2024 Tenth Indian Control Conference (ICC)
9 - 11 December, 2024
IISER Bhopal, India

Reinforcement Learning with Quasi-Hyperbolic Discounting (Extended Abstract)

S.R. Eshwar Mayank Motwani Nibedita Roy Gugan Thoppe

***Abstract*— Human decision-making is better modeled via discounting schemes such as quasi-hyperbolic and hyperbolic than classical ones such as exponential or average reward. In [3], we initiated the study of reinforcement learning with Quasi-Hyperbolic (QH) discounting and proposed the first model-free algorithm for finding a type of policy called the Markov perfect equilibrium. This paper provides a summary of that algorithm and our other findings on its convergence.**

I. Introduction

Humans tend to be more impatient when making short-term decisions, while exhibiting greater patience over longer time horizons [2]. In the behavioral sciences, this shift in impatience is known to be better captured by hyperbolic and quasi-hyperbolic (QH) discounting models, as opposed to classical models like exponential discounting or average reward [2]. Despite this, much of the Reinforcement Learning (RL) literature has remained focused on classical discounting approaches. In [3], we initiated the study of RL with QH discounting, whose findings we summarize in this article.

We first define QH discounting [7], [6] and then discuss its non-trivial implications in RL. Recall that, in RL, an agent interacts with a system via a sequence (a_n) of actions. After each action a_n, the system's state transitions from s_n to s_{n+1}, and the agent receives a reward r_n. The overall quality of the action sequence (a_n) is evaluated through the expectation of the infinite-horizon discounted sum, $\sum_{n=0}^{\infty} d_n r_n$. Here, d_n represents the discount factor, indicating the perceived value at time 0 of 1 unit of reward received after the n-th interaction. In classical exponential discounting, $d_n = \gamma^n$ for some fixed $\gamma \in [0, 1)$. In contrast, QH discounting sets $d_n = 1$ for $n = 0$, and $d_n = \sigma\gamma^n$ for $n \geq 1$. Thus, QH discounting is a two-parameter model, with $\sigma \in [0, 1]$ representing the short-term discount factor and $\gamma \in [0, 1)$ the long-term one. Their interplay allows us to model the variability in the agent's impatience levels [2].

The foremost implication of QH discounting in RL is that an optimal policy is no longer guaranteed to be time consistent. That is, if π_* is a policy that is optimal from some time t_1, then the action sequence prescribed by it for time t_2 onwards need not be optimal for the horizon starting from t_2. Therefore, 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 avoid the scenario described above, it is advisable to design policies that eliminate any incentive for deviation, even when the agent may lack self-control. Such a policy is referred to as a Markov Perfect Equilibrium (MPE) [5]. In general, for Markov Decision Processes (MDPs) under quasi-hyperbolic (QH) discounting, the existence and uniqueness of an MPE are not guaranteed. Where they do exist, to the best of our knowledge, only analytical methods have been employed to find them [5], and these methods typically assume complete information about the MDP.

Our goal in [3] was to design a model-free RL algorithm to find an MPE in a finite state, finite action MDP. Such a method would eliminate the need for prior knowledge of the MDP transition probabilities, and instead rely on samples of state-action-next state tuples and their associated rewards. We note that the existence of a stationary MPE in finite state-action MDPs is guaranteed [5].

The challenges in finding an MPE in the framework of QH discounting are (i) there is no known Bellman-type contraction operator for which an MPE's value function is a fixed point; (ii) MPEs may be stochastic, increasing the search space compared to traditional discounting where optimal policies are deterministic; and (iii) MPEs (and also their value functions) need not be unique.

Our main contributions in [3] are as follows:

1) We present the first model-free RL algorithm for finding an MPE in the QH discounting framework.
2) We provide guarantees that if the algorithm converges to an isolated point, then it must be an MPE.
3) We provide numerical experiments on the standard inventory control problem to showcase the algorithm's effectiveness in identifying multiple MPEs.

In the rest of the paper, we provide a brief overview of these key contributions.

II. Setup and Goal

We begin by describing our setup and our goal.

Let $\Delta(\mathcal{U})$ denote the set of all probability distributions over a set $\mathcal{U}$. The Markov Decision Process (MDP) considered in this work is represented as $M \equiv (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \sigma, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces, respectively. The function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ defines the

S.R. Eshwar, Nibedita Roy, and Gugan Thoppe are with Department of Computer Science and Automation, Indian Institute of Science, Bengaluru, India. S.R. Eshwar was supported by the Prime Minister's Research Fellowship (PMRF). Nibedita Roy and Gugan Thoppe were supported in part by the Walmart Centre for Tech Excellence at IISc (CSR Grant WMGT-23-0001) and the Indo-French Centre for the Promotion of Advanced Research—CEFIPRA (7102-1). Gugan Thoppe's research was also supported in part by DST-SERB's Core Research Grant (CRG/2021/008330), and the Pratiksha Trust Young Investigator Award. Email: eshwarsr@iisc.ac.in, nibeditaroy@iisc.ac.in, gthoppe@iisc.ac.in

Mayank Motwani is with Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, India Email: 22b1052@iitb.ac.in

transition probability matrix, while $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the immediate reward function. The parameters $\sigma \in [0, 1]$ and $\gamma \in [0, 1)$ are as defined in Section I.

Next, we define an MPE [5]. For a stationary policy π, let

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

be its QH Q-value function starting from the state-action pair (s_0, a_0), where $s_{n+1} \sim \mathcal{P}(\cdot|s_n, a_n)$ and $a_{n+1} \sim \pi(\cdot|s_{n+1})$. This policy is said to be an MPE if it satisfies

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

where $Q_\pi^{\sigma,\gamma}(s, \nu) = \sum_{a\in\mathcal{A}(s)} \nu(a|s) Q_\pi^{\sigma,\gamma}(s, a)$ is the average of π's QH Q-values under a starting action distribution $\nu : \mathcal{S} \to \Delta(\mathcal{A})$.

Our goal formally is to design a model-free RL algorithm for finding an MPE, i.e., a policy π that satisfies (1).

III. Key Contributions: MPE-learning Algorithm and its Convergence

Our proposed MPE learning algorithm has the following update rule at iteration $n \geq 0$: for all (s, a),

$$\begin{aligned} W_{n+1}(s, a) &\leftarrow W_n(s, a) + \alpha_n \Big[r(s, a) \\ &\qquad - (1 - \sigma)\gamma r(s', a') + \gamma W_n(s', a') \Big], \\ \theta_{n+1}(s, a) &\leftarrow \theta_n(s, a) + \beta_n \Big[W_n(s, a) \\ &\qquad - \sum_{\overline{a}} W_n(s, \overline{a}) \pi_{\theta_n}(\overline{a}|s) \Big], \end{aligned} \tag{2}$$

where $s' \sim \mathcal{P}(\cdot|s, a)$ and $a' \sim \pi_{\theta_n}(\cdot|s')$. The policy $\pi_{\theta_n} = \text{softmax}(\theta_n)$, with $\theta_n \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$, represents our current estimate of an MPE, while the vector $W_n \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ can be seen as an estimate of π_{θ_n}'s QH Q-value function. We presume that the stepsize sequences (α_n) and (β_n) satisfy $\alpha_n/\beta_n \to 0$. Hence, our method falls within the family of critic-actor methods.

Our proposed algorithm is similar in form to the critic-actor method for traditional RL with exponential discounting [1]. In that, the actor updates along the QH advantage function estimate, and the critic tries to minimize a suitable TD error defined in the QH context. The reasons for using these update rules in our work [3], though, differ significantly from those in [1]. In the latter, the actor's update rule is along the advantage function since it aligns with the gradient of the value function ([8]). However, in the QH case, a policy π_θ's advantage function and its value function's gradient with respect to θ are not aligned. Hence, unlike in classical RL, our actor-update's purpose is not to improve the policy in terms of its QH value function. Instead, it is to ensure that our policy is asymptotically greedy with respect to the limit of the (W_n) sequence whenever such a limit exists. Additional details on these remarks can be found in [3].

Our main theorem from [3] on the convergence of our proposed MPE learning algorithm can be stated as follows.

Let $\lambda(W)$ be the set of stochastic policies given by

$$\begin{aligned} \lambda(W) := \Big\{ \pi : \mathcal{S} \to \Delta(\mathcal{A}) : \sum_{a\in\mathcal{A}} \pi(a|s) = 1 \text{ and} \\ \text{supp}(\pi(\cdot|s)) \subseteq \arg\max W(s, \cdot) \ \forall s \in \mathcal{S} \Big\}. \end{aligned} \tag{3}$$

Theorem 3.1: Suppose α_n and β_n satisfy the standard Robbins-Monro condition with $\lim_{n\to\infty} \alpha_n/\beta_n = 0$, and rewards are bounded (i.e., there exists $r_{\max} > 0$ such that $|r(s, a)| \leq r_{\max}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$). Then the following statements hold for the iterates (W_n) and (θ_n) obtained from our algorithm (2):

(i) (W_n) is stable, i.e., $\sup_n \|W_n\|_\infty < \infty$ a.s.;
(ii) If (W_n) converges to W_*, then $\pi_{\theta_n} \to \lambda(W_*)$; and
(iii) 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^*.

This result can be interpreted as follows. The first claim is that our (W_n) iterates are bounded almost surely. The second guarantees that, whenever (W_n) converges to a limit W_*, the limit of π_{θ_n} must be greedy with respect to W_*. Our last and final statement shows that, whenever our algorithm converges, the limit of the (π_{θ_n}) sequence must be an MPE and that of (W_n) must be this MPE's QH Q-value function.

The detailed proof of the above result can be found [4].

Finally, in the numerical illustrations section in [3], we demonstrated the utility of our algorithm in identifying various MPEs for the standard stochastic inventory control problem. Alongside and in contrast, we showed that the classical critic-actor method can help in estimating only the optimal policy under QH discounting. The optimal policy we identified, in fact, provided incentives for deviation. For this policy, we quantified the sub-optimality of returns for the case where the agent has self-control issues.

References

[1] S. Bhatnagar, V. S. Borkar, and S. Guin. Actor-critic or critic-actor? a tale of two time scales. *IEEE Control Systems Letters*, 2023.
[2] S. Dhami. *The foundations of behavioral economic analysis*. Oxford University Press, 2016.
[3] S. R. Eshwar, M. Motwani, N. Roy, and G. Thoppe. Reinforcement learning with quasi-hyperbolic discounting, 2024. Accepted for publication in 63rd IEEE Conference on Decision and Control (CDC).
[4] S. R. Eshwar, M. Motwani, N. Roy, and G. Thoppe. Reinforcement learning with quasi-hyperbolic discounting. *arxiv 2409.10583*, 2024.
[5] A. Jaśkiewicz and A. S. Nowak. Markov decision processes with quasi-hyperbolic discounting. *Finance and Stochastics*, 25(2):189–229, 2021.
[6] D. Laibson. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478, 1997.
[7] E. Phelps and R. Pollak. On second-best national saving and game-equilibrium growth. *The Review of Economic Studies*, 35(2):185, 1968.
[8] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.