Reinforcement Learning with Quasi-Hyperbolic Discounting

Anonymous Authors[1]

Abstract

Reinforcement learning has traditionally been studied with exponential discounting or the average rewards setup, mainly due to their mathematical tractability. However, such frameworks fall short of accurately capturing human behavior, which often 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 then formally show that any MPE would be an invariant set of our identified DI. Finally, we validate our findings numerically for the standard inventory system with stochastic demands.

1. Introduction

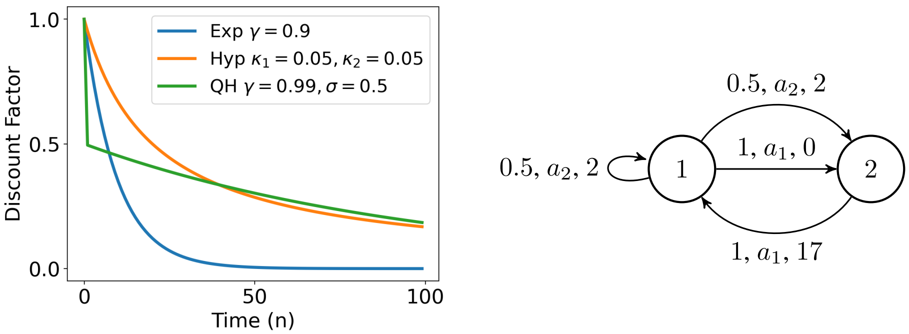
Reinforcement Learning (RL) (Sutton & Barto., 2018; Bertsekas, 2019) looks at identifying a policy/ strategy for an agent to optimally solve a task with sequential decisions. Since ages, a strategy $\bar{\pi}$'s optimality has been decided based on either the expected exponentially discounted sum or the long-term average of the 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. Under both discounting schemes, the optimal policy is time-consistent, i.e., the optimal policy starting from time t remains optimal when reconsidered from some later time as well. Despite their long history, evidence is now growing that these types of optimal policies 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 optimal policy under the two traditional models because of their time-consistent nature, which demonstrates the limitations of these 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 under this discounting model is $\sum_{n=0}^{\infty} b_n r_n$, where $b_n = (1 + \kappa_1 n)^{-\kappa_2/\kappa_1}$ for some $\kappa_1, \kappa_2 > 0$. This form of discounting is 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 $\sigma \in [0, 1]$. 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 discounting rates under exponential, hyperbolic, and quasi-hyperbolic discounting is given in Figure 1a. Unlike exponential discounting, note that there is a sharp decrease 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.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the Workshop on Foundations of Reinforcement Learning and Control at the International Conference on Machine Learning (ICML). Do not distribute.



(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 policies $\bar{f}, \bar{g}$, and $\bar{h}$. The rows refer to the (s, a) pairs. The columns represent the corresponding policies.

Under exponential discounting, the optimal policy $\bar{\pi}_*$ is deterministic and stationary, sharing a greedy relationship with its Q-value function (Sutton & Barto., 2018). However, this relationship does not hold under QH-discounting, leading to non-trivial behaviors. We illustrate this fact using a simple two-state Markov Decision Process (MDP) setup in Figure 1b from (Jaśkiewicz & Nowak, 2021). There are two deterministic policies possible for this MDP: f, which maps state 1 to action a_1, and g, which maps state 1 to action a_2; both map state 2 to action a_1. For $\sigma = 0.5$ and $\gamma = 0.8$, observe that the stationary policies $\bar{f}$ and $\bar{g}$, acting according to f and g respectively at all times, do not share a greedy relationship with their Q-value functions under QH-discounting (see Table 1c). Hence, they are not optimal. Instead, the non-stationary policy $g\bar{f}$ (acting as g at $n = 0$ and f for $n \geq 1$) yields the highest returns from state 1, while $\bar{f}$ is optimal when starting from state 2. The policy $f\bar{g}$ is time-inconsistent, meaning that re-evaluating the optimal policy at state 1 results in the agent following g for that time instant (as $g\bar{f}$ is optimal in state 2). This repeats at each time instant when agent reaches state 1, resulting in following $\bar{g}$ in state 1 everytime, which is suboptimal and leads to significantly lower returns.

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

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

where $Q_{\bar{\pi}}^{\sigma,\gamma}(s, \pi) = \sum_{a \in \mathcal{A}(s)} \pi(a|s) Q_{\bar{\pi}}^{\sigma,\gamma}(s, a)$. 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, it follows from Table 1c that the stationary policy $\bar{h}$, where $\bar{h}(a_1|1) = \bar{h}(a_2|2) = 0.5$, is an MPE. For a general MDP and a general value of σ, an MPE is neither guaranteed to exist nor be unique. So far, MPEs have been found only using analytical techniques, and that too only for simple MDPs.

Our goal in this work is to design a model-free RL algorithm for finding an MPE in an MDP with finite states and finite actions. The sufficient conditions for MPE existence which are stated in (Jaśkiewicz & Nowak, 2021) trivially hold in this setting and, hence, an MPE is guaranteed. Nevertheless, finding such an MPE poses the following challenges. One, there is *no known Bellman-type* operator for which an MPE is a fixed point. Hence, fixed-point-type iterations cannot be used to find an MPE. Two, an MPE is often *stochastic*. This implies that there are infinitely many candidate solutions even with finite state and finite actions. Thus, the goal of finding an MPE in the QH setting is not equivalent to finding the optimal policy in the exponential setting.

Our main contributions can be summarized as follows. We provide the first model-free RL algorithm for finding an MPE. Using a two-timescale analysis based on (Gopalan & Thoppe, 2022) and (Ramaswamy & Bhatnagar, 2016), we provide evidence to show that our algorithm converges to an invariant set of a suitable Differential Inclusion (DI), a standard analysis tool used in control theory. Thereafter, we formally show that any MPE must be an invariant set of our DI. Finally, we provide numerical experiments in an inventory control setup to show that our algorithm succeeds in extracting out the various MPEs.

2. Setup, Algorithm, and Main Result

2.1. Setup and Algorithm

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 some 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. The definition of Q-value function of a stationary policy[1] $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ under QH discounting is given by

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

where, for $n \geq 0$, $s_{n+1} \sim \mathcal{P}(\cdot|s_n, a_n)$, and $a_{n+1} \sim \pi(\cdot|a_n)$. In this work, we consider a policy π parameterized by θ where $\pi(.|s) = softmax(\theta(s,.))$. We refer the reader to Figure 1a to understand the differences between exponential discounting and QH discounting.

We now present our main contribution, Algorithm 1, which finds a policy satisfying (1), that is, an MPE. In our algorithm, θ_n is the parameter corresponding to the policy that estimates an MPE at time n, while W_n is the estimate of the QH Q-value function of that policy. The (θ_n)- updates improve the policy, thus, it is referred to as the actor update. On the other hand, the (W_n)- updates attempt to evaluate the Q-value function of π_{θ_n} under QH discounting; hence, it is referred to as the critic update. We need to learn both θ and Q-value function, as the MPE policy cannot be derived always from its Q-values as in exponential discounting. In this work we consider the case where the stepsizes α_n and β_n for updating W_n and θ_n respectively satisfy $\alpha_n/\beta_n \to 0$. Because the θ_n-iterates get updated on a faster timescale, our algorithm falls under the category of Critic-Actor algorithms (Bhatnagar et al., 2023).

We now motivate the design of our algorithm from the perspective of traditional Critic-Actor algorithm. Let $\eta^{\sigma,\gamma}_{\pi_\theta}(\mu) := \mathbb{E}_{s\sim\mu, a\sim\pi_\theta}[Q^{\sigma,\gamma}_{\pi_\theta}(s,a)]$ denote the policy value of π_θ for a fixed initial state distribution μ. Similar to exponential discounting (Sutton et al., 1999), the policy gradient for QH discounting is

$$\begin{aligned} \frac{\partial \eta^{\sigma,\gamma}_{\pi_\theta}(\mu)}{\partial \theta(s,a)} =& (1-\sigma)\mu(s)\pi_\theta(a|s)A^0_{\pi_\theta}(s,a) \\ &+ \frac{\sigma}{1-\gamma} d^{\pi_\theta}_{\mu}(s)\pi_\theta(a|s)A^{\gamma}_{\pi_\theta}(s,a). \end{aligned} \tag{2}$$

When $\sigma = 1$, the RHS above reduces only to the second term, which depends on the advantage function ($A^{\gamma}_{\pi_\theta}(s,a) = Q^{\gamma}_{\pi_\theta}(s,a) - \langle \pi_\theta(\cdot|s),\ Q^{\gamma}_{\pi_\theta}(s,\cdot)\rangle$) with respect to exponential discounting. However, in the policy gradient of QH discounting, the state distribution in the first term (i.e., $\mu(s)$) and the second term (i.e., $d^{\pi_\theta}_{\mu}(s)$) are not equal so they cannot be combined to get the QH advantage function $A^{\sigma,\gamma}_{\pi_\theta}$. With this observation in mind, we designed our update rule of θ_n, i.e., the actor's behavior, to directly depend on the advantage function $A^{\sigma,\gamma}_{\pi_\theta}$ instead of (2).

Algorithm 1 Synchronous MPE-learning

```
Input: $\{\alpha_n\}_{n\geq 0}, \{\beta_n\}_{n\geq 0}$ satisfying $\mathcal{A}_1$, discount factors $\sigma, \gamma$
Initialize: $\theta_0, W_0 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$
for $n = 0, 1, 2, ...$ do
    Initialize $r'_n, W'_n, \hat{W}^{\theta_n}_n \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$
    for $(s,a) \in \mathcal{S} \times \mathcal{A}$ do
        Observe $s' \sim \mathcal{P}(\cdot|s,a)$
        Sample $a' \sim \pi_{\theta_n}(\cdot|s')$
        $r'_n(s,a) \leftarrow r(s',a'), W'_n(s,a) \leftarrow W_n(s',a')$
        $\hat{W}^{\theta_n}_n(s,a) \leftarrow \langle \pi_{\theta_n}(\cdot|s),\ W_n(s,\cdot)\rangle$
    end for
    $W_{n+1} = W_n + \alpha_n[r - (1-\sigma)\gamma r'_n + \gamma W'_n - W_n]$
    $\theta_{n+1} = \theta_n + \beta_n \left[W_n - \hat{W}^{\theta_n}_n\right]$
end for
```

2.2. Main Result

We first state all our our assumptions.

$\mathcal{A}_1$. **Step sizes:** $(\alpha_n)_{n\geq 0}$ and $(\beta_n)_{n\geq 0}$ are two sequences of positive real numbers satisfying (i) $\alpha_0 \leq 1, \beta_0 \leq 1$ and $(\alpha_n), (\beta_n)$ are monotonically decreasing, (ii) $\lim_{n\to\infty}(\alpha_n/\beta_n) = 0$, and (iii) $\sum_{n=0}^{\infty} \alpha_n = \sum_{n=0}^{\infty} \beta_n = \infty$; further, $\sum_{n=0}^{\infty}(\alpha_n^2 + \beta_n^2) < \infty$.

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

For $W \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, define the set-valued maps

$$\begin{aligned} \lambda(W) &:= \overline{co}\{g : \mathcal{S} \to \mathcal{A} : g(s) \in \arg\max W(s,\cdot)\}, \\ T_\lambda(W) &:= \{r + \gamma\mathcal{P}g\left[-(1-\sigma)r + W\right] : g \in \lambda(W)\}, \end{aligned}$$

where $\overline{co}$ is the convex closure.

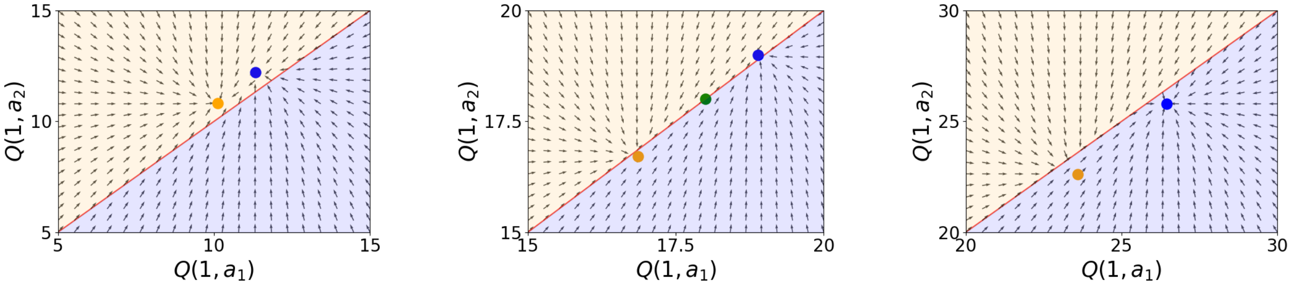
Our main result can now be stated as follows. Let $\|\cdot\|$ be the Euclidean norm.

Theorem 2.1 (Main Result)**.** *Suppose $\mathcal{A}_1$ and $\mathcal{A}_2$ hold. Then, we have the following statements:*

(i) The (W_n) iterates of Algorithm 1 are stable, i.e., $\sup_n ||W_n|| < \infty$ a.s.;

(ii) $D(\pi_{\theta_n}, \lambda(W_n)) \to 0$, where $D(x,Y) := \inf_{y\in Y} ||x - y||$ and θ_n, W_n are as in Algorithm 1; and

[1] Henceforth, we denote a stationary policy $\bar{\pi} \equiv (\pi, \pi, \ldots)$ by π itself.



(a) $\sigma = 0.3$

(b) $\sigma = 0.5$

(c) $\sigma = 0.7$

Figure 2: Vector fields for the DI (3). Here orange dot is the $Q_{\pi_{a_2}}$ value, blue dot is the $Q_{\pi_{a_1}}$ value and green dot is the $Q_{\pi_{MPE}}$ value.

(iii) The (W_n) sequence of Algorithm 1 converges to a connected internally chain transitive set of the Differential Inclusion (DI)

$$\dot{W}(t) \in T_\lambda(W(t)) - W(t). \tag{3}$$

Remark 2.2. Due to $\mathcal{A}_1$(ii), (θ_n) is updated on a faster timescale than (W_n). Hence, (W_n) appears quasi-static to (θ_n). If W_n is indeed static, i.e., $W_n \equiv W$, then it is easy to check that $\pi_{\theta_n} \to \lambda(W)$. Statement (ii) extends this claim to show that the distance between θ_n and $\lambda(W_n)$ asymptotically decreases.

Remark 2.3. $\mathcal{A}_1$ is a standard assumption in the SA literature for two time scale algorithms. $\mathcal{A}_2$ is also not very restrictive and is used to show a.s. boundedness or the stability of the (W_n)-iterates in Statement (i). This stability is crucial to show Statement (iii).

Remark 2.4. While we show that the (W_n)-sequence is bounded (see Theorem 2.1(i)), (θ_n) will be unbounded, i.e., $\|\theta_n\| \to \infty$. Existing works on critic-actor or actor-critic methods use projection to a bounded set to forcefully ensure (θ_n)'s stability to enable analysis using existing techniques. In contrast, we use the stability of the policies π_{θ_n} to show Theorem 2.1(iii), which is novel.

Proposition 2.5. *The $Q^{\sigma,\gamma}_{\pi^*}$ values of every MPE (π^*) of an MDP is a zero of the DI (3).*

2.3. Control Theory connection to Main Result

To explain our algorithm's behavior, we examine the vector fields of (3) for various σ values, as depicted in Figure 2 for the MDP in Figure 1b. The blue region below the line $x = y$ shows where a_1 is the greedy action, while the orange region shows where a_2 is the greedy action. The dynamics in each region move towards the Q-value function of the respective greedy policy, resulting in different dynamics and a discontinuous nature, which makes an ODE approach unsuitable. Instead, we use DI, a standard tool to analyze discontinuous dynamics in control theory. For a DI, the trajectories may converge to a point within a cone or on the boundary between cones. For instance, Figures 2a and 2c show that for $\sigma = 0.3$ and $\sigma = 0.7$, the optimal policies g and f lie inside their respective cones. However, this is not always true, as shown in Figure 2b for $\sigma = 0.5$, where the dynamics push trajectories toward specific points within each region, ultimately converging on the boundary at the MPE, represented by the green dot.

While the dynamics in tabular exponential discounting are discontinuous, the presence of a contraction operator makes DI analysis unnecessary. However, in QH discounting, the absence of a known contraction operator requires the adoption of the new DI perspective for analysis. Additionally, the optimal policy is always deterministic in exponential discounting, meaning the optimal point lies within a cone. In contrast, as shown in Figure 2b, the optimal policy in QH discounting may lie on the boundary. Therefore, we need tools from control theory, such as sliding mode attractors and Lyapunov functions for DI, to establish results in QH discounting, which we are currently pursuing.

3. Brief of Experimental Results

In this paper, we use the famous inventory control problem to empirically test our algorithm, with a detailed discussion available in Appendix A. We ran our algorithm to find the MPEs for an inventory control problem with a maximum storage capacity of 2 converged to three different points, indicating that MPEs are not unique, even in a small case of 3 states and 3 actions. Unlike in exponential discounting, where multiple actions can have the highest $Q^{\sigma,\gamma}$ values, these actions in MPE must be chosen according to an MPE policy to achieve the expected profits. Any deviation from an MPE policy will result in sub-optimal profits. However, within an MPE, no action has a higher $Q^{\sigma,\gamma}$ value than the one suggested, so there is no incentive for the agent to deviate from the MPE (see Tables in Appendix A).

References

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

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.

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 rl with $epsilon$-greedy exploration: A differential inclusion view. 2022.

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. In Solla, S., Leen, T., and Müller, K. (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

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

Zhao, X., Zhou, Y., and Xie, J. An inventory system with quasi-hyperbolic discounting rate. *IISE Transactions*, 49 (6):593–602, 2017.

Zipkin, P. Foundations of inventory management mcgraw-hill. *Irwin, New York, USA*, 2000.

A. Experiments

This section has two main objectives. First, to empirically demonstrate that our proposed Algorithm 1 finds an MPE, a policy favored by an agent aware of time-inconsistency (sophisticate agent). Second, to show that the vanilla policy gradient-based algorithm produces a time-inconsistent optimal policy, favored by an agent unaware of this time-inconsistency (naive agent). We will briefly describe the inventory control problem, the chosen environment to illustrate these claims.

A.1. Inventory Control: Problem Setup

In this study, we focus on the inventory control problem, a widely researched topic in economics (Zipkin, 2000). It's a classic problem in Reinforcement Learning, where the manager's preferences are typically represented using exponential discounting. Building on previous work such as (Zhao et al., 2017), we model the manager's preferences using quasi-hyperbolic discounting. We will now introduce the problem by framing it as an infinite horizon discrete-time Markov Decision Process.

State space: The state space $\mathcal{S}$ is defined as $\{1, 2, ..., M\}$, where M represents the maximum storage capacity of the inventory. Each state $s_t \in \mathcal{S}$ corresponds to the number of items in the inventory on day t.

Action space: The action space $\mathcal{A}$ is defined as $\{1, 2, ..., M\}$. Each action $a_t \in \mathcal{A}$ represents the number of items the manager orders on day t, with the assumption that the order will be delivered on the same day.

System dynamics: Each day, the manager observes customer demand d_t, assumed to be independently sampled from an unknown demand distribution D, driving system dynamics. For a given state s, action a, and demand d, the next state s' is determined by $s' = \max(\min(s + a, M) - d, 0)$. Here, $\min(s + a, M)$ ensures the ordered items do not exceed the maximum capacity M, with excess items being discarded if the inventory is full. The expression $\max(\min(s + a, M) - d, 0)$ accounts for partial fulfillment of demand based on available stock. Considering the stochastic demand, the transition dynamics are described by $\mathcal{P}(s'|s, a) = \mathbb{E}_{d\sim D}[I\{s' = \max(\min(s + a, M) - d, 0)\}]$.

Reward function: The daily reward, on day t, encompasses three components: 1) The cost of purchasing a_t items at a unit cost of c. 2) Revenue from sales, where each item fetches a price of p. 3) Holding costs for remaining inventory items, with a per-item holding cost of h.

In this work, we consider an inventory system with a maximum capacity of $M = 2$. The procurement cost per item is $c = 500$, while the selling price is $p = 900$, and the holding cost is $h = 50$. For this analysis, we employ a short-term discount factor of $\sigma = 0.3$ and a long-term discount factor of $\gamma = 0.9$.

A.2. Optimal Policy of the Sophisticate Manager

We now discuss the optimal policy of a sophisticate manager, i.e an MPE. From Theorem 2.1, it is clear that our algorithm tracks a DI whose invariant sets contain MPEs. Hence, we run the Algorithm 1 on the inventory control problem under consideration. We observe that the algorithm converges to three different points from different runs. The three points where our algorithm converged are indeed MPEs. We denote policy π by a matrix A, where $\pi(a|s) = A[s, a]$. Table 1, 2 and 3 represents the $Q^{\sigma,\gamma}$ values for the 3 MPEs π_1^*, π_2^* and π_3^* respectively.

Here are some key insights from this experiment:

1. MPEs under QH discounting may not be unique.
2. Unlike in exponential discounting, where multiple actions in MPE may have the highest $Q^{\sigma,\gamma}$ values, in QH discounting, actions must adhere to the MPE policy to achieve the expected profits outlined in the corresponding tables. Any deviation from this policy results in suboptimal profits. However, within an MPE, no action yields a greater $Q^{\sigma,\gamma}$ value than the one recommended by it, eliminating any incentive for the manager to deviate.
3. As shown in Tables 1, 2, and 3, the profits from each MPE differ. When knowledge of multiple MPEs is available, selecting a specific one may be advantageous. For instance, in our setup, π_3^* is preferable as it yields the maximum $Q^{\sigma,\gamma}(s, a)$ for all s and a, compared to π_1^* and π_2^*.

$$\pi_1^* = \begin{bmatrix} 0.00 & 0.53 & 0.47 \\ 0.53 & 0.47 & 0.00 \\ 1.00 & 0.00 & 0.00 \end{bmatrix}, \pi_2^* = \begin{bmatrix} 0.0 & 0.8 & 0.2 \\ 0.0 & 1.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \end{bmatrix},$$

$$\pi_3^* = \begin{bmatrix} 0.0 & 0.3 & 0.7 \\ 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \end{bmatrix}$$

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

Table 1: $Q^{\sigma,\gamma}$ values for policy π_1^*

a s	0	1	2
0	873.75	**1040.5**	**1053**
1	1540.5	**1540.5**	1040.5
2	**2040.5**	1540.5	1040.5

Table 2: $Q^{\sigma,\gamma}$ values for policy π_2^*

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

Table 3: $Q^{\sigma,\gamma}$ values for policy π_3^*

A.3. Optimal Policy of the Naive Manager

We now discuss finding the optimal policy for the naive manager. Note that the vanilla policy gradient with QH discounting as in (2) optimizes the policy value from the current time and ignores the change in the discounting factor of the agent in future times. This is precisely the behavior of a naive manager, where they maximize the value function without considering the time-inconsistent nature of the optimal policy. Hence, we claim that the solution obtained by performing gradient ascent using vanilla policy gradient is optimal for a naive agent and not for a sophisticated one. For the setup under consideration, the optimal policy of the naive manager is:

$$\pi_N^* = \begin{bmatrix} 0.0 & 0.0 & 1.0 \\ 0.0 & 1.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \end{bmatrix}$$

a s	0	1	2
0	1080	$\underline{1235.5}$	**1228**
1	$\underline{1735.5}$	**1728**	1228
2	**2228**	1728	1228

Table 4: $Q^{\sigma,\gamma}$ values for policy π_N^*

We now show that the optimal policy π_N^* is time-inconsistent. Table 4 represents the $Q^{\sigma,\gamma}$ values for the optimal policy of the naive manager π_N^*. The numbers marked in bold represent the profit the manager gets by following π_N^*. However, in state 0, the $Q^{\sigma,\gamma}_{\pi_N^*}(0,1) > Q^{\sigma,\gamma}_{\pi_N^*}(0,2)$ (the one suggested by the policy π_N^*). Hence, when the inventory is at state 0, the naive manager decides to deviate from the optimal policy for the current day and follow the optimal policy from the next day onwards. Similarly, in state 1, the manager deviates from action 1 to action 0. The naive manager, the next day, unaware that the previous day's decision was to continue with the optimal policy from today, again re-evaluates the optimal policy and deviates from the optimal policy for one more day. This process continues, and the naive manager finally follows the below-mentioned policy:

$$\pi_N = \begin{bmatrix} 0.0 & 1.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \end{bmatrix}$$

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

Table 5: $Q^{\sigma,\gamma}$ values for policy π_N

Table 5 represents the $Q^{\sigma,\gamma}$ values of the policy π_N which is finally followed by the naive manager. The values mentioned in boldface are the profits gained by the manager by following π_N. A key point to note here is that the profits realized by the naive agent is less than both the underlined and boldfaced profits mentioned in Table 4. This indicates that the naive agent earns less than the profits suggested by the time-inconsistent policy, even though it initially seemed that deviating from it would lead to higher profits.