What is your discount factor?

Shadi Tasdighi Kalat, Sriram Sankaranarayanan, and Ashutosh Trivedi

University of Colorado Boulder, Boulder CO, USA

Abstract. We study the problem of inferring the discount factor of an agent optimizing a discounted reward objective in a finite state Markov Decision Process (MDP). Discounted reward objectives are common in sequential optimization, reinforcement learning, and algorithmic game theory. The discount factor is an important parameter used in formulating the discounted reward. It captures the "time value" of the reward - i.e., how much reward at hand would equal a promised reward at a future time. Knowing an agent's discount factor can provide valuable insights into their decision-making, and help predict their preferences in previously unseen environments. However, pinpointing the exact value of the discount factor used by the agent is a challenging problem. Ad-hoc guesses are often incorrect.

This paper focuses on the problem of computing the range of possible discount factors for a rational agent given their policy. A naive solution to this problem can be quite expensive. A classic result by Smallwood shows that the interval $[0, 1)$ of possible discount factor can be partitioned into finitely many sub-intervals, such that the optimal policy remains the same for each such sub-interval. Furthermore, optimal policies for neighboring sub-intervals differ for a single state. We show how Smallwood's result can be exploited to search for discount factor intervals for which a given policy is optimal by reducing it to polynomial root isolation. We extend the result to situations where the policy is suboptimal, but with a value function that is close to optimal. We develop numerical approaches to solve the discount factor elicitation problem and demonstrate the effectiveness of our algorithms through some case studies.

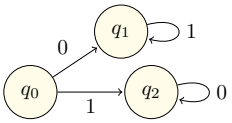
Keywords: Discount factor · Markov Decision Process · Optimization

1 Introduction

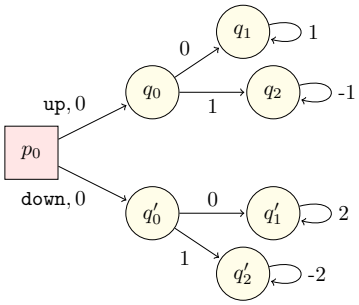
The recent success of Reinforcement Learning (RL) algorithms [24] in discovering creative solutions with superhuman performance [18, 26] has contributed to the popularity of the discounted-sum as a canonical optimization objective. Discounted-sum aggregators over Markov decision processes (MDP) [20] offer several theoretical advantages, including the existence of stationary optimal policies, a contractive improvement operator, and an effective formalism to approximate other aggregators such as limiting-average and total-sum. The discounted-sum aggregator formalizes the notion of time preference (or rate of impatience) [8] that rational agents exhibit when weighing immediate small rewards against larger rewards later.

However, characterizing this preference is challenging because it is subjective and depends on the agent's ability to delay gratification. This paper aims to develop algorithms that identify both exact and approximate ranges of discount factors to which an agent might subscribe by observing their behavior in a finite-state MDP.

The need for discount factor elicitation. The need for discount factor elicitation arises in several situations. In single-agent optimization scenarios, it is helpful for an agent to know its discount factor in order to select among different behaviors. For instance, in the decision-making scenario shown on the right, the decision maker has to choose between sequences $\sigma_1 : 01^\omega$ and $\sigma_2 : 10^\omega$ with resulting discounted sums of $\gamma/(1-\gamma)$ and 1, respectively. Thus, if the agent's discount factor lies in the range $[0, 1/2)$, σ_2 is optimal; for discount factor of $1/2$ both sequences have the same discounted reward, and for discount factors in the interval $(1/2, 1)$, the sequence σ_1 is optimal. Moreover, in situations where one can assume a subjective and constant discount factor, elicitation of the discount factor can be the canonical "inverse-RL" problem and may characterize a key hyperparameter for transfer learning. In other words, we focus on uncovering characteristics of the agent which are affecting their decision making dynamics in order to be able to predict their behavior in unfamiliar situations.



In multi-agent interaction scenarios, the assumption that adversarial agents subscribe to the same discount factor as the ego agent may not hold. Consider a strategic interaction as a two-player turn-based discounted game depicted in the figure wrapped to the left of this text, with circular nodes controlled by Player Min and a box node controlled by Player Max. Let's consider two extreme situations with respect to Player Min's choice of discount factors: myopic (discount factor close to 0) and foresight (discount factor close to 1). Also, assume that Player Max is foresighted, i.e., subscribes to a discount factor close to 1. Intuitively, a myopic player will select a locally-optimal action, while a foresighted player will play for the average reward. Hence, in this case, against a myopic Player Min, Player Max will select action `down`, while against a foresighted Player Min, Player Max will select action `up`. This example demonstrates that knowing the discount factor of the other agent may allow the ego agent to achieve a better payoff.



The landscape of discount factors. For finite-state MDPs, Blackwell [3] introduced the notion of discount-optimal policies that are optimal for all sufficiently large discount factors. A policy is called Blackwell-optimal if there exists a

discount factor γ_* such that for all $\gamma \in [\gamma_*, 1)$, it is also discount-optimal. Gurvich and Miltersen [11] bound γ_* from above by $1-1/((n!)^2 2^{2n+3} M^{2n^2})$ where n is the number of states in the MDP and M is a bound for the absolute values of probabilities and rewards. Puterman [20, Section 10.1.2] presents an accessible proof for the existence of Blackwell-optimal policies based on a rational characterization of discount-factor parameterized value function and finiteness of the stationary policies. The proof can easily be extended to show that, for every finite MDP, there exists a finite sequence $\mathcal{L} = \langle a_0(=0) \leq a_1 \leq a_2 \leq \cdots \leq a_N = \gamma_* \leq 1 \rangle$ such that within each region $[a_i, a_{i+1}]$, for $0 \leq i < N$, the discount optimal policies remain invariant.

From Policies to Discount-Factors. Smallwood [23] refined the understanding of this landscape $\mathcal{L}$ by observing a remarkable fact: for every discount factor a_i in this landscape, the optimal policies on either side differ in only a single state. Using this observation, Smallwood developed a procedure to determine the value of discount factors for which one is indifferent between two optimal policies. Given an optimal policy, this observation can be leveraged to compute a range of discount factors for which the given policy is optimal. However, as discussed in the paper, a straightforward approach of this kind is computationally expensive. To address this challenge, we propose algorithms to compute such ranges for optimal, and sub-optimal policies. We also present experimental results to demonstrate the effectiveness of the proposed approach.

Organization. We begin the technical exposition with prior work in Section 2. We formalize the problem of discount factor elicitation in Section 3. Our approach builds upon Smallwood's work that is presented in Section 4. The key algorithm is developed in Section 5 and its extension to accommodate sub-optimal policies in Section 6. In Section 7 we discuss a case study for elicitation of discount factors for varying environment before concluding in Section 8.

2 Related work

Discounted Optimization. Puterman [20] and Filar & Vrieze [7] provide comprehensive collections of technical results on optimization over MDPs and stochastic games. Discounted, total, and average reward are among the most studied optimization objectives. Among these, the discounted-sum objective is the best understood, theoretically elegant, and allows effective optimization [20] and reinforcement learning [24] algorithms. The notion of Blackwell-optimality [3] allows us to reduce average optimization to discounted-sum optimization.

Discount Factor. Starting from the work of Fisher [8], the discount factor has received several interpretations, including rate of impatience [8], delayed gratification [9, 17], geometrically distributed horizon [16], system failure [20], timing of consumption & expenditure of resources [13], and stochastic shortest path with proper-policy assumption [19]. Given their importance in optimization and games, it is surprising that the problem of selecting the discount factor has not received sufficient attention, and there is no consensus among practitioners

on how to select a discount factor [2, 9, 12, 25]. Smallwood's work [23] is perhaps closest to our problem; however, his goal is to find the discount factors under which two optimal policies are equally profitable. Giwa et al [10] present an approach to inverse reinforcement learning that simultaneously optimizes the rewards and the discount factor of the agent from observations of the agent's behavior. The key differences between our work and Giwa et al include: (a) we assume that the entire policy is given for a given MDP whose rewards are given and learn a discount factor, whereas Giwa et al do not assume that the rewards are known; and (b) we identify all discount factors for which the input policy is optimal whereas Giwa et al minimize a log-likelihood using gradient-based optimization that yields a local minimum.

Mismatching Discount Factors. In the context of strategic games, the need for strategic agents to have different planning horizons and consequently different discount factors is well documented [1, 4, 14]. The issue of different discounting was studied extensively by Lehrer et al. [14], and they showed that in the case of complete information zero-sum game, the equilibrium payoff is zero-sum. They also demonstrated that despite the purely competitive nature of the games, different discounting can give rise to some cooperation, known as inter-temporal trades [1].

3 Problem Definition

We focus on inferring the unknown discount factor for a single agent based on their policy in a finite MDP. We assume that the agent's policy selects actions to maximize the discounted sum of rewards over an infinite time horizon.

Definition 1. *An MDP $\mathcal{M}$ is a tuple $(S, A, (P_a)_{a\in A}, (R_a)_{a\in A})$, where:*

- *$S = \{1, \ldots, N\}$ is a finite set of states,*
- *A is a finite set of actions,*
- *for each action $a \in A$, P_a is the transition probability matrix of size $|S| \times |S|$, wherein $P_a(i, j)$ represents the probability of moving from state i to state j upon action a, and*
- *for each action $a \in A$, the reward $R_a(i, j)$ is the reward obtained by applying action a at state i and reaching state j.*

Assumption 1 *We will assume that for any state $i \in S$ and any two actions $a_1, a_2 \in A$ such that $a_1 \neq a_2$, the next state distributions $P_{a_1}(i, j) \neq P_{a_2}(i, j)$ for some $j \in S$. Although this assumption is not strictly necessary for our approach but will be useful in avoiding checking for special cases.*

The agent is interested in maximizing a discounted-sum of reward signals. It is well known [3, 20] that for discounted-sum objective, deterministic and memoryless policies suffice for optimality. Thus, we restrict our focus to such policies.

A deterministic, memoryless *policy* $\pi : S \to A$ for $\mathcal{M}$ is a mapping from the set of states S to the set of actions A. We write Π for the set of all possible policies

of the observed agent. We represent every policy as a $|S|$-vector that maps each state to an action. Given a policy π, we define the transition matrix P_π as $P_\pi(i,j) = P_{\pi(i)}(i,j)$. In other words, $P_\pi(i,j)$ is the probability of moving from i to j upon the action $\pi(i)$. Similarly, let $R_\pi(i,j)$ denote the reward $R_{\pi(i)}(i,j)$.

The *value* associated with policy $\pi \in \Pi$, denoted $\mathbf{v}_\pi$, maps every $s \in S$ to the discounted sum of rewards gained by following policy π starting from s. The value function is characterized by the following equations:

$$\mathbf{v}_\pi(i) = \sum_{j=1}^{N} P_\pi(i,j)\left(R_\pi(i,j) + \gamma \mathbf{v}_\pi(j)\right) \tag{1}$$

wherein $\gamma \in [0,1)$ is the *discount factor*. The discount factor controls the effect of immediate versus future rewards on the decisions made by the agent.

*Problem 1 (***Discount Factor Elicitation).** Assume that the agent plays according to a given policy $\pi \in \Pi$, which is an optimal policy corresponding to some unknown discount factor γ. We seek to find a (or union of) closed interval I of the form $[\gamma_l, \gamma_u]$ for $0 \leq \gamma_l \leq \gamma_u < 1$ or a half-open interval of the form $I: [\gamma_l, 1)$ such that for all $\gamma \in I$, the given policy π is optimal for the value of γ provided. I.e., for each $\gamma \in I$, and for any policy $\xi \in \Pi$ and for all states i:

$$\mathbf{v}_\pi(i) = \sum_{j=1}^{N} P_\pi(i,j)\left(R_\pi(i,j) + \gamma \mathbf{v}_\pi(j)\right) \geq \mathbf{v}_\xi(i)\,. \tag{2}$$

This problem can be reduced to a decision problem involving univariate rational functions through a standard linear-programming (LP) based approach. The value function for a given discount factor γ is obtained as a vector $\mathbf{v}$, wherein $\mathbf{v}(i)$ is the value associated with state i, via the following LP [20]:

$$\begin{aligned} \min \quad & \textstyle\sum_{i=1}^{n} \mathbf{v}(i) \\ \text{s.t. } & \mathbf{v}(i) - \gamma \textstyle\sum_{j=1}^{n} P_a(i,j)\mathbf{v}(j) \geq \mathbf{q}_a(i) \quad \forall a \in A, i \in \{1,\ldots,N\} \end{aligned}$$

where $\mathbf{q}_a(i) = \sum_{j=1}^{n} R_a(i,j)P_a(i,j)$ for $a \in A$ is the vector of expected immediate rewards. Given the value function, the policy π is extracted from the optimal solution of the LP by noting that whenever $\pi(i) = a$, the constraint corresponding to state i and action $\pi(i)$ in the LP above is saturated, i.e., satisfied with an equality. Note that if there are multiple optimal policies π, π' for a given discount factor, then the constraints corresponding to the actions chosen by each of the policies are saturated by the optimal solution to the LP above.

Lemma 1. *A policy π is optimal for a given discount factor $\gamma \in [0,1)$ if and only if the following constraints hold for each action $a \in A$:*

$$(I - \gamma P_a)(I - \gamma P_\pi)^{-1}\mathbf{q}_\pi \geq \mathbf{q}_a\,,$$

wherein $\mathbf{q}_\pi(i) = \mathbf{q}_{\pi(i)}(i)$.

Proof. Let π be the optimal policy and $\mathbf{v}^*$ be the corresponding value function. It holds that $\mathbf{v}^*$ must be a feasible solution to the LP above that satisfies:

$$\mathbf{v}^*(i) - \gamma \sum_{j=1}^{n} P_\pi(i,j)\mathbf{v}^*(j) = \mathbf{q}_\pi(i) \;\; \forall i \in \{1,\ldots,N\}\,.$$

In other words, $\mathbf{v}^* - \gamma P_\pi \mathbf{v}^* = \mathbf{q}_\pi$. Since $I - \gamma P_\pi$ is an invertible matrix for $\gamma \in [0,1)$ and P_π being a stochastic matrix, we have $\mathbf{v}^* = (I - \gamma P_\pi)^{-1}\mathbf{q}_\pi$. Also, $\mathbf{v}^*$ must be primal feasible as well, yielding $(I - \gamma P_a)\mathbf{v}^* \geq \mathbf{q}_a$ for every $a \in A$. Therefore, we may eliminate v^* to obtain

$$(I - \gamma P_a)(I - \gamma P_\pi)^{-1}\mathbf{q}_\pi \geq \mathbf{q}_a\,, \forall a \in A.$$

This completes the proof. □

As a result, a policy π is optimal for some discount factor $\gamma \in [0,1)$ iff the following assertion over the reals is valid.

$$\exists \gamma \in [0,1) \bigwedge_{a \in A} (I - \gamma P_a)(I - \gamma P_\pi)^{-1}\mathbf{q}_\pi \geq \mathbf{q}_a\,. \tag{3}$$

Note that each $(I - \gamma P_\pi)^{-1}$ is a $N \times N$ matrix whose entries are rational functions (ratio of two polynomials in γ), wherein the degrees of the numerator and denominator are at most N. The denominators are all given by $\det(I - \gamma P_\pi)$, which is positive over $\gamma \in [0,1)$.

Lemma 2. *The polynomial* $\det(I - \gamma P_\pi) > 0$ *for* $\gamma \in [0,1)$.

Proof. Since P_π is a stochastic matrix, its eigenvalues λ are such that $|\lambda| \leq 1$. The real roots of $\det(I - \gamma P_\pi)$ are in fact $\frac{1}{\lambda}$ where λ is a non-zero real eigenvalue of P_π. Since $|\lambda| \leq 1$, we conclude that $\det(I - \gamma P_\pi)$ has no real roots in $[0,1)$ and is thus sign invariant. Thus, $\det(I - \gamma P_\pi) > 0$ since it is positive for $\gamma = 0$. □

The problem of checking if a given policy π is optimal for some discount factor γ can be reduced to checking whether for some $\gamma \in [0,1)$, a given list of $N|A|$ polynomials involving γ of degree at most $N+1$ are all non-negative. The roots of a given polynomial delineate intervals where the sign of the polynomial does not change. Therefore, to find intervals where the polynomials are all non-negative, we first identify their roots. In practice, this can be solved using real-root isolation for the polynomials using Sturm sequences or the Descartes rule of signs to count the number of real roots in an interval and bisection to refine these intervals [21, 27]. Having identified all such sign invariant intervals, the procedure checks if the intervals corresponding to the polynomials have a non-empty intersection. In what follows, we will provide a more elegant approach using a result from Smallwood [23].

4 Optimal Policy Regions

Our approach to Problem 1 builds upon the result of Smallwood [23], which shows that as we vary the discount factor γ over the interval $[0, 1)$, we effectively partition the interval $[0, 1)$ into finitely many intervals $I_1, \ldots, I_K$ such that for each interval I_l, there is an associated policy π_l which is optimal for any discount factor $\gamma \in I_l$. Furthermore, two neighboring intervals I_l, I_{l+1} that share an endpoint have associated policies π_l, π_{l+1} that differ only at a single state.

Theorem 1 (Smallwood [23]). *For a given MDP $\mathcal{M}$, there exists finitely many points $0 = a_0 \leq a_1 \leq a_2 \cdots \leq a_N < 1$ and policies $\pi_0, \pi_1, \ldots, \pi_N$ such that*

(a) for $i < N$, π_i is an optimal policy for any discount factor $\gamma \in [a_i, a_{i+1}]$,
(b) π_N is an optimal policy for $[a_N, 1)$,
(c) "neighboring" policies π_i, π_{i+1} differ only at a single state, and
(d) the optimal values associated with π_i, π_{i+1} are the same for all states at discount factor a_{i+1}.

Following Smallwood, we call an interval of the form $[a_i, a_{i+1}]$ (or $[a_N, 1)$) an *optimal policy region*. For any two neighboring optimal policies π_{i-1}, π_i for $i \geq 1$, the point a_i that is common to their respective optimal policy regions is called an *indifference point*. Note that a policy π can be optimal for two separate intervals. Therefore, there could be multiple indifference points between two policies. At an indifference point, the optimal action for one state of the MDP changes, resulting in a change from a policy π_i to π_{i+1}. Note that at the indifference point, the value functions associated with π_i and π_{i+1} are the same.

Based on Smallwood's Theorem, we propose the following simple strategy for identifying the optimal policy intervals corresponding to a given policy π:

1. Iterate through all policies π' that differ from π at just a single state. There are $N(|A| - 1)$ such policies.
2. For each such policy π' find if an indifference point exists between π, π'.
3. If π, π' have an indifference point γ for which they are optimal, then we have discovered one of the end points of an interval for which π is optimal.

The key observation [23] is that we can check if two policies π, π' have an indifference point in common using polynomial root solving. Let π, π' be two policies that differ in just one state. The value function $\mathbf{v}$ for π satisfies the equation: $\mathbf{v} - \gamma P_\pi \mathbf{v} = \mathbf{q}_\pi$. Likewise, π' must have the same value function that also satisfies: $\mathbf{v} - \gamma P_{\pi'} \mathbf{v} = \mathbf{q}_{\pi'}$. Subtracting, we obtain:

$$\mathbf{q}_\pi - \mathbf{q}_{\pi'} + \gamma(P_\pi - P_{\pi'})\mathbf{v} = 0$$

Since, π and π' differ from each other in one state, $\mathbf{q}_\pi$, and $\mathbf{q}_{\pi'}$ differ in just one entry corresponding to the state where the policies diverge. Similarly, P_π, and $P_{\pi'}$ also differ just in a single row where the two policies diverge. Therefore, we obtain a single equation pertaining to just the row where the two policies differ:

$$\Delta q + \gamma \Delta P (I - \gamma P_\pi)^{-1} \mathbf{q}_\pi = 0\,. \tag{4}$$

Here Δq is a scalar value that represents the difference between immediate expected rewards obtained by the two policies, and $\Delta P = (P_\pi - P_{\pi'})$ is a row vector. Assumption 1 ensures that $\Delta P \neq 0$. Otherwise, note that π, π' cannot share an indifference point. From Cramer's rule, the function $\gamma \to (I - \gamma P)^{-1}$ is a rational function whose denominator is $\det(I - \gamma P)$ [15]. Therefore, solving for γ in Eq.(4) is equivalent to finding the roots of the resulting rational function in γ within the interval $[0, 1)$.

The numerators and common denominator for the entries of $(I - \gamma P_\pi)^{-1}$ can be computed efficiently by computing the characteristic polynomial of P_π and noting that P_π satisfies its own characterisitic polynomial using the Cayley-Hamilton theorem. The details are explained in Smallwood's paper [23] wherein the method of [6] is employed. Computing the polynomial in Eq. (4) will require $O(N^3)$ additions and multiplications involving entries of $\Delta q, \Delta P, P_\pi$ and $\mathbf{q}_\pi$.

Example 1. Consider an MDP with states $S = \{0, 1, 2, 3\}$ and actions $A = \{0, 1\}$ where transition probabilities and expected immediate rewards are the following:

$$\text{Action 0: } P_0 = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/2 & 0 & 1/2 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}, \; q_0 = \big[8,\, 20,\, 2,\, 3\big],$$

$$\text{Action 1: } P_1 = \begin{bmatrix} 1/16 & 3/8 & 3/16 & 3/8 \\ 1/16 & 7/16 & 1/16 & 7/16 \\ 1/8 & 3/8 & 1/8 & 3/8 \\ 1/8 & 3/8 & 1/8 & 3/8 \end{bmatrix}, \; q_1 = \big[3,\, 15,\, 4,\, 40\big]$$

As we change the value of γ for this MDP, we observe that the optimum policy changes at certain values of γ. This partitions the whole range of discount factor $[0, 1)$ to sub-intervals where the optimum policy is the same. For this example, the optimum policy is $[0, 0, 1, 1]$ for $\gamma \in [0, 0.38)$, is $[0, 1, 0, 1]$ for $\gamma \in [0.38, 0.91)$, and is $[1, 1, 1, 1]$ for $\gamma \in [0.91, 1)$. So, in order to find these intervals, we can focus on their boundary points where the value of two different policies (optimum policies for two neighboring intervals) are the same. From Equation 4, for two policies $\pi_1 :\ [0, 1, 1, 1]$, and $\pi_2 :\ [0, 0, 1, 1]$ we have:

$$d_{\pi_1} = [0, 1, 1, 1],\; P_{\pi_1} = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/16 & 7/16 & 1/16 & 7/16 \\ 1/8 & 3/8 & 1/8 & 3/8 \\ 1/8 & 3/8 & 1/8 & 3/8 \end{bmatrix},\; q_{\pi_1} = \big[8,\, 15,\, 4,\, 40\big],$$

and $d_{\pi_2} = [0, 0, 1, 1],\ \Delta P = \big[-7/16,\, 7/16,\, -7/16, 7/16\,\big],\ \Delta q = -5$. This equation has a root $\gamma_0 = 0.38$ in the interval $[0, 1)$, which is the indifference point between d and d'. Figure 1.(a) depicts these intervals and the plot of the optimum value function versus γ.

5 Computing Discount Factor Ranges

Algorithm 1 returns a union of intervals of the discount factor γ for which a given policy $\hat{\pi}$ is optimal for a given MDP $\mathcal{M} = (S, A, (P_a)_{a \in A}, (R_a)_{a \in A})$ with

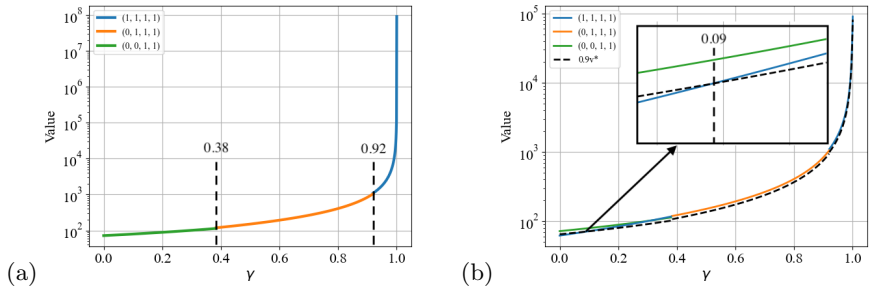


Fig. 1. (a) Optimum policy regions for the MDP in Example 1 (b) Range of discount factor for which policy $[1, 1, 1, 1]$ is near optimal for $\epsilon = 0.1$ is $[0.09, 1)$.

$|S| = N$ states and $|A| = M$ actions. Let $H_1(\hat{\pi})$ denote the set of all policies that differ from $\hat{\pi}$ at precisely one state. Line 2 iterates through all the policies π_n in $H_1(\hat{\pi})$. For each policy, it computes the intervals of discount factors where it is optimal by solving a polynomial equation as explained in Section 4. We will assume for now that we are able to find a precise root of the polynomial. However, we can adapt the algorithm for the more realistic case when we can isolate the root of the polynomial within an interval $\gamma \pm \delta$ for some tolerance $\delta > 0$. Section 5.1 delves into the details of this approach.

Next, we check if the policy $\hat{\pi}$ is optimal for γ_n using the isOptimal subroutine. If yes, we add γ_n to the list of indifference points Γ. Next, we sort the list of points in Γ in ascending order. We consider two consecutive elements in Γ (line 11) and check if $\hat{\pi}$ is optimal at the midpoint of the interval $[\gamma_i, \gamma_{i+1}]$. If yes, we add this interval to our list of intervals. Finally, we consider the interval $[\gamma_{\max}, 1)$, wherein γ_m is the largest number in Γ. This checks if $\hat{\pi}$ is a *Blackwell-optimal policy* [3].

The subroutine isOptimal for checking of the policy $\hat{\pi}$ is optimal for a given discount factor γ is as follows:

1. Compute the value function $\mathbf{v} = (I - \gamma P_{\hat{\pi}})^{-1}\mathbf{q}_{\hat{\pi}}$.
2. For each $a \in A$, check if $(I - \gamma P_a)\mathbf{v} \geq \mathbf{q}_a$.

The complexity of isOptimal procedure is $O(N^3 + N^2M)$ for N states and M actions. Algorithm 1 computes isOptimal at most $O(N^2M)$ times in the worst case and computeIndifferencePoint at most NM times. The overall complexity of Algorithm 1 is therefore $O(NM \times \mathsf{Poly}(N) + N^5M + N^3M^2)$, wherein $\mathsf{Poly}(d)$ is the complexity of finding the roots of a univariate polynomial of degree d.

Theorem 2. *For a given MDP $\mathcal{M}$ and policy $\hat{\pi}$, the result $\mathcal{I}$ obtained by running Algorithm 1 yields all optimal policy intervals for which $\hat{\pi}$ is the optimal policy.*

Proof. Let $[\ell, u]$ for $0 < \ell \leq u < 1$ be any maximal interval wherein $\hat{\pi}$ is optimal for all values of the discount factor in the interval. By the results from section 4, the end points ℓ, u are indifference points between $\hat{\pi}$ and a policy in $H_1(\hat{\pi})$.

Algorithm 1: Compute discount factor range(s) for which $\hat{\pi}$ is optimum

Data: MDP: $\mathcal{M}$ and policy: $\hat{\pi}$.
Result: Union of intervals $\mathcal{I} = \bigcup_j [\gamma_l^{(j)}, \gamma_u^{(u)}]$ for which $\hat{\pi}$ is optimal.

1 $\Gamma \leftarrow [0]$
2 **for** $\pi_n \in H_1(\hat{\pi})$ **do**
3 $\quad R_n \leftarrow \mathsf{computeIndifferencePoint}(\hat{\pi}, \pi_n)$ `/* R_n is a set of numbers in [0,1) */`
4 $\quad$ **for** $\gamma_n \in R_n$ **do**
5 $\quad\quad$ **if** $\mathsf{isOptimal}(\hat{\pi}, \gamma_n)$ **then append** γ_n to Γ
6 $\quad$ **end**
7 **end**
8 **sort** Γ in ascending order
9 $\mathcal{I} \leftarrow \emptyset$
10 **for** $i = 1, \ldots |\Gamma| - 1$ **do**
11 $\quad \gamma_i, \gamma_{i+1} \leftarrow \Gamma[i], \Gamma[i+1]$
12 $\quad \gamma_m = \frac{\gamma_i + \gamma_{i+1}}{2}$
13 $\quad$ **if** $\mathsf{isOptimal}(\hat{\pi}, \gamma_m)$ **then** $\mathcal{I} \leftarrow \mathcal{I} \cup \{[\gamma_i, \gamma_{i+1}]\}$
14 **end**
15 $\gamma_{\max} \leftarrow \Gamma[|\Gamma|]$ `/* Last element of Γ */`
16 $\gamma_b = \frac{\gamma_{\max}+1}{2}$ `/* Check if π̂ is a Blackwell optimal policy. */`
17 **if** $\mathsf{isOptimal}(\hat{\pi}, \gamma_b)$ **then** $\mathcal{I} \leftarrow \mathcal{I} \cup \{[\gamma_{\max}, 1)\}$
18 **return** $\mathcal{I}$

Therefore, ℓ, u will belong to the list of indifference points Γ. Suppose there are other indifference points $\ell < \ell_1 < \ell_2 < \ldots < \ell_k < u$ in the interval $[\ell, u]$ discovered by our algorithm. This is possible since there could be other policies in $H_1(\hat{\pi})$ whose value intersects with the value of $\hat{\pi}$ is just a single point. As a result, the for-loop in Line 10 will iterate through $\ell, \ell_1, \ldots, \ell_k, u$ and for each successive pair of points, it will discover that $\hat{\pi}$ is optimal for the mid-point. Therefore, the intervals $[\ell, \ell_1], \ldots, [\ell_k, u]$ will be part of $\mathcal{I}$. A similar argument holds if $\hat{\pi}$ is optimal over an interval of the form $[0, u]$ or $[\ell, 1)$. □

5.1 Approach Using Root Isolation

Thus far, Algorithm 1 assumes that we can compute the roots of a polynomial precisely. This requires working with algebraic numbers and furthermore, finding roots of a polynomial precisely is hard, when its degree exceeds 5. We modify Algorithm 1 to work with a polynomial root isolation procedure. Given a polynomial $p(\gamma)$ of degree N and tolerance factor δ, the root isolation yields approximate roots $\{\gamma_1, \ldots, \gamma_k\}$ such that the interval $[\gamma - \delta, \gamma + \delta]$ contains a root of $p(\gamma)$. We will assume that δ is small enough that any two roots of the polynomial are separated by at least δ. Such a procedure can be implemented using an algorithm similar to that of Collins and Akritas [5] that uses the Descartes rule of signs to check if an interval has no real-roots, one real root or multiple roots. The approach uses bisection by splitting the interval in half if it contains a

root until the size of the interval is smaller than a desired bound. This approach has been improved substantially in the recent past. The ANewDsc algorithm of Sagralof and Mehlhorn combines the bisection method based on the rule of signs with Newton's method for iterative computation of polynomial roots in time $\tilde{O}(N(N^2 + N\tau + \log(\delta)))$ where τ is the maximum number of bits needed to store the coefficients of the polynomial and $\tilde{O}$ denotes that terms that are logarithmic in the polynomial size and coefficients are omitted [22].

Suppose computeIndifferencePoints called from Algorithm 1 returns intervals of the form $(\gamma - \delta, \gamma + \delta)$ guaranteed to find a single indifference point, we will modify our approach as follows: Line 5 checks the optimality of both end points $\gamma - \delta, \gamma + \delta$ and inserts the ones for which $\hat{\pi}$ is optimal into the list Γ.

Let us assume Algorithm 1 was executed and the indifference points $\gamma \in \Gamma$ were found precisely. Let δ be chosen so that $2\delta < \min_{\gamma_1, \gamma_2 \in \Gamma, \gamma_1 \neq \gamma_2} |\gamma_2 - \gamma_1|$. In other words, no two indifference points discovered by the exact version of Algorithm 1 are closer than 2δ.

Theorem 3. *Let interval $[\ell, u]$ be a maximal interval wherein $\ell < u$ and $\hat{\pi}$ is optimal over the interval. The result $\mathcal{I}$ obtained by running Algorithm 1 modified with interval based root isolation for interval width δ will contain the interval $[\ell + \delta, u - \delta]$.*

Proof. Since $[\ell, u]$ is maximal, we note that the end points ℓ, u are indifference points. By our assumption on the minimum separation between indifference points, we have $u - l > 2\delta$. Therefore, $\ell + \delta, u - \delta$ are added to the set Γ in the modified version of Algorithm 1. Consider if there are indifference points $\ell < \ell_1 < \ell_2 < \cdots < \ell_k < u$. Due to the separation assumptions, note that we will find $\hat{\pi}$ to be optimal for $\ell_j \pm \delta$ for $j = 1, \ldots, k$. As a result, the intervals $[\ell + \delta, \ell_1 - \delta], [\ell_1 - \delta, \ell_1 + \delta], \cdots, [\ell_i + \delta, \ell_{i+1} - \delta], \cdots, [\ell_k + \delta, u - \delta]$ are all part of the result $\mathcal{I}$. Thus, $\mathcal{I}$ will contain the interval $[\ell + \delta, u - \delta]$. □

6 Near-Optimal Policy Regions

In previous sections, we formulated the problem of finding the range of discount factors for which a given policy is optimal. Here, we extend the problem to find ϵ-optimal regions for a given policy $\hat{\pi}$. Assume that $\hat{\pi}$ is a policy that is ϵ-far away from being optimal. We say that for a discount factor γ, $\hat{\pi}$ is ϵ optimal if the value function $\mathbf{v}_{\hat{\pi}}$ and the optimal value function $\mathbf{v}^*$ satisfy the inequality for some given $\epsilon \in [0, 1)$: $\mathbf{v}_{\hat{\pi}} \geq (1 - \epsilon)\mathbf{v}^*$.

Problem 2. Assume the agent plays according to known policy $\hat{\pi}$, find all discount factors such that $\hat{\pi}$ is ϵ optimal.

Clearly, $\hat{\pi}$ is ϵ optimal for all regions where it is optimal. However, (a) it need not be optimal anywhere and (b) it can be ϵ-optimal for discount factors that are "far away" from regions where $\hat{\pi}$ is optimal. From results in Section 4, we can compute all regions of discount factors as well as their respective

optimum policies for a given MDP. Suppose π is an optimal policy over some region $[\gamma_1, \gamma_2]$. Let $\hat{P} = P_{\hat{\pi}}, \hat{\mathbf{q}} = \mathbf{q}_{\hat{\pi}}, P = P_\pi$ and $\mathbf{q} = \mathbf{q}_\pi$. Our goal is to compute a subset of the interval $[\gamma_1, \gamma_2]$ where $\hat{\pi}$ is ϵ-optimal. We seek to find γ such that

$$(I - \gamma\hat{P})^{-1}\hat{\mathbf{q}} \geq (1-\epsilon)(I - \gamma P)^{-1}\mathbf{q}$$

or equivalently by defining $\Delta q = \mathbf{q} - \hat{\mathbf{q}}$, and $\Delta P = P - \hat{P}$, Equation 5 below can be solved for γ:

$$\Delta q + (1-\epsilon)(I - \gamma P)^{-1}\mathbf{q} - \gamma \Delta P (I - \gamma\hat{P})^{-1}\hat{\mathbf{q}} \geq 0. \tag{5}$$

However, we must note that $\hat{\pi}$ and π are not necessarily neighboring policies. Therefore, ΔP and Δq may have multiple non-zero entries, and Eq (5) involves multiple polynomials of degree $2N$ that need to be non-negative for the value of γ. This can be solved through univariate polynomial quantifier elimination.

Example 2. Consider the MDP in Example 1 with its optimum policy regions. Let us assume we want to find the range of discount factor $[\gamma_l^{0.1}, \gamma_u^{0.1}]$ for which the difference between value of policy $[1, 1, 1, 1]$, and the optimum policy stays within $0.9v^*$. Figure 1.(b) shows this range.

7 Eliciting Discount Factor for Varying Environments

Thus far, we have seen how a given policy can be optimal for a range of discount factors. Therefore, it is hard to elicit a fixed value discount factor from a single policy. However, if the underlying MDPs rewards and transition probabilities change in time and the player adapts their policy according to each change, we can use this information to "zero-in" on a small range of discount factors. We illustrate our approach through a simple illustrative study wherein we consider an agent making decisions against an environment that changes in time. We will assume that after each environment change, the agent learns an optimal policy through a reinforcement learning scheme that converges faster than the rate at which the environments change. We further assume that we can observe the optimal policy employed by the agent at each time.

Formally, consider a repeated decision making process for an agent which at every stage is modeled by a MDP with $|S|$ states and $|A|$ actions. The agent is considered rational and makes decisions according to a fixed but unknown value of discount factor throughout the experiment. We assume the number of states of the MDP graph and the actions remain the same. However, the rewards associated with state action pairs, and the transition probabilities between states can change at each stage. We are interested in zeroing-in on the unknown discount factor of the agent and therefore predicting the next policy of the agent at every stage by observing its past policies. We use the methodology in this paper to compute the optimum range of discount factors for every policy that the agent adopts at each stage. We compute the intersection of these intervals to narrow down the discount factor range. This refined range is then used to predict the policy of agent in the next stage.

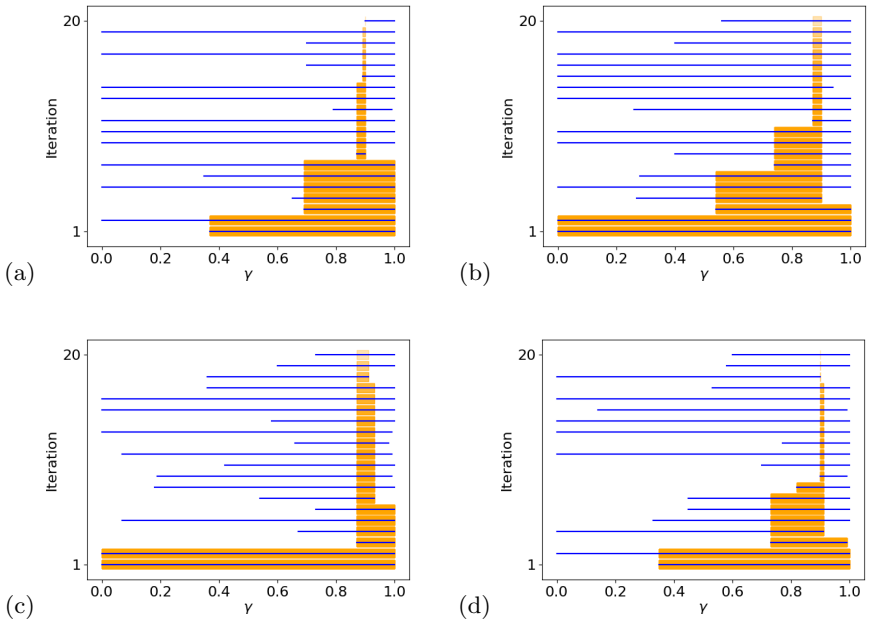


Fig. 2. Simulation results of implementing Algorithm 1 in a MDP with (a) 10 states and 3 actions results in discount factor range $(0.90, 0.90)$, (b) 20 states and 5 actions results $(0.87, 0.90)$, (c) 50 states and 7 actions results $(0.87, 0.91)$, and (d) 100 states and 10 actions result $(0.90, 0.90)$. The discount factor used by the agent in all experiments was 0.9.

Figure 2 shows an illustration of this over four different MDPs with varying number of states. For each instance, we move through 20 different stages that consist of varying transition probabilities and rewards. The discount factor is unknown but is assumed to be fixed for all stages. Algorithm 1 was used with a numerical root finding procedure based on Newton-Raphson method to find ranges of the discount factor. In each case, we identified a single range of discount factors for which the given policy was optimal. In all simulations, the agent's actual discount factor was fixed to 0.9. The estimated range for every simulation is reported in Figure 2. We note that our approach rapidly converges on the underlying discount factor by exploiting the information on the discount factor ranges given from varying environments.

8 Conclusion

This paper studies the problem of inferring the discount factor of an agent by observing their behavior given as a policy on a finite MDP. We employed the notion of *optimum policy regions* for discount factors and characterized the intervals where the optimum policies remain invariant. The boundaries of such

intervals are points at which two *neighboring* optimum policies have the same value. Therefore, the range of discount factors for which a given policy is optimal is identified by points where its value intersects with the value of neighboring optimum policies. We develop numerical approaches to compute these boundary points under different assumptions on the optimality of the policy. We demonstrate the effectiveness of our algorithms through some case studies.

As future directions for this study, we plan to investigate how the discount factor can be inferred by observing partial policies, where the actions for some states are not known. Another possible application of this study is in inverse reinforcement learning and transfer learning, where knowing the discount factor of an agent provides high-quality information regarding their preferences in previously unseen environments. While this paper considers the discount rate to be constant, the need for stateful discounting has been well-argued. For instance, following a sequence of winning streaks, an agent may change its outlook on the horizon. Similarly, institutions may change interest rates based on historical transactions. Similar to the notion of reward machines, our approach can be extended to learn a so-called discount machine to supply discount factors based on the history of an interaction. These discounted machines can be learned by combining ideas presented in this work with active or passive learning of automata.

9 Acknowledgement

We would like to thank the anonymous reviewers for their insightful comments. This work was supported in part by the NSF CAREER Award CCF-2146563, and the NSF IUCRC Center for Autonomous Air Mobility and Sensing (CAAMS).

References

1. Agranov, M., Kim, J., Yariv, L.: Coordination with differential time preferences: Experimental evidence. Tech. rep., National Bureau of Economic Research (2023)
2. Amit, R., Meir, R., Ciosek, K.: Discount factor as a regularizer in reinforcement learning. In: International conference on machine learning. pp. 269–278. PMLR (2020)
3. Blackwell, D.: Discrete dynamic programming. The Annals of Mathematical Statistics pp. 719–726 (1962)
4. Chen, B., Takahashi, S.: A folk theorem for repeated games with unequal discounting. Games and Economic Behavior **76**(2), 571–581 (2012)
5. Collins, G.E., Akritas, A.G.: Polynomial real root isolation using descarte's rule of signs. In: Proceedings of the Third ACM Symposium on Symbolic and Algebraic Computation. p. 272–275. SYMSAC '76, Association for Computing Machinery, New York, NY, USA (1976). https://doi.org/10.1145/800205.806346, `https://doi.org/10.1145/800205.806346`
6. Faddeev, D.K., Faddeeva, V.N., Williams, R.C.: Computational methods of linear algebra (1963)

7. Filar, J., Vrieze, K.: Competitive Markov decision processes. Springer Science & Business Media (2012)
8. Fisher, I.: The theory of interest. New York **43**, 1–19 (1930)
9. François-Lavet, V., Fonteneau, R., Ernst, D.: How to discount deep reinforcement learning: Towards new dynamic strategies. arXiv preprint arXiv:1512.02011 (2015)
10. Giwa, B.H., Lee, C.G.: A marginal log-likelihood approach for the estimation of discount factors of multiple experts in inverse reinforcement learning. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7786–7791 (2021). https://doi.org/10.1109/IROS51168.2021.9636479
11. Gurvich, V., Miltersen, P.B.: On the computational complexity of solving stochastic mean-payoff games. CoRR **abs/0812.0486** (2008), `http://arxiv.org/abs/0812.0486`
12. Hu, H., Yang, Y., Zhao, Q., Zhang, C.: On the role of discount factor in offline reinforcement learning. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 9072–9098. PMLR (17–23 Jul 2022)
13. Jackson, M.O., Yariv, L.: Collective dynamic choice: the necessity of time inconsistency. American Economic Journal: Microeconomics **7**(4), 150–178 (2015)
14. Lehrer, E., Pauzner, A.: Repeated games with differential time preferences. Econometrica **67**(2), 393–412 (1999)
15. Lehrer, E., Solan, E., Solan, O.N.: The value functions of Markov decision processes. Operations Research Letters **44**(5), 587–591 (2016)
16. Littman, M.L., Topcu, U., Fu, J., Isbell, C., Wen, M., MacGlashan, J.: Environment-independent task specifications via gltl. arXiv preprint arXiv:1704.04341 (2017)
17. Mischel, W., Ebbesen, E.B., Raskoff Zeiss, A.: Cognitive and attentional mechanisms in delay of gratification. Journal of personality and social psychology **21**(2), 204 (1972)
18. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013)
19. Patek, S.D., Bertsekas, D.P.: Stochastic shortest path games. SIAM Journal on Control and Optimization **37**(3), 804–824 (1999)
20. Puterman, M.L.: Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons (2014)
21. Roy, M.F., Basu, S., Pollack, R.: Algorithms in real algebraic geometry. Algorithms and Computation in Mathematics **10** (2006)
22. Sagraloff, M., Mehlhorn, K.: Computing real roots of real polynomials. Journal of Symbolic Computation **73**, 46–86 (2016). https://doi.org/https://doi.org/10.1016/j.jsc.2015.03.004, `https://www.sciencedirect.com/science/article/pii/S0747717115000292`
23. Smallwood, R.D.: Optimum policy regions for Markov processes with discounting. Operations Research **14**(4), 658–669 (1966)
24. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
25. Tessler, C.: Deep reinforcement learning works - now what? `https://tesslerc.github.io/posts/drl_works_now_what/` (2020)
26. Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., et al.: Grandmaster level in

starcraft ii using multi-agent reinforcement learning. Nature **575**(7782), 350–354 (2019)
27. Wilf, H.S.: Mathematics for the physical sciences. Courier Corporation (2013)