an agent may deviate from π^*_{naive}.

instead of π^*_{naive}. In that case, the agent gets significantly low returns as can be seen from Table 2.

Acknowledgement

Eshwar S. R. was supported by the Prime Minister's Research Fellowship (PMRF), while Nibedita Roy was supported in part by the Walmart Centre for Tech Excellence. Gugan Thoppe's research was supported in part by DST-SERB's Core Research Grant CRG/2021/008330, the Indo-French Centre for the Promotion of Advanced Research—CEFIPRA (7102-1), the Walmart Center for Tech Excellence, the Kotak-IISc AI/ML Centre, and the Pratiksha Trust Young Investigator Award.

References

Ainslie, G. Specious reward: a behavioral theory of impulsiveness and impulse control. *Psychological bulletin*, 82 (4):463, 1975.

Aubin, J. and Cellina, A. Differential inclusions: Set-valued maps and viability theory, springer science & business media. *Berlin, Germany*, 2012.

Bertsekas, D. Reinforcement learning and optimal control. *Athena Scientific*, 2019.

Bhatnagar, S., Borkar, V. S., and Guin, S. Actor-critic or critic-actor? a tale of two time scales. *IEEE Control Systems Letters*, 2023.

Borkar, V. S. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Cropper, M. L., Aydede, S. K., and Portney, P. R. Rates of time preference for saving lives. *The American Economic Review*, 82(2):469–472, 1992.

Dhami, S. *The foundations of behavioral economic analysis*. Oxford University Press, 2016.

Frederick, S., Loewenstein, G., and O'donoghue, T. Time discounting and time preference: A critical review. *Journal of economic literature*, 40(2):351–401, 2002.

Gopalan, A. and Thoppe, G. Demystifying approximate value-based rl with ϵ-greedy exploration: A differential inclusion view, 2023. URL https://arxiv.org/abs/2205.13617.

Jaśkiewicz, A. and Nowak, A. S. Markov decision processes with quasi- hyperbolic discounting. *Finance and Stochastics*, 25(2):189–229, 2021.

Laibson, D. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478, 1997.

Loewenstein, G. and Prelec, D. Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics*, 107(2):573–597, 1992.

Phelps, E. and Pollak, R. On second-best national saving and game-equilibrium growth. *The Review of Economic Studies*, 35(2):185, 1968.

Ramaswamy, A. and Bhatnagar, S. Stochastic recursive inclusion in two timescales with an application to the lagrangian dual problem. *Stochastics*, 88(8):1173–1187, 2016.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Thaler, R. Some empirical evidence on dynamic inconsistency. *Economics letters*, 8(3):201–207, 1981.

Yaji, V. G. and Bhatnagar, S. Stochastic recursive inclusions in two timescales with nonadditive iterate-dependent markov noise. *Mathematics of Operations Research*, 45 (4):1405–1444, 2020.

A. Details of MPE's for Inventory Control Problem

For the inventory control setup considered in Section 5, our Algorithm 1 identified two additional MPEs,

$$\pi^*_{MPE_2} = \begin{bmatrix} 0.0 & 0.8 & 0.2 \\ 0.0 & 1.0 & - \\ 1.0 & - & - \end{bmatrix}, \pi^*_{MPE_3} = \begin{bmatrix} 0.0 & 0.3 & 0.7 \\ 1.0 & 0.0 & - \\ 1.0 & - & - \end{bmatrix}.$$

The corresponding QH Q-values are presented in Tables 4 and 5, respectively.

Table 4: $Q^{\sigma,\gamma}_{\pi^*_{MPE_2}}$ values

s \ a	0	1	2
0	873.75	**1040.5**	**1040.5**
1	1540.5	**1540.5**	-
2	**2040.5**	-	-

Table 5: $Q^{\sigma,\gamma}_{\pi^*_{MPE_3}}$ values

s \ a	0	1	2
0	918.64	**1064.125**	**1064.125**
1	**1564.125**	1564.125	-
2	**2064.125**	-	-