Online Reinforcement Learning with Uncertain Episode Lengths

Debmalya Mandal[1], **Goran Radanović**[1], **Jiarui Gan**[2], **Adish Singla**[1], **and Rupak Majumdar**[1]

[1]Max Planck Institute for Software Systems
[2] University of Oxford
dmandal@mpi-sws.org, gradanovic@mpi-sws.org, jiarui.gan@cs.ox.ac.uk, adishs@mpi-sws.org, rupak@mpi-sws.org

Abstract

Existing episodic reinforcement algorithms assume that the length of an episode is fixed across time and known a priori. In this paper, we consider a general framework of episodic reinforcement learning when the length of each episode is drawn from a distribution. We first establish that this problem is equivalent to online reinforcement learning with general discounting where the learner is trying to optimize the expected discounted sum of rewards over an infinite horizon, but where the discounting function is not necessarily geometric. We show that minimizing regret with this new general discounting is equivalent to minimizing regret with uncertain episode lengths. We then design a reinforcement learning algorithm that minimizes regret with general discounting but acts for the setting with uncertain episode lengths. We instantiate our general bound for different types of discounting, including geometric and polynomial discounting. We also show that we can obtain similar regret bounds even when the uncertainty over the episode lengths is unknown, by estimating the unknown distribution over time. Finally, we compare our learning algorithms with existing value-iteration based episodic RL algorithms on a grid-world environment.

Introduction

We consider the problem of *episodic reinforcement learning*, where a learning agent interacts with the environment over a number of episodes (Sutton and Barto 2018). The framework of episodic reinforcement learning usually considers two types of episode lengths: either each episode has a fixed and invariant length H, or each episode may have a varying length controlled by the learner. The fixed-length assumption is relevant for recommender systems (Aggarwal et al. 2016) where the platform interacts with a user for a fixed number of rounds. Variable length episodes arise naturally in robotics (Kober, Bagnell, and Peters 2013), where each episode is associated with a learning agent completing a task, and so the length of the episode is entirely controlled by the learner. Fixed horizon lengths make the design of learning algorithms easier, and is the usual assumption in most papers on theoretical reinforcement learning (Azar, Osband, and Munos 2017; Jin et al. 2018).

In this paper, we take a different perspective on episodic reinforcement learning and assume that the length of each episode is drawn from a distribution. This situation often arises in online platforms where the length of an episode (i.e., the duration of a visit by a user) is not fixed a priori, but follows a predictable distribution (Onah, Sinclair, and Boyatt 2014). Additionally, various econometric and psychological evidence suggest that humans learn by maintaining a risk/hazard distribution over the future (Sozou 1998), which can be interpreted as a distribution over the horizon length. Despite a large and growing literature on episodic reinforcement learning, except for (Fedus et al. 2019), uncertain epsiodic lengths or settings with general survival rates of agents have not been studied before.

Our Contributions: In this paper, we describe reinforcement learning algorithms for general distributions over episode lengths. Our main contribution is a general learning algorithm which can be adapted to a given distribution over episode lengths to obtain sub-linear regret over time. In particular, our contributions are the following.

- We first establish an equivalence between maximization of expected total reward with uncertain episode lengths and maximization of expected (general) discounted sum of rewards over an infinite horizon. In particular, we show that minimization of regret is equivalent in these two environments.
- Next we design a learning algorithm for the setting with arbitrary distribution over the episode lengths. Our algorithm generalizes the value-iteration based learning algorithm of Azar, Osband, and Munos (2017) by carefully choosing an effective horizon length and then updating the backward induction step based on the distribution over episode lengths. In order to analyze its regret, we use the equivalence result above, and bound its regret for a setting with general discounting.
- We instantiate our general regret bound for different types of discounting (or equivalently episode distributions), including geometric and polynomial discounting, and obtain sub-linear regret bounds. For geometric discounting with parameter γ, we bound regret by $\widetilde{O}(\sqrt{SAT}/(1-\gamma)^{1.5})$ which matches the recently established minimax optimal regret for the non-episodic setting (He, Zhou, and Gu 2021). For the polynomial discounting of the form h^{-p} we upper bound regret by $\widetilde{O}(\sqrt{SA}T^{\frac{1}{2-1/p}})$.

- Finally, we show that we can obtain similar regret bounds even when the uncertainty over the episode lengths is unknown, by estimating the unknown distribution over time. In fact, for geometric discounting, we recover the same regret bound (i.e. $\widetilde{O}(\sqrt{SAT}/(1-\gamma)^{1.5})$ up to logarithmic factors, and for the polynomial discounting we obtain a regret bound of $\widetilde{O}(\sqrt{SA}T^{\frac{p}{1+2p}})$, which asymptotically matches the previous regret bound.

Our results require novel and non-trivial generalizations of episodic learning algorithms and straightforward extensions to existing algorithms do not work. Indeed, a naive approach would be to use the expected episode length as the fixed horizon length H. However, this fails with heavy-tailed distributions which often appear in practice. Alternately, we could compute an upper bound on the episode length so that with high probability the lengths of all the T episodes are within this bound. Such an upper bound can be computed with the knowledge of distribution over episode lengths and using standard concentration inequalities. However, these upper bounds become loose either with a large number of episodes or for heavy-tailed distributions.

Related Work

Episodic Reinforcement Learning: Our work is closely related to the UCB-VI algorithm of Azar, Osband, and Munos (2017), which achieves $O(\sqrt{HSAT})$ regret for episodic RL with fixed horizon length H. The main difference between our algorithm and UCB-VI is that we use a different equation for backward-induction where future payoffs are discounted by a factor of $\gamma(h+1)/\gamma(h)$ at step h, where γ is a general discount function. Beyond (Azar, Osband, and Munos 2017), several papers have considered different versions of episodic RL including changing transition function (Jin et al. 2018; Jin and Luo 2020), and function approximation (Jin et al. 2020; Wang, Salakhutdinov, and Yang 2020; Yang and Wang 2020).

General Discounting: Our work is also closely related to reinforcement learning with general discounting. Even though geometric discounting is the most-studied discounting because of its theoretical properties (Bertsekas 2012), there is a wealth of evidence suggesting that humans use general discounting and time-inconsistent decision making (Ainslie 1975; Mazur 1985; Green and Myerson 2004). In general, optimizing discounted sum of rewards with respect to a general discounting might be difficult as we are not guaranteed to have a stationary optimal policy. Fedus et al. (2019) study RL with hyperbolic discounting and learn many Q-values each with a different (geometric) discounting. Our model is more general, and our algorithm is based on a modified value iteration. We also obtain theoretical bounds on regret in our general setting. Finally, Pitis (2019) introduced more general state, action based discounting but that is out of scope of this paper.

Stochastic Shortest Path: Our work is related to the stochastic shortest path (SSP), introduced by Bertsekas and Tsitsiklis (1991). In SSP, the goal of the learner is to reach a designated state in an MDP, and minimize the expected total cost of the policy before reaching that goal. Recently, there has been a surge of interest in deriving online learning algorithms for SSP (Rosenberg et al. 2020; Cohen et al. 2021; Tarbouriech et al. 2021). Our setting differs from SSP in two ways. First, the horizon length is effectively controlled by the learner in SSP, once she has a good approximation of the model. But in our setting, the horizon length is drawn from a distribution at the start of an episode by the nature, and is unknown to the learner during that episode. Second, when the model is known in SSP, different policies induce different distributions over the horizon length. Therefore, in contrast to our setting, minimizing regret in SSP is not the same as minimizing regret under general discounting.

Other Related Work: Note that uncertainty over episode lengths can also be interpreted as hazardous MDP (Howard and Matheson 1972), where hazard rate is defined to be the negative rate of change of log-survival time. Sozou (1998) showed that different prior belief over hazard rates imply different types of discounting. We actually show equivalence between general discounting and uncertain episode lengths, even in terms of regret bounds. Finally, this setting is captured by the partially observable Markov decision processes (Kaelbling, Littman, and Cassandra 1998), where one can make the uncertain parameters hidden and/or partially observable.

Model

We consider the problem of episodic reinforcement learning with uncertain episode length. An agent interacts with an MDP $\mathcal{M} = (S, \mathcal{A}, r, \mathbb{P}, \mathtt{P}_{\mathtt{H}})$, where $\mathtt{P}_{\mathtt{H}}$ denotes the probability distribution over the episode length. We assume that the rewards are bounded between 0 and 1. The agent interacts with the environment for T episodes as follows.

- At episode $k \in [T]$, the starting state $x_{k,1}$ is chosen arbitrarily and the length of the episode $H_k \sim \mathtt{P}_{\mathtt{H}}(\cdot)$. [1]
- For $h \in [H_k]$, let the state visited be $x_{k,h}$ and the action taken be $a_{k,h}$. Then, the next state $x_{k,h+1} \sim \mathbb{P}(\cdot | x_{k,h}, a_{k,h})$.

The agent interacts with the MDP $\mathcal{M}$ for T episodes and the goal is to maximize the expected undiscounted sum of rewards. Given a sequence of k episode lengths $\{H_k\}_{k\in[T]}$ the expected cumulative reward of an agent's policy $\boldsymbol{\pi} = \{\pi_k\}_{k\in[T]}$ is given as

$$\text{Rew}\left(\boldsymbol{\pi}; \{H_k\}_{k\in[T]}\right) = \sum_{k=1}^{T} \mathbb{E}\left[\sum_{h=1}^{H_k} r(x_{k,h}, a_{k,h})\right]$$

Since each H_k is a random variable drawn from the distribution $\mathtt{P}_{\mathtt{H}}(\cdot)$, we are interested in expected reward with respect to distribution $\mathtt{P}_{\mathtt{H}}$.

$$\begin{aligned} &\mathbb{E}\left[\text{Rew}\left(\boldsymbol{\pi}; \{H_k\}_{k\in[T]}\right)\right] \\ &= \mathbb{E}\left[\sum_{k=1}^{T}\sum_{H_k=1}^{\infty} \mathtt{P}_{\mathtt{H}}(H_k) \sum_{h=1}^{H_k} r(x_{k,h}, a_{k,h})\right] \\ &= \mathbb{E}_{\pi}\left[\sum_{k=1}^{T}\sum_{h=1}^{\infty} \mathtt{P}_{\mathtt{H}}(H \geq h) r(x_{t,h}, a_{t,h})\right] \qquad (1) \end{aligned}$$

[1] The parameter H_k is unknown to the learner during episode k.

As is standard in the literature on online learning, we will consider the problem of minimizing regret instead of maximizing the reward. Given an episode length H_k and starting state $x_{k,1}$ let $\pi_k^\star$ be the policy that maximizes the expected sum of rewards over H_k steps i.e. $\pi_k^\star \in \text{argmax}_\pi \mathbb{E}_\pi\left[\sum_{h=1}^{H_k} r(x_{k,h}, a_{k,h})|x_{k,1}\right]$. We will write $V^{\pi_k}(x_{k,1}; H_k)$ to write the (undiscounted) value function of a policy π_k over H_k steps starting from state $x_{k,1}$. Then $\pi_k^\star$ is also defined as $\pi_k^\star \in \text{argmax}_\pi V^\pi(x_{k,1}; H_k)$. We will also write $V^\star(x_{k,1}; H_k)$ to denote the corresponding value of the optimal value function. Now we can define the regret over T steps as follows.

Definition 1. *The regret of a learning algorithm* $\boldsymbol{\pi} = \{\pi_k\}_{k \in [T]}$ *over* T *steps with episode lengths* $\{H_k\}_{k\in[T]}$ *is*

$$\textit{Reg}\left(\boldsymbol{\pi}; \{H_k\}\right) = \sum_{k \in [T]} V^\star(x_{k,1}; H_k) - V^{\pi_k}(x_{k,1}; H_k) \quad (2)$$

Note that the regret as defined in eq. (2) is actually a random variable as the episode lengths are also randomly generated from the distribution $\mathtt{P_H}(\cdot)$. So we will be interested in bounding the expected regret. Let $V^\star(x_{k,1})$ be the expected value of $V^\star(x_{k,1}; H_k)$ i.e. $V^\star(x_{k,1}) = \sum_\ell V^\star(x_{k,1}; \ell)\mathtt{P_H}(\ell)$. Then the expected regret of a learning algorithm is given as

$$\text{Reg}(\boldsymbol{\pi}; \mathtt{P_H}(\cdot)) = \sum_{k\in[T]} V^\star(x_{k,1}) - \mathbb{E}_{H_k}\left[V^{\pi_k}(x_{k,1}; H_k)\right]$$

An Equivalent Model of General Discounting

We first establish that the problem of minimizing regret in our setting is equivalent to minimizing regret in a different environment, where the goal is to minimize discounted reward over an infinite horizon with a general notion of discounting. By setting $\gamma(h) = \mathtt{P_H}(H \geq h)$, the expected reward in eq. (1) becomes a sum of T expected rewards under the general discounting function $\{\gamma(h)\}_{h=1}^\infty$.

$$\begin{aligned}&\mathbb{E}\left[\text{Rew}(\boldsymbol{\pi}; \{H_k\}_{k\in[T]})\right]\\ &= \sum_{t=1}^T \mathbb{E}\left[\sum_{h=1}^\infty \gamma(h) r(x_{t,h}, a_{t,h})|x_{k,1}\right]\end{aligned}$$

Therefore, we consider the equivalent setting where the agent is interacting with the MDP $\mathcal{M} = (S, \mathcal{A}, r, \mathbb{P}, \boldsymbol{\gamma})$ where $\boldsymbol{\gamma} = \{\gamma(h)\}_{h=1}^\infty$ is a general discounting factor. We will require the following two properties from the discounting factors:

1. $\gamma(1) = 1$,
2. $\sum_{h=1}^\infty \gamma(h) \leq M$ for some universal constant $M > 0$.

The first assumption is without loss of generality as we can normalize all the discount factors without affecting the maximization problem. The second assumption guarantees that the optimal policy is well-defined. Note that this assumption rules out hyperbolic discounting $\gamma(h) = \frac{1}{1+h}$, but does allow discount factors of the form $\gamma(h) = h^{-p}$ for any $p > 1$. Finally, note that our original reformulation of $\gamma(h) = \mathtt{P_H}(H \geq h)$ trivially satisfies the first assumption. The second assumption essentially ensures that the horizon length has a finite mean. We will also write $\Gamma(h)$ to define the sum of the tail part of the series starting at h i.e.

$$\Gamma(h) = \sum_{j \geq h} \gamma(j) \quad (3)$$

In this new environment, the learner solves the following episodic reinforcement learning problem over T episodes.

Environment: General Discounting

1. The starting state $x_{k,1}$ is chosen arbitrarily.
2. The agent maximizes $\mathbb{E}\left[\sum_{h=1}^\infty \gamma(h) r(x_{k,h}, a_{k,h})|x_{k,1}\right]$ over an infinite horizon.

Notice that even though the new environment is episodic, the length of each episode is infinite. So this environment is not realistic, and we are only introducing this hypothetical environment to design our algorithm and analyze its performance.

Suppose that we are given a learning algorithm $\boldsymbol{\pi} = \{\pi_k\}_{k\in[T]}$. We allow the possibility that π_k is a non-stationary policy as each π_k is used to maximizing a discounted sum of rewards with respect to a general discounting factor and in general the optimal policy need not be stationary. A non-stationary policy π_k is a collection of policies $\{\pi_{k,h}\}_{h=1}^\infty$ where $\pi_{k,h} : (S \times \mathcal{A})^{h-1} \times S \to \Delta(\mathcal{A})$. Given a non-stationary policy π_k at episode k, we define the state-action Q function and the value function as

$$Q^{\pi_k}(x, a; \boldsymbol{\gamma}) = \mathbb{E}\left[\sum_{h=1}^\infty \gamma(h) r(x_{k,h}, a_{k,h}) |\ x, a\right]$$

$$V^{\pi_k}(x; \boldsymbol{\gamma}) = \mathbb{E}\left[\sum_{h=1}^\infty \gamma(h) r(x_{k,h}, a_{k,h}) |\ x\right]$$

Here $a_{k,h} \sim \pi_{k,h}(x_{k,1}, a_{k,1}, \ldots, x_{k,h-1}, a_{k,h-1}, x_{k,h})$ and the conditioning event is $x_{k,1} = x$ and $a_{k,1} = a$ (or just $x_{k,1} = x$). In this environment, we again measure the regret as the sum of sub-optimality gaps over the T episodes.

Definition 2. *Let the optimal value function be defined as* $V^\star(x; \boldsymbol{\gamma}) = \sup_\pi V^\pi(x; \boldsymbol{\gamma})$. *Then we define regret for a learning algorithm* $\boldsymbol{\pi} = \{\pi_k\}_{k\in[T]}$ *as*

$$\textit{Reg}(\boldsymbol{\pi}, \boldsymbol{\gamma}) = \sum_{k=1}^T V^\star(x_{k,1}; \boldsymbol{\gamma}) - V^{\pi_k}(x_{k,1}; \boldsymbol{\gamma}) \quad (4)$$

Our next result shows that it is sufficient to minimize regret with respect to the new environment of episodic reinforcement learning. In fact, if any algorithm has regret $\mathcal{R}(T)$ with respect to the new benchmark, then it has regret at most $\mathcal{R}(T)$ with respect to the original environment with uncertain episode lengths.

Lemma 1. *For any learning algorithm* $\boldsymbol{\pi} = \{\pi_k\}_{k\in[T]}$ *we have the following guarantee:*

$$\textit{Reg}(\boldsymbol{\pi}; \mathtt{P_H}(\cdot)) \leq \textit{Reg}(\boldsymbol{\pi}; \boldsymbol{\gamma}).$$

We also show that a converse of lemma 1 holds with additional restrictions on the discount factor $\boldsymbol{\gamma}$.

Lemma 2. *Suppose the discount factor $\boldsymbol{\gamma}$ is non-increasing. Then there exists a distribution $\mathtt{P}_\mathtt{H}(\cdot)$ over the episode lengths so that*

$$Reg(\boldsymbol{\pi}; \boldsymbol{\gamma}) \le Reg(\boldsymbol{\pi}; \mathtt{P}_\mathtt{H}(\cdot)).$$

Because of lemma 1, it is sufficient to bound a learning algorithm's regret for the environment with infinite horizon and general discounting. Therefore, we now focus on designing a learning algorithm that acts in an episodic setting with uncertain episode lengths, but analyze its regret in the infinite horizon setting with general discounting.

Algorithm: Regret Minimization under General Discounting

We now introduce our main algorithm. Given a non-stationary policy π_k, we define the state-action function and value function at step h as follows.

$$Q_h^{\pi_k}(x,a) = \mathbb{E}\left[\sum_{j=1}^{\infty} \gamma(j) r(x_{k,h+j-1}, a_{k,h+j-1}) \mid \right.$$
$$\left. \mathcal{H}_{h-1}, x_{k,h} = x, a_{k,h} = a\right]$$
$$V_h^{\pi_k}(x) = \mathbb{E}\left[\sum_{j=1}^{\infty} \gamma(j) r(x_{k,h+j-1}, a_{k,h+j-1}) \mid \right.$$
$$\left. \mathcal{H}_{h-1}, x_{k,h} = x\right]$$

where $\mathcal{H}_{h-1} = (x_{k,1}, a_{k,1}, \ldots, a_{k,h-1})$ and $a_{k,h+j} \sim \pi_{k,h+j}(\mathcal{H}_{h+j-1}, x_{k,h+j})$. Note that, both the state-action Q-function and the value function depend on the history $\mathcal{H}_{h-1}$. Moreover, conditioned on the history, we are evaluating the total discounted reward as if the policy $\{\pi_{k,h+j}\}_{j \ge 0}$ was used from the beginning. We first establish some relations regarding the above state-action and value functions. We drop the episode index k for ease of exposition. Given a non-stationary policy $\pi = \{\pi_h\}_{h \ge 1}$ let

$$Q_h^{\pi}(x,a) = r(x,a) + \gamma(2) \cdot \mathbb{E}\left[\sum_{j=1}^{\infty} \gamma(j+1)/\gamma(2) \cdot \right.$$
$$\left. r(x_{h+j}, a_{h+j}) | \mathcal{H}_{h-1}, x_h = x, a_h = a\right]$$
$$= r(x,a) + \gamma(2) \mathbb{E}_{x_{h+1} \sim \mathbb{P}(\cdot|x,a)}\left[\mathbb{E}\left[\sum_{j=1}^{\infty} \gamma(j+1)/\gamma(2) \cdot \right.\right.$$
$$\left.\left. r(x_{h+j+1}, a_{h+j+1}) | \mathcal{H}_h, x_{h+1}\right]\right]$$
$$= r(x,a) + \gamma(2) \mathbb{E}_{x_{h+1} \sim \mathbb{P}(\cdot|x,a)}\left[V_{h+1}^{\pi}(x_{h+1}; \boldsymbol{\gamma}_2)\right]$$

where in the last line we write $\boldsymbol{\gamma}_2$ to denote the discount factor $\boldsymbol{\gamma}_2(j) = \frac{\gamma(j+1)}{\gamma(2)}$ and $V_{h+1}^{\pi}(x_{h+1}; \boldsymbol{\gamma}_2)$ is the value function at time-step h with respect to the new discount factor $\boldsymbol{\gamma}_2$. By a similar argument one can write the action-value function with respect to the discount factor $\boldsymbol{\gamma}_2$ as the following expression.

$$Q_h^{\pi}(x,a; \boldsymbol{\gamma}_2)$$
$$= r(x,a) + \boldsymbol{\gamma}_2(2) \mathbb{E}_{x_{h+1} \sim \mathbb{P}(\cdot|x,a)}\left[V_{h+1}^{\pi}(x_{h+1}; \boldsymbol{\gamma}_2)\right]$$
$$= r(x,a) + \frac{\gamma(3)}{\gamma(2)} \mathbb{E}_{x_{h+1} \sim \mathbb{P}(\cdot|x,a)}\left[V_{h+1}^{\pi}(x_{h+1}; \boldsymbol{\gamma}_3)\right]$$

where the discount factor $\boldsymbol{\gamma}_3$ is given as $\boldsymbol{\gamma}_3(j) = \frac{\gamma(j+2)}{\gamma(3)}$. In general, we have the following relation.

$$Q_h^{\pi}(x,a; \boldsymbol{\gamma}_k) = r(x,a)$$
$$+ \frac{\gamma(k+1)}{\gamma(k)} \mathbb{E}_{x_{h+1} \sim \mathbb{P}(\cdot|x,a)}\left[V_{h+1}^{\pi}(x_{h+1}; \boldsymbol{\gamma}_{k+1})\right] \quad (5)$$

where the discount factor γ_k is defined as $\gamma_k(j) = \frac{\gamma(j+k-1)}{\gamma(k)}$ for $j = 1, 2, \ldots$. Notice that when $\boldsymbol{\gamma}$ is a geometric discounting, we only need equation.

$$Q_h^{\pi}(x,a) = r(x,a) + \gamma \mathbb{E}_{x_{h+1} \sim \mathbb{P}(\cdot|x,a)}\left[V_{h+1}^{\pi}(x_{h+1})\right] \quad (6)$$

ALGORITHM 1: UCB-VI Generalized

```
Input: Discount factor $\{\gamma(h)\}_{h=1}^{\infty}$, parameter $\Delta$
$\mathcal{H} \leftarrow \emptyset$.
for $h = 1, \ldots, N(\Delta)$ do
    Set $Q_{1,h}(x,a) \leftarrow \sum_{j=1}^{\infty} \gamma_h(j) =$
      $\frac{1}{\gamma(h)} \sum_{j=1}^{\infty} \gamma(j+h-1)$ for all $x \in S$ and $a \in \mathcal{A}$.
for $t = 1, \ldots, T$ do
    Update-Q-values($\mathcal{H}, \boldsymbol{\gamma}, \Delta$).
    Receive state $x_{t,1}$.
    for $h = 1, \ldots$ do
        if $h \le N(\Delta)$ then
            Take action $a_{t,h} = \mathrm{argmax}_a Q_{t,h}(x_{t,h}, a)$
            Update $\mathcal{H} = \mathcal{H} \cup (x_{t,h}, a_{t,h}, x_{t,h+1})$
            if $x_{t,h+1}$ is a terminal state then
                Continue to the next episode.
        Take an arbitrary action.
```

Description of the Learning Algorithm : The sequence of recurrence relations eq. (5) motivates our main algorithm (1). Our algorithm is based on the upper confidence value iteration algorithm (`UCBVI` (Azar, Osband, and Munos 2017)). In an episodic reinforcement learning setting with fixed horizon length H, `UCBVI` uses backward induction to update the Q-values at the end of each episode, and takes greedy action according to the Q-table.

However, in our setting, there is no fixed horizon length and the Q-values are related through an infinite sequence of recurrence relations. So, algorithm 1 considers a truncated version of the sequence of recurrence relations eq. (5). In particular, given an input discount factor $\{\gamma(h)\}_{h=1}^{\infty}$[2] and a parameter Δ, algorithm 1 first determines $N(\Delta)$ as a measure of effective length of the horizon. In particular, we set $N(\Delta)$ to be an index so that $\Gamma(N(\Delta)) = \sum_{j \ge N(\Delta)} \gamma(j) \le \Delta$. Note that, such an index $N(\Delta)$ always exists as we assumed that the total sum of the discounting factors converges. Then algorithm 1 maintains an estimate of the Q value for all possible discount factors up to $N(\Delta)$ i.e. $\boldsymbol{\gamma}_k$ for $k = 1, \ldots, N(\Delta)$.

The details of the update procedure is provided in the appendix. In the update procedure, we first set the $(N(\Delta)+1)$-th Q-value to be $\Delta/\gamma(N(\Delta)+1)$ which is always an upper bound on the Q-value with discount factor $\gamma_{N(\Delta)+1}$ because

[2] Recall that $\gamma(h) = \mathtt{P}_\mathtt{H}(H \ge h)$.

of the way algorithm 1 sets the value $N(\Delta)$. Then starting from level $N(\Delta)$, we update the Q-values through backward induction and eq. (5).

Note that our algorithm needs to maintain $N(\Delta)$ action-value tables. We will later show that in order to obtain sublinear regret we need to choose Δ based on the particular discount factor. In particular, for the geometric discount factor $\gamma(h) = \gamma^{h-1}$ we need to choose $N(\Delta) = \frac{\log T}{\log(1/\gamma)}$. On the other hand, discounting factor of the form $\gamma(h) = 1/h^p$ requires $N(\Delta) = O\left(T^{1/(2p-1)}\right)$.

Analysis

The next theorem provides an upper bound on the regret $\mathrm{Reg}(\boldsymbol{\pi};\boldsymbol{\gamma})$. In order to state the theorem, we need a new notation. Let the function $t : N \to \mathbb{R}$ be defined as

$$t(h) = \begin{cases} 1 & \text{if } h = 1 \\ \frac{\gamma(h)}{\gamma(1)} \prod_{j=2}^{h} \left(1 + \frac{\gamma(j)}{j^{\beta}\Gamma(j)}\right) & \text{o.w.} \end{cases}$$

Note that the function t is parameterized by the parameter β and depends on the discount factor $\gamma(\cdot)$.

Theorem 1 (Informal)**.** *With probability at least $1-\delta$, Algorithm 1 has the following regret.*

$$\begin{aligned} Reg(\boldsymbol{\pi};\boldsymbol{\gamma}) &\le \frac{\Delta T}{\gamma(N(\Delta)+1)} t(N(\Delta)+1) \\ &+ \max_{h \in [N(\Delta)]} t(h) \frac{\Gamma(h+1)}{\gamma(h)} \widetilde{O}\left(\sqrt{SATN(\Delta)}\right) \end{aligned}$$

Theorem 1 states a generic bound that holds for any discount factor. The main terms in the bound are $O\left(\sqrt{SATN(\Delta)}\right)$, ΔT, and several factors dependent on the discount factor γ. We now instantiate the bound for different discount factors by choosing appropriate value of Δ and the parameter β.

Corollary 2. *Consider the discount factor $\gamma(h) = h^{-p}$. For $p \ge 2$ and $T \ge O(S^3 A)$ we have*

$$Reg(T) \le \widetilde{O}\left(S^{1/2} A^{1/2} T^{\frac{1}{2-1/p}}\right)$$

and for $1 < p < 2$ and $T \ge O\left((S^3 A)^{\frac{2p-1}{p-1}}\right)$ we have

$$Reg(T) \le \widetilde{O}\left((p-1)^{-\frac{p}{p-1}} S^{1/2} A^{1/2} T^{\frac{1}{2-1/p}}\right)$$

We prove corollary 2 by substituting $\beta = p-1$ and $\Delta = O\left(T^{-\frac{p-1}{2p-1}}\right)$. Note that this result suggests that as p increases to infinity, the regret bound converges to $O(\sqrt{T})$. This also suggests that for exponentially decaying discounting factor, our algorithm should have exactly $O(\sqrt{T})$ regret. We verify this claim next.

Corollary 3. *Consider the discount factor $\gamma(h) = \gamma^{h-1}$ for $\gamma \in [0,1)$ and suppose $T \ge \frac{S^3 A}{(1-\gamma)^4}$. Then algorithm 1 has regret at most*

$$Reg(T) \le \widetilde{O}\left(\sqrt{SAT}/(1-\gamma)^{1.5}\right)$$

Here we substitute $\beta = 3/2$ and $\Delta = T^{-1}/(1-\gamma)$. Our regret bound for the geometric discounting matches the minimax optimal regret bound of the non-episodic setting of (He, Zhou, and Gu 2021).

ALGORITHM 2: Estimating Unknown Discount Factor

```
Input: Horizon Length $H^\star = N(\Delta)$.
Set block length $B = \sqrt{T}\log T \log(\log(T)/\delta)$.
Set $\hat{\gamma}_0$ to be an arbitrary discount factor.
for $j = 0, 1, \ldots, \log(T/B) - 1$ do
    if $j > 0$ then
        $\hat{\gamma}_j(h) = 1 - \hat{F}_H(h-1)$ forall $h$.
    $\hat{\Delta}_j = \sum_{h \ge H^\star + 1} \hat{\gamma}_j(h)$.
    Run algorithm 1 for $2^j B$ episodes with inputs $\hat{\gamma}_j$
      and $\hat{\Delta}_j$.
    /* update empirical distribution
       function                                  */
    $\hat{F}_H(h) = \frac{1}{2^j B}\sum_{t=0}^{2^j B} 1\{H_t \le h\}$.
```

Proof Sketch of Theorem 1 : We now give an overview of the main steps of the proof. Although the proof is based upon the proof of the UCB-VI algorithm (Azar, Osband, and Munos 2017), there are several differences.

- Let $V_h^\star(\cdot)$ be the optimal value function under discounting factor $\gamma_h(\cdot)$ i.e. $V_h^\star(x) = \sup_\pi V^\pi(x;\gamma_h)$. We first show that the estimates $V_{k,h}$ maintained by Algorithm 1 upper bound the optimal value functions i.e. $V_{k,h}(x) \ge V_h^\star(x)$ for any $k, h \in [N(\Delta)]$.
- Let $\widetilde{\Delta}_{k,h} = V_{k,h} - V_h^{\pi_k}$. Then regret can be bounded as

  $$\begin{aligned} &\mathrm{Reg}(\boldsymbol{\pi};\boldsymbol{\gamma}) = \textstyle\sum_{k=1}^T V^\star(x_{k,1}) - V_1^{\pi_k}(x_{k,1}) \\ &\le \textstyle\sum_{k=1}^T V_{k,1}(x_{k,1}) - V_1^{\pi_k}(x_{k,1}) \le \sum_{k=1}^T \widetilde{\Delta}_{k,1}(x_{k,1}) \end{aligned}$$

- Let $\widetilde{\delta}_{k,h} = \widetilde{\Delta}_{k,h}(x_{k,h})$. Then, the main part of the proof of theorem 1 is establishing the following recurrent relation.

  $$\begin{aligned} \widetilde{\delta}_{k,h} &\le \tfrac{\gamma(h+1)}{\gamma(h)}\left(1 + \tfrac{\gamma(h+1)}{(h+1)^\beta \Gamma(h+1)}\right)\widetilde{\delta}_{k,h+1} \\ &+ \sqrt{2L}\bar{\varepsilon}_{k,h} + e_{k,h} + b_{k,h} + \varepsilon_{k,h} + f_{k,h} \end{aligned}$$

  Here $\bar{\varepsilon}_{k,h}$ and $\varepsilon_{k,h}$ are Martingale difference sequences and $b_{k,h}, e_{k,h}, f_{k,h}$ are either the bonus term or behave similarly as the bonus term.
- We complete the proof by summing the recurrence relation above over all the episodes and from $h = 1$ to $N(\Delta)$. Although (Azar, Osband, and Munos 2017) established a similar recurrence relation, there are two major differences. First the multiplicative factor in front of $\widetilde{\delta}_{k,h+1}$ is changing with time-step h and is not a constant. This is because the backward induction step uses eq. (5) in our setting. Second, after expanding the recurrence relation from $h = 1$ to $N(\Delta)$ the final term is no longer zero and an extra $O(\Delta T)$ term shows up in the regret bound.

Estimating the Discount Function

In this section we consider the situation when the discount function $\gamma(h) = \mathtt{P}_{\mathtt{H}}(H \ge h)$ is not unknown. We start with the assumption that the optimal value of $N(\Delta)$ (say $H^\star$) is known. The next lemma bounds the regret achieved by running an algorithm with $N(\Delta) = H^\star$ with the true discounting γ and an estimate of the discounting $\hat{\gamma}$. Our algorithm partitions the entire sequence of T episodes into blocks of lengths $B, 2B, 2^2B, \ldots, 2^sB$ for $s = \log(T/B) - 1$. At the end of each block the algorithm recomputes an estimate of γ. Recall that we defined $\gamma(h) = \Pr(H \geq h)$. Since every episode we get one sample from the distribution of H (the random length of the current episode) we can use the empirical distribution function of horizon length to obtain $\hat{\gamma}$. At the end of block B, the algorithm computes $\hat{\gamma}_B$, and runs algorithm 1 with this estimate and $\hat{\Delta}_B = \hat{\Gamma}_B(H^\star + 1) = \sum_{h \geq H^\star + 1} \hat{\gamma}_B(h)$ for the block $B + 1$.

Theorem 4 (Informal)**.** *When run with horizon length $H^\star$, algorithm 2 has the following regret bound with probability at least $1 - \delta$*

$$\begin{aligned} \textit{Reg}(\pi; \gamma) &\leq \min_{L \in [T]} \left(T\Gamma(L+1) + 2L \log(T)\sqrt{T}) \right) \\ &+ \max_{h \in [H^\star]} \frac{t(h)}{\gamma(h)} g(h) \left(1 + \frac{O(T^{-1/4})}{\Gamma(h+1)}\right) \widetilde{O}\left(\sqrt{SATH^\star}\right) \\ &+ \Gamma(H^\star)T \end{aligned}$$

where $g(h) = \exp\left\{ O\left(\sum_{k=2}^{h} \frac{T^{-1/4}}{\gamma(k) + k^\beta \Gamma(k)} \right) \right\}$.

Our proof relies on bounding the estimation error of $\hat{\gamma}$ and $\hat{\Gamma}$. We can use the classical DKW inequality (Dvoretzky, Kiefer, and Wolfowitz 1956) to bound the maximum deviation between empirical CDF ($\widehat{\mathtt{P}}_{\mathtt{H}}(\cdot)$) and true CDF ($\mathtt{P}_{\mathtt{H}}$). Through a union bound over the $\log(T)$ blocks, this immediately provides a bound between $\left\|\hat{\gamma}_j - \gamma_j\right\|_\infty$ for all $j \in [\log(T/B)]$. However, we also need to bound the distance between $\hat{\Gamma}_j(\cdot)$ and $\Gamma(\cdot)$ for all j (defined in (3)). A naive application of DKW inequality results in an additive bound between $\hat{\Gamma}_j(h)$ and $\Gamma(h)$ that grows at a rate of h. This is insufficient for our case to get a sublinear regret bound. However, we show that we can use the multiclass fundamental theorem (Shalev-Shwartz and Ben-David 2014) to derive an error bound that grows at a rate of $\sqrt{\log h}$ and this is sufficient for our proof.

The main challenge in the proof of theorem 4 is controlling the growth of the term $t(h)/\gamma(h)$. Notice that this term is a product of h terms of the form $1 + \frac{\gamma(k)}{k^\beta \Gamma(k)}$, so any error in estimating γ could blow up the product by a factor of h. We could show that the regret is multiplied by an additional function $g(h)$ which is parameterized by β. We next instantiate theorem 4 for different discount factors and show that we can obtain regret bounds similar to corollary 2, and 3 up to logarithmic factors.

Corollary 5. *Consider the discount factor $\gamma(h) = h^{-p}$ for $p \geq 2$. Then the regret of algorithm 2 is*

$$\textit{Reg}(T) \leq \begin{cases} \widetilde{O}\left(\sqrt{SAT}^{\frac{p+1}{2p}}\right) & \textit{if } T \geq O\left((S^{3/2}A^{1/2})^p\right) \\ \widetilde{O}\left(S^2AT^{\frac{1}{2-1/p}}\right) & \textit{if } T \leq O\left((S^{3/2}A^{1/2})^p\right) \end{cases}$$

Corollary 6. *Suppose $\frac{T}{\log^3 T} \geq \frac{S^3A}{(1-\gamma)^4}$. Then algorithm 2 has regret at most $\widetilde{O}\left(\sqrt{SAT}/(1-\gamma)^{1.5}\right)$ for geometric discounting $\gamma(h) = \gamma^{h-1}$ for $\gamma \in [0, 1)$.*

For the polynomial discounting we get a regret of the order of $T^{(p+1)/2p}$ which is worse than the regret bound of theorem 1 by a factor of $T^{1/2p}$. However, the difference goes to zero as p increases and approaches the same limit of $\widetilde{O}(\sqrt{T})$. On the other hand, for geometric discounting we recover the same regret as corollary 3. Interestingly, He, Zhou, and Gu (2021) obtained a similar bound on regret for the non-episodic setting where the learner maximizes her long-term geometrically distributed reward.

Unknown $N(\Delta)$: Note that algorithm 2 takes as input the optimal value of $N(\Delta)$ or $H^\star$. However, this problem can be handled through a direct application of model selection algorithms in online learning (Cutkosky et al. 2021). Let $\text{Reg}(H^\star)$ be the regret when algorithm 2 is run with true $H^\star$. We now instantiate algorithm 2 for different choices of $H^\star$ and perform model selection over them. In particular, we can consider $H^\star = 2, 2^2, \ldots, 2^{O(\log T)}$ as it is sufficient to consider $H^\star = O(T)$. Moreover, given true $H^\star$ there exists $\widetilde{H} \leq 2H^\star$ for which the regret is increased by at most a constant. This step requires bounding $\frac{t(H^\star)}{\gamma(H^\star)} / \frac{t(\widetilde{H})}{\gamma(\widetilde{H})}$ and is constant for the discounting factors considered in the paper. We now apply algorithm 1 from (Cutkosky et al. 2021) to the collection of $O(\log T)$ models and obtain a regret bound of at most $O\left(\sqrt{\log T}\text{Reg}(\widetilde{H})\right) = \widetilde{O}(\text{Reg}(H^\star))$.

Experiments

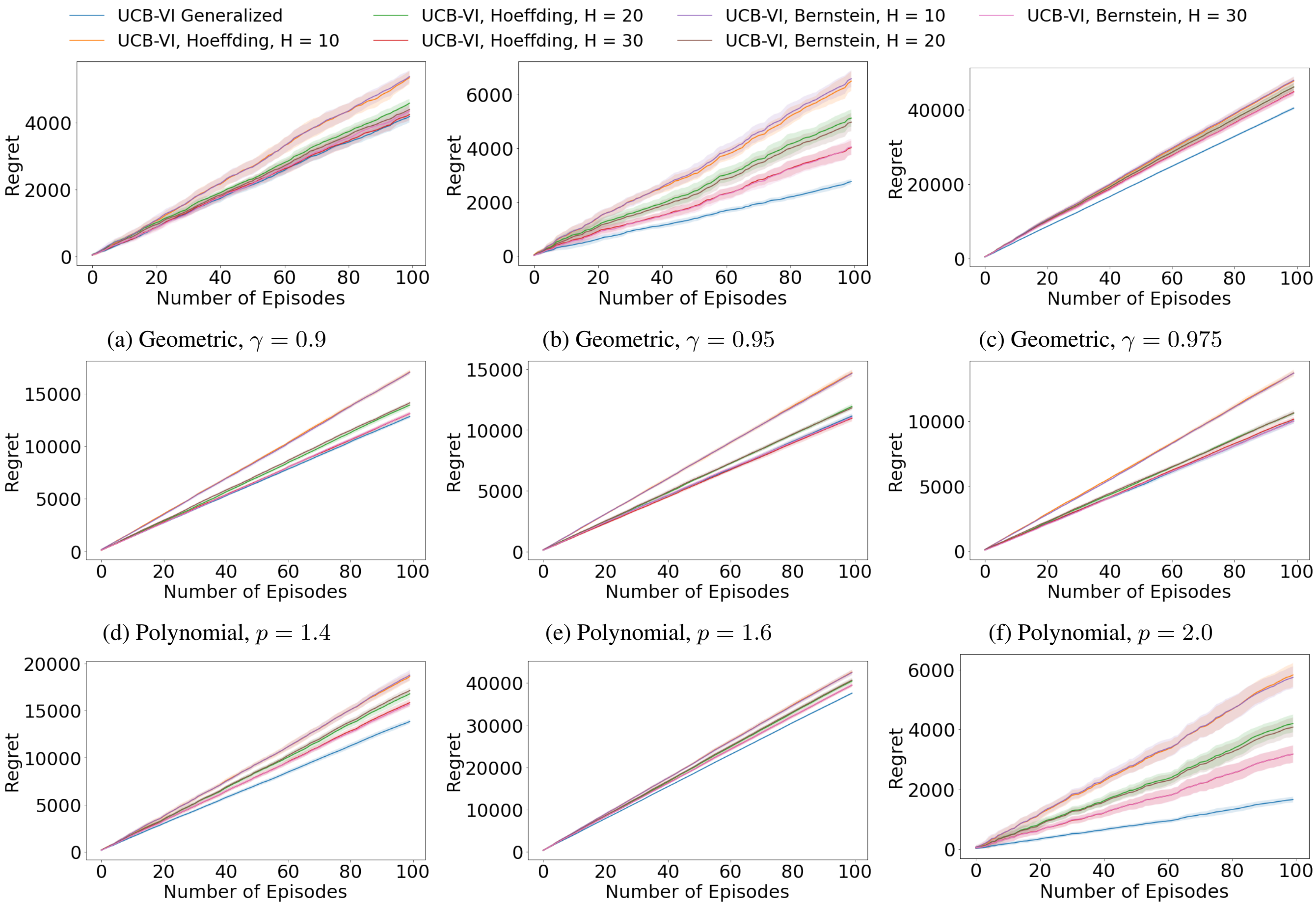
We evaluated the performance of our algorithm on the Taxi environment, a 5×5 grid-world environment introduced by (Dietterich 2000). The details of this environment is provided in the appendix, since the exact details are not too important for understanding the experimental results. We considered 100 episodes and each episode length was generated uniformly at random from the following distributions. [3]

1. Geometric discounting $\gamma(h) = \gamma^{h-1}$.
2. Polynomial discounting $\gamma(h) = h^{-p}$.
3. Quasi-Hyperbolic discounting $\gamma(h) = \beta^{1\{h>1\}}\gamma^{h-1}$

Figure 1 shows some representative parameters for three different types of discounting. For the geometric discounting, we show $\gamma = 0.9, 0.95$ and 0.975. For the polynomial discounting we generated the horizon lengths from a polynomial with $p \in \{1.4, 1.6, 2.0\}$ and added an offset of 20. Finally, for the Quasi-hyperbolic discounting, we fixed γ at 0.95 and considered three values of β: $0.7, 0.8$, and 0.9.

We compared our algorithm (1) with two variants of UCB-VI (Azar, Osband, and Munos 2017) – (a) UCB-VI-Hoeffding computes bonus terms using Chernoff-Hoeffding inequality, and (b) UCB-VI-Bernstein computes bonus terms using Bernstein-Freedman inequality. It is known that

[3] Here $\gamma(h)$ refers to probability that the episode lengths exceeds h i.e. $\gamma(h) = \Pr(H \geq h)$.



(a) Geometric, $\gamma = 0.9$

(b) Geometric, $\gamma = 0.95$

(c) Geometric, $\gamma = 0.975$

(d) Polynomial, $p = 1.4$

(e) Polynomial, $p = 1.6$

(f) Polynomial, $p = 2.0$

(g) Quasi-Hyperbolic, $\beta = 0.7$

(h) Quasi-Hyperbolic, $\beta = 0.8$

(i) Quasi-Hyperbolic, $\beta = 0.9$

Figure 1: Comparison of our algorithm with different variants of UCB-VI on the Taxi environment (Dietterich 2000). The regret is measured over 100 episodes, and the length of each episode is drawn independently from a given distribution. Each plot shows average regret and standard error from 10 trials.

when the horizon length is fixed and known, UCB-VI-Bernstein achieves minimax optimal regret bounds. We implemented two versions of UCB-VI with three different assumed horizon lengths.

Figure 1 shows that, for several situations, our algorithm strict improves in regret compared to all the other variants of UCB-VI. These include Geometric discounting ($\gamma = 0.95$ and 0.975) and Quasi-Hyperbolic discounting (all possible choices of β). For the other scenarios (e.g. polynomial discounting), our algorithm performs as well as the best version UCB-VI. Figure 1 also highlights the importance of choosing not only the right horizon length but also the correct update equation in backward induction. Consider for example, figure 1b for the geometric discounting with $\gamma = 0.95$. Here the expected horizon length is $\frac{1}{1-\gamma} = 20$. However, different UCB-VI variants (horizon lengths $10, 20, 30$ and Bernstein and Hoeffding variants) perform worse. Our algorithm benefits by choosing the right effective horizon length, and also the correct update equation (6).

Conclusion

In this paper, we have designed reinforcement learning algorithms when the episode lengths are uncertain and drawn from a fixed distribution. Our general learning algorithm (1) and result (theorem 1) can be instantiated for different types of distributions to obtain sub-linear regret bounds. An interesting direction of future work is to extend our algorithm to function approximation (Jin et al. 2020). For the standard linear MDP model, the least squares value iteration based algorithm (LSVI-UCB) (Jin et al. 2020) solves a regularized least squares to learn a parameter vector w_h for each step h. We can follow a similar approach for estimation, however we would need to re-weight the outcome variables by the discount factor in the least-squares problem. We are also interested in other models of episode lengths. For example, one can consider a setting where the lengths are adversarially generated but there is a limit on the total amount of change. This is similar to the notion of variation budget (Besbes, Gur, and Zeevi 2014) considered in the literature on non-stationary multi-armed bandits.

Acknowledgements

Goran Radanović acknowledges that his research was, in part, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

References

Aggarwal, C. C.; et al. 2016. *Recommender systems*, volume 1. Springer.

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

Azar, M. G.; Osband, I.; and Munos, R. 2017. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 263–272. PMLR.

Bertsekas, D. 2012. *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific.

Bertsekas, D. P.; and Tsitsiklis, J. N. 1991. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3): 580–595.

Besbes, O.; Gur, Y.; and Zeevi, A. 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27.

Cohen, A.; Efroni, Y.; Mansour, Y.; and Rosenberg, A. 2021. Minimax regret for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34: 28350–28361.

Cutkosky, A.; Dann, C.; Das, A.; Gentile, C.; Pacchiano, A.; and Purohit, M. 2021. Dynamic balancing for model selection in bandits and rl. In *International Conference on Machine Learning*, 2276–2285. PMLR.

Dietterich, T. G. 2000. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of artificial intelligence research*, 13: 227–303.

Dvoretzky, A.; Kiefer, J.; and Wolfowitz, J. 1956. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 642–669.

Fedus, W.; Gelada, C.; Bengio, Y.; Bellemare, M. G.; and Larochelle, H. 2019. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*.

Green, L.; and Myerson, J. 2004. A discounting framework for choice with delayed and probabilistic rewards. *Psychological bulletin*, 130(5): 769.

He, J.; Zhou, D.; and Gu, Q. 2021. Nearly minimax optimal reinforcement learning for discounted MDPs. *Advances in Neural Information Processing Systems*, 34: 22288–22300.

Howard, R. A.; and Matheson, J. E. 1972. Risk-sensitive Markov decision processes. *Management science*, 18(7): 356–369.

Jin, C.; Allen-Zhu, Z.; Bubeck, S.; and Jordan, M. I. 2018. Is Q-learning provably efficient? *Advances in neural information processing systems*, 31.

Jin, C.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2020. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2137–2143. PMLR.

Jin, T.; and Luo, H. 2020. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in neural information processing systems*, 33: 16557–16566.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2): 99–134.

Kober, J.; Bagnell, J. A.; and Peters, J. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11): 1238–1274.

Mazur, J. E. 1985. Probability and delay of reinforcement as factors in discrete-trial choice. *Journal of the Experimental Analysis of Behavior*, 43(3): 341–351.

Onah, D. F.; Sinclair, J.; and Boyatt, R. 2014. Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 proceedings*, 1: 5825–5834.

Pitis, S. 2019. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7949–7956.

Rosenberg, A.; Cohen, A.; Mansour, Y.; and Kaplan, H. 2020. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, 8210–8219. PMLR.

Shalev-Shwartz, S.; and Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Sozou, P. D. 1998. On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1409): 2015–2020.

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

Tarbouriech, J.; Zhou, R.; Du, S. S.; Pirotta, M.; Valko, M.; and Lazaric, A. 2021. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in Neural Information Processing Systems*, 34: 6843–6855.

Wang, R.; Salakhutdinov, R. R.; and Yang, L. 2020. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33: 6123–6135.

Yang, L.; and Wang, M. 2020. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, 10746–10756. PMLR.