TÉCNICO LISBOA

Discounting Strategies for Planning in Hazardous Environments

Guilherme Eduardo Roque Salvador

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisors: Prof. Francisco António Chaves Saraiva de Melo
Eng. Pedro Pinto Santos

Examination Committee

Chairperson: Prof. João António Madeiras Pereira
Supervisor: Prof. Francisco António Chaves Saraiva de Melo
Member of the Committee: Prof. José Alberto Rodrigues Pereira Sardinha

November 2024

Declaration
I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

During this long and challenging journey, I have been surrounded by people who have supported me in many different ways. I would like to express my gratitude to all of them and the next lines are dedicated to them.

My first words could not be to anyone other than my supervisors, Prof. Francisco Melo and Pedro Santos. I would like to deeply thank them for their guidance and support during this journey. I am extremely grateful for their patience, availability, and the time they dedicated to me in meetings, discussions, and feedback.

I would like to thank each and everyone of my family for their unconditional support, love, and for always believing in me. A special thanks to my parents, who have always been there for me, providing me with the best conditions to pursue my objectives, and to my brother, who has always been a great companion.

I also would like to express my gratitude to my friends, who shared with me the good and the bad moments, and who have always been there to support me. A particular thanks to my housemates, who have been a great company during this time and made this journey much more enjoyable. As it could not be different, I would like to express my appreciation to my friends and course colleagues, who have been the best support during this time. Thank you for the great moments we shared procrastinating and having fun.

Finally, I want to acknowledge the partial financial support provided by the national funds through the Portuguese Fundação para a Ciência e a Tecnologia (FCT) under project UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020) (INESC-ID multi-annual funding) and PTDC/CCI-COM/5060/2021 (RELEvaNT).

To all of you, thank you for being part of this journey.

Abstract

In this work, we study the development of agents to deal with environments featuring unknown, possibly non-stationary, hazard rates using their discounting process. First, we study the benefits of an emerging discounting technique, the hyperbolic discounting, in the context of hazardous tasks and compare it with the standard exponential discounting. We show that hyperbolic discounting can improve the performance of agents in hazardous tasks, especially when the hazard rates are non-deterministic, being sampled from random distributions at each episode. Second, we propose a framework incorporating an adaptive discount factor, aiming to adapt the standard exponential discounting to the uncertainty of the environment. The proposed framework allows the agents to make online adjustments between episodes of interaction, increasingly improving their performance. Our solution has shown to be robust in the presence of both stationary and non-stationary hazard rates, having a close performance to agents acting under the premise of full knowledge of the hazard rates.

Keywords

Planning; Hyperbolic Discounting; Adaptive Discount Factor; Uncertainty; Hazard.

Resumo

Neste trabalho, estudamos o desenvolvimento de agentes para lidar com a incerteza em ambientes que apresentam taxas de risco desconhecidas, possivelmente não estacionárias, adaptando seu método de desconto. Numa primeira fase, estudamos as vantagens da utilização de um método de desconto alternativo, o desconto hiperbólico, no que consta à sua aplicação em ambientes com taxas de risco desconhecidas, comparando ainda o seu desempenho com o desconto exponencial. Mostramos que o desconto hiperbólico pode melhorar o desempenho dos agentes em tarefas com taxas de risco, especialmente quando as taxas de risco são não determinísticas, sendo estas amostradas de distribuições aleatórias no início de cada episódio. Propomos ainda um algoritmo que incorpora um fator de desconto adaptável, visando adaptar o desconto exponencial padrão à incerteza inerente ao ambiente. O algoritmo proposto permite que os agentes façam ajustes dinâmicos entre episódios de interação, melhorando progressivamente o seu desempenho. A nossa proposta revelou-se robusta na presença de taxas de risco estacionárias e não estacionárias, apresentando um desempenho próximo ao de agentes que atuam sob a premissa de conhecimento total das taxas de risco inerente às tarefas.

Palavras Chave

Planeamento; Desconto Hiperbólico; Fator de Desconto Adaptativo; Incerteza; Risco.

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

GVF	General Value Function
MDP	Markov Decision Process
RL	Reinforcement Learning
VI	Value Iteration

1

Introduction

Contents

Planning in hazardous environments is a challenging task that requires agents to balance the risk and uncertainty associated with distant-future rewards, aiming to maximize the obtained rewards. However, in continuing tasks, those without a definitive terminal state that could persist indefinitely, the expected cumulative reward can be infinite, resulting in issues regarding expected reward maximization. In order to address this problem, conventional planning algorithms take advantage of temporal discounting, a key aspect of the decision-making process that involves decreasing the value of future rewards in contrast to immediate rewards. Moreover, it can be seen as a mechanism that allows agents to balance the inherent risk and uncertainty associated with distant-future rewards. Therefore, the discounting process has a clear connection with the concept of risk and uncertainty, since the devaluation of future rewards can be seen as a way to quantify the amount of risk that the agent expects to face until the moment of reward receipt.

Most planning algorithms apply exponential discounting, reducing the value of future rewards exponentially. This approach assumes that the discount factor is constant over time and it is very convenient

since it simplifies the learning process and gives theoretical convergence guarantees. Furthermore, exponential discounting is time-consistent, i.e., the preference of a reward over another is maintained regardless of the delay to their receipt, as long as the temporal delay between them remains the same. However, it is usually assumed that the discount factor, being part of the problem definition, is known and fixed. This assumption may not always be realistic, especially in scenarios where the agents face hazardous environments, where they can at any time be exposed to hazardous events that terminate the undergoing task. Therefore, relying on a predetermined constant discount factor may not be the best approach since it also implies the assumption of a known and constant hazard rate, the risk of a hazardous event occurring at each time step.

Moreover, some works on psychology and economics have shown that humans and animals exhibit dynamic shifts in their preferences over time [6], not discounting future rewards exponentially. Since the human/animal decision-making process has an intrinsic concern with risk and uncertainty, modeling these behaviors can be beneficial in addressing problems where there is a need to balance the risk and uncertainty associated with distant-future rewards. Even planning algorithms are powerful approaches for addressing decision-making problems, the standard temporal preference setting used in these algorithms does not accurately capture the actual behavior observed in humans and animals, falling short in modeling correctly the preference shifts observed in real-world scenarios. Since the standard exponential discounting cannot accurately model the human decision-making process, several works propose the adoption of hyperbolic discounting [3, 7, 8] to better model behavior observed in humans and animals.

Considering the preceding discussion, in this work, we investigate the following research question: *How might the discounting process be adapted to improve the performance of agents in hazardous environments?* Therefore, we explore the development of agents for hazardous environments, where the agents are exposed to unknown and possibly non-stationary hazard rates. We address scenarios involving stochasticity in the hazard rate of the environment, where the hazard rate can change over time being sampled from an underlying random distribution. We also consider scenarios where the hazard rate being deterministic, can be either stationary, remaining constant over time, or non-stationary, changing over time. The aforementioned scenarios are modeled as a hazardous variant of the standard Markov Decision Process (MDP) model, known as hazardous MDPs. This modified version, introduced by Fedus et al. [2], differs from the standard MDP model by incorporating a hazard rate, λ, which determines the risk of occurrence of a hazardous event that leads to the abrupt end of the episode and the loss of all future rewards.

In this work, we focus on improving the performance of agents in hazardous planning tasks, modeled as hazardous MDPs. Therefore, we consider settings where the agent has full knowledge of the environment's dynamics, except the hazard rate. Nonetheless, we expect that the proposed methodologies can be easily extended to traditional Reinforcement Learning (RL) settings, where environments'

dynamics are not fully known and agents must learn from experience. We explore two main approaches to improve the performance of agents in planning tasks involving hazardous MDPs. First, we study the practical application of hyperbolic discounting in standard algorithms, exploring its benefits in hazardous tasks, especially when involving stochasticity in the environments' hazard rates. Second, we propose a framework that takes advantage of an online adaptation of the discount factor to deal with uncertainty over the hazard rate of the environment, being well suited to stationary or non-stationary deterministic hazard rates. Therefore, we propose the development of an agent, based on a tabular method, Value Iteration (VI), that incorporates an adaptive discounting framework, allowing the tuning of its estimate of the discount factor over time. By adapting the discount factor to the environment, the agent can have a better performance in environments with associated hazards since they can adopt a more short-sighted or long-sighted behavior depending on their estimation of the hazard rate.

1.1 Contributions

Taking into account the limitations of standard exponentially discounted agents and the theoretical benefits of both hyperbolic discounting and discount factor adaptation techniques in hazardous environments, the main contributions of this work are the following:

1. Perform a systematic study comparing exponential and hyperbolic discounting models, discussing the main differences between both discounting approaches and empirically demonstrating the advantages of hyperbolic discounting in hazardous environments with stochastic hazard rates.

2. Propose an adaptive discounting framework that allows the agent to estimate the hazard rate of the environment and adjust the discount factor accordingly, introducing an alternative approach that outperforms conventional exponentially discounted agents in tasks involving unknown, either stationary or non-stationary, hazard rates.

1.2 Document Outline

The remainder of this document starts by presenting, in Chapter 2, the theoretical background that is essential to understanding the main concepts and techniques used in this work. In Chapter 3, we provide a comprehensive review of the literature, where we present some of the most relevant works involving discounting mechanisms that deal with the presence of uncertainty in the environment, enhancing the role of hyperbolic discounting and discount factor adaptation techniques. In Chapter 4, we perform a comparative analysis of both exponential discounting and hyperbolic discounting, evaluating their performance in different environments with several hazardous settings. Following this, in Chapter 5, we

present an adaptive discounting framework and the mechanisms used to estimate the hazard rate of the environment, detailing all the implementation steps. In Chapter 6, we present the experimental results obtained by applying the proposed adaptive discounting framework in different environments, comparing its performance with standard algorithms using both exponential and hyperbolic discounting. Finally, in Chapter 7, we present the conclusions of this work, discussing the main achievements and addressing some of the limitations and future work directions.

2

Background

Contents

This section provides a brief overview of the main concepts used as the basis for the development of our work.

2.1 Markov Decision Processes

A MDP is a mathematical framework that models decision-making problems where an agent interacts with an environment over a series of discrete time steps [1].

2.1.1 Formal Definition

Formally, we define an MDP as a tuple $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$, composed of the following elements:

- $\mathcal{S}$ - state space, a set of possible states of the environment;

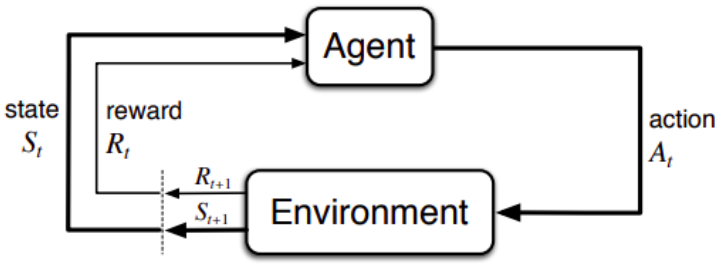

Figure 2.1: Agent/environment interaction (from Sutton and Barto [1]).

- $\mathcal{A}$ - action space, a set of actions that the agent can select during the interaction;
- $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ - transition probabilities, a function that defines the probability of transitioning from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$ when taking action $a \in \mathcal{A}$. It is important to note that transition probabilities must satisfy the condition $\sum_{s' \in \mathcal{S}} p(s'|s, a) = 1, \forall s \in \mathcal{S}, a \in \mathcal{A}$. By $p(s'|s, a)$ we denote the probability of transitioning from state s to state s' when taking action a;
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ - reward function, a function that defines the expected reward received by the agent when taking action $a \in \mathcal{A}$ in the state $s \in \mathcal{S}$;
- $\gamma \in [0, 1[$ - discount factor, a parameter that defines the rate at which future rewards are devalued in the agent's decision-making process.

The agent's interaction with an environment, modeled as an MDP, unfolds through discrete time steps. Illustrated by the agent-environment interaction depicted in Figure 2.1, at each time step t, the agent observes the current state of the environment, $S_t \in \mathcal{S}$, and chooses an action, $A_t \in \mathcal{A}$. Subsequently, at time step $t + 1$, the environment provides a numerical reward R_{t+1} and transitions to the new state, $S_{t+1} \in \mathcal{S}$, according to the transition probabilities denoted as $p(.|S_t, A_t)$.

It is also important to note that, as suggested by the name, the decisions involving MDPs follow the *Markov property*

$$\mathbb{P}[S_{t+1} = s'|S_{0:t}, A_{0:t}] = \mathbb{P}[S_{t+1} = s'|S_t, A_t]. \tag{2.1}$$

In simple words, this property assures that, at each time step, the system's future state is determined exclusively by the current state and selected action, being independent of the path leading to that state.

2.1.2 Learning Policies

Given all the elements that characterize an MDP, the agent's ultimate goal is to learn a policy that allows it to maximize its expected cumulative reward, during the interaction with the environment. Therefore, the notion of policy is central to the MDP framework. A policy consists essentially of a mapping between each possible environment's state and an action. Formally, a policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ corresponds to

a probability distribution over the action space $\mathcal{A}$ for each state $s \in \mathcal{S}$, representing the probability of taking each action a in each state s.

A fundamental concept in the process of solving MDPs is the concept of value function. The value function assigns a value to each state of the environment, representing the expected cumulative reward that the agent expects to receive from that state onwards, provided that it follows a given policy π. The value function can be defined as

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | S_0 = s\right]. \tag{2.2}$$

Regarding the inclusion of discounting in the computation of the value function, this definition features an exponential discounting factor $\gamma \in [0, 1[$ that is used to balance the trade-off between short-term and long-term rewards and to avoid infinite expected cumulative rewards during the learning process.

It is also possible to define a function returning the expected cumulative reward that the agent can expect to receive from a given state onwards when taking an initial state-action pair and following a given policy π afterwards. This function is commonly referred to as Q-function and can be defined as

$$Q^{\pi}(s, a) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | S_0 = s, A_0 = a\right]. \tag{2.3}$$

Since the agent's goal is to maximize its expected cumulative reward, many algorithms are built upon the computation of optimal value functions from which it is possible to derive the optimal policy π^*. In the concrete case of state-value functions, the optimal function, V^*, must satisfy

$$V^*(s) \geq V^{\pi}(s), \quad \forall s \in \mathcal{S}, \quad \forall \pi. \tag{2.4}$$

The same reasoning can be applied to the computation of the optimal Q-function, Q^*.

When all the elements of an MDP are fully known, it is possible to compute the optimal value functions, such as V^* and Q^*. Being optimal value functions, they satisfy the Bellman optimality equation, [9], defined as

$$V^*(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s'} p(s'|s, a) V^*(s') \right\}, \tag{2.5}$$

and derive the optimal policy π^* defined as

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s, a), \quad \forall s \in \mathcal{S}. \tag{2.6}$$

2.1.2.A Value Iteration

One of the most popular dynamic programming algorithms for solving MDPs is the VI algorithm, which iteratively computes the optimal value functions by applying the Bellman optimality equation [9]. This method maintains an estimate of the optimal value function and updates iteratively the value until the estimate converges to the optimal value function or a predefined stopping criterion is met. After the convergence of the value function, the optimal policy can be derived by selecting the action that maximizes the Q-value of each state.

The algorithm is widely used in practice due to its simplicity and efficiency in solving MDPs with a small number of states and actions and can be used to solve several types of planning problems. A pseudo-code of the VI algorithm is presented in Algorithm 2.1.

Algorithm 2.1: Value Iteration Algorithm

Input: $\mathcal{S}$: Set of all states, γ: Discount factor, $p(s'|s,a)$: Transition probabilities, $r(s,a)$: Reward function
Output: V: Optimal value function
begin
  $V(s) \longleftarrow 0, \forall s \in \mathcal{S}$
  $\epsilon \longleftarrow$ small positive number
  $\Delta \longleftarrow \infty$
  while $\Delta > \epsilon$ **do**
    $\Delta \longleftarrow 0$
    for $s \in \mathcal{S}$ **do**
      $temp \longleftarrow V(s)$
      $V(s) \longleftarrow \max_a \{r(s,a) + \gamma \sum_{s'} p(s'|s,a) V(s')\}$
      $\Delta \longleftarrow \max(\Delta, |temp - V(s)|)$
  return V

2.2 Hazardous Markov Decision Processes

In order to model environments with uncertain hazard rates, Fedus et. al [2] introduced the concept of hazardous MDPs. In this hazardous version of MDPs, the agent is not required to discount future rewards but is subject to a constant probability of death at each time step by the introduction of a hazard rate, λ, defined by Sozou [3] as the risk per unit time of a hazard occurring, provided that it has not occurred in a previous time step.

Formally, this adaptation of the MDP framework can be defined, similarly to the original MDP, as a tuple $(\mathcal{S}, \mathcal{A}, p_\lambda, r, \mathcal{H})$. The main difference to the original MDP is the introduction of the hazard rate distribution $\mathcal{H}$ from which the environment samples the hazard rate $\lambda \in [0, \infty[$ at the beginning of each episode. Moreover, the transition probabilities p_λ are conditioned on the hazard rate, λ, to take into

account the possibility of the agent's death during the interaction with the environment, being defined as $p_\lambda(s'|s,a) = e^{-\lambda}p(s'|s,a)$ [2] where p are the transition probabilities of a conventional MDP. In practice, this adaptation can be implemented by introducing a new absorbing state in the state space, reachable from any other state with a probability $1 - e^{-\lambda}$. This absorbing state represents the agent's death, meaning that the agent will not be able to receive any rewards after reaching this state. Although this modification of the MDP is more suitable for modeling hazardous environments, it is still closely related to the concept of discounting in the original MDP framework due to the tight relation between the concepts of risk and discounting.

2.3 Bayesian Inference

Bayesian inference is a statistical framework that allows the extraction and update of beliefs about unobserved quantifiers by combining prior knowledge about such quantifiers with the observation of new data, being highly suitable for decision-making problems involving uncertainty. This framework is fundamentally grounded on Bayes' theorem, which provides a way to update the prior beliefs about a problem with the observation of new data. The theorem can be defined as

$$p(\theta|D) = \frac{p(\theta) \cdot p(D|\theta)}{p(D)}, \tag{2.7}$$

being composed of the following elements:

- $p(\theta)$ - prior distribution, represents the prior knowledge about the problem being a density function that quantifies the probability of each possible value of the parameter θ;
- $p(\theta|D)$ - posterior distribution, represents the updated distribution of the parameter θ after incorporating the observation of the data D;
- $p(D|\theta)$ - likelihood function, a density function that measures how likely it is to observe the data D given the parameter θ;
- $p(D)$ - marginal likelihood, a normalization factor representing the overall probability of observing the data D, integrated over all possible values of the parameter θ.

By applying the Bayes' theorem, prior beliefs about a problem can be updated as new data is observed. This iterative process refines the posterior distribution, leading to a progressively greater level of certainty about the underlying problem.

As later detailed in Chapter 5, we apply a Bayesian approach to estimate the unknown hazard rate of the environments. We now introduce the distributions used in the rest of this work.

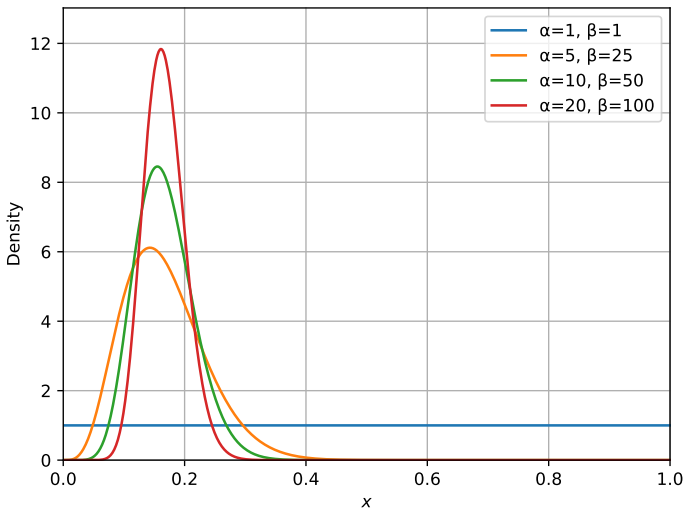

Figure 2.2: Probability density function of the Beta distribution for different values of the shape parameters, α and β.

The first distribution we introduce is the geometric distribution, which is commonly used to represent the number of trials required to achieve the first success in a sequence of Bernoulli trials and is defined as

$$P(X = k) = (1 - p)^{k-1} \cdot p, \quad k = 1, 2, 3, ..., \tag{2.8}$$

where p corresponds to the probability of success in each trial, and X is the random variable representing the number of trials before the first success occurs.

We also use the Beta distribution, whose corresponding probability density function is illustrated in Figure 2.2, being defined as

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}, \tag{2.9}$$

where α and β are the shape parameters of the distribution, and Γ is the gamma function. This distribution serves as a conjugate prior for the geometric distribution [10].

The proper computation of the Bayes rule for an arbitrary combination of prior and likelihood functions can be hard due to the absence of a simple closed-form solution. Therefore, the adoption of a conjugate prior for the geometric distribution, the Beta distribution, simplifies the computation by providing a closed-form solution for the posterior distribution. This property not only allows the posterior distribution to be used as the prior for the next update but also simplifies the computation by making the posterior a straightforward adjustment of the shape parameters, α and β. After each new observation, k, the parameters are updated according to

$$\begin{cases} \alpha' = \alpha + 1, \\ \beta' = \beta + k \end{cases},$$

where α' and β' are the updated shape parameters of the Beta distribution.

As the number of coherent observations increases, the posterior distribution becomes more peaked and eventually converges to a Dirac delta function.

3

Related Work

Contents

Over the years, several researchers have addressed the role of discounting, as well as the limitations of standard exponential discounting models when applied to scenarios involving risk and uncertainty. This section provides an overview of the most relevant research works regarding the topic.

3.1 Discounting and human behavior

Experimental studies in humans and animals have shown that the value attributed by each individual to a reward does not remain constant over time. Raclin et al. [7] showed that animals tend to prefer immediate small rewards over larger delayed rewards, due to either the uncertainty of receiving the

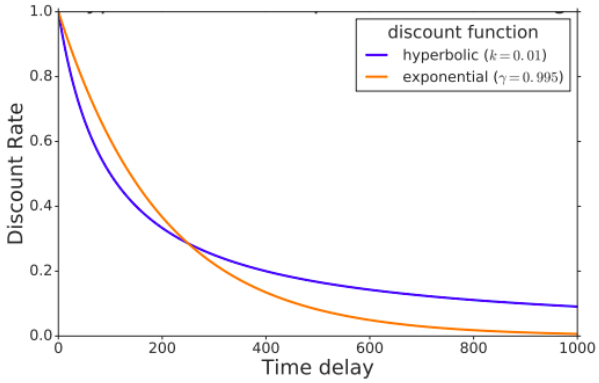

Figure 3.1: Exponential and hyperbolic discounting curves (adapted from Fedus et al. [2]).

delayed reward or the extra effort required to obtain it. This leads to the conclusion that the subjective value of a reward decreases as the delay to its receipt increases [11] This phenomenon is the basis of temporal discounting models.

One of the most prevalent and well-studied discounting models is exponential discounting, a broadly applied model in economics that states that the value of a reward, r_0, over time is "perceived" as

$$r(t) = r_0 \gamma^t, \tag{3.1}$$

after a delay of t time steps. The discount factor $\gamma < 1$ is the parameter that controls the rate at which the value of future rewards decreases. It is worth noting that, when the interaction between agents and environments occurs in a set of discrete time steps, this discounting model is more properly referred to as geometric discounting.

Nonetheless, studies involving both humans and animals have yielded evidence endorsing the idea that the standard exponential discounting model is not a good fit for the observed behavior [7, 8]. More precisely, it has been shown that individuals do not discount future rewards exponentially by applying a constant discount rate γ to all rewards, independently of the delay to their receipt. Instead, the discount rate seems to be higher for rewards with shorter delays, and lower for rewards with longer delays. To accommodate the observed behavior, researchers have introduced the hyperbolic discounting model, which assigns to a reward a current value of

$$r(t) = r_0 \frac{1}{1 + kt}, \tag{3.2}$$

where k is a predefined constant influencing the curvature of the discounting curve.

The differences between the exponential and hyperbolic discounting curves are shown in Figure 3.1.

Furthermore, there are significant distinctions in the temporal consistency of the decision-making

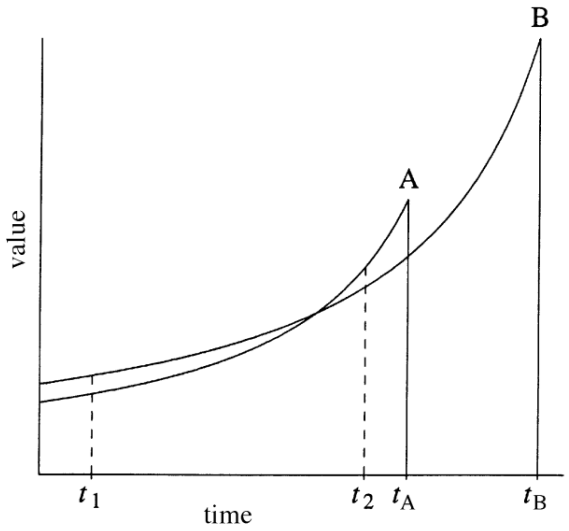

Figure 3.2: Preference reversals. The rewards A and B are received at time steps t_A and t_B respectively. The agent prefers B over A at time t_1, but prefers A over B at time t_2 (adapted from Sozou [3]).

.

process between hyperbolic discounting and exponential discounting.

The conventional exponential discounting model is based on the assumption that the discount rate is constant over time. This implies that, from an agent's perspective, the probability of surviving in the environment between two time steps stays constant over time. This assumption leads to a time-constant preference ordering, meaning that if the agent prefers a reward A over a reward B at some moment in time, this preference remains unchanged regardless of the time step at which the rewards are received.

On the other hand, in the hyperbolic discounting model, the assumption of a time-independent discount factor does not hold, leading to situations where the agent's actions result in non-stationary preference orderings since the preference between two rewards can change solely due to different delays in their receipt. This phenomenon is known as preference reversal and is referred to by Green et al. [6] as a trace in human behavior and can be modeled by hyperbolic discounting [12] as shown in Figure 3.2.

3.2 Survival analysis and uncertainty

Discounting functions play a crucial role in decision-making processes since their shape directly impacts the agent's choices. While certain functions may lead the agent to prioritize short-term rewards, others can promote a more forward-thinking approach, emphasizing the accumulation of long-term rewards. Obtaining an equilibrium between immediate and delayed rewards is also interesting within the realm of survival analysis as the agent's lifespan significantly influences the choices made during action selection.

The discount factor can alternatively be represented as a hazard rate. Sozou [3] defines this concept

as the risk per unit time of the hazard occurring, provided that it has not occurred in a previous time step. The author also states that, assuming a constant hazard rate, λ, the probability of receiving a reward after a delay of t time steps is given by

$$\sigma(t) = e^{-\lambda t}. \tag{3.3}$$

Moreover, the value of a reward after a delay, t, can be defined as a function of, λ, by rewriting (3.1) as

$$r(t) = r_0 e^{-\lambda t}, \quad \lambda \in [0, \infty[, \tag{3.4}$$

establishing the direct relation between the discount factor γ and the hazard rate λ as $e^{-\lambda} = \gamma$.

Following the same line of thought, in addition to the advantages hyperbolic discounting offers in modeling preference reversals, several researchers suggest that it is more suitable when the hazard rate characterizing the target environment is unknown and stochastic [2, 3]. Given that the discount factor is typically a predetermined parameter set before the agent engages with the environment, an inappropriate choice for the discount factor when the hazard rate is not known can result in suboptimal policies. In contrast, hyperbolic discounting avoids assuming a constant discount factor. Moreover, some studies have demonstrated that the behavior of agents employing hyperbolic discounting can be approximated by combining a group of agents assuming distinct discount factors [2, 3, 13], each one associated with a distinct hazard rate. This approach, by requiring the agent to learn from a set of different discount factors, enhances a better adaptability of hyperbolic discounting to a wider range of hazard rates.

3.3 Hyperbolic discounting models

In order to leverage the benefits of hyperbolic discounting, several lines of work have been developed to adapt conventional algorithms to incorporate hyperbolic discounting.

Learning over multiple timescales

To address the issue of learning in environments with constant but unknown hazard rates, Sozou [3] proposes the *direct superposition method* which allows the agent to learn using a set of different hazard rates simultaneously, combining the resulting policies to obtain a final policy. For instance, in a scenario where the environment can draw hazard rates from a discrete set of values $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, the survival function $\sigma(t)$, i.e, the probability of the agent surviving until time t, can be expressed as

$$\sigma(t) = \sum_{i=1}^{n} p_i e^{-\lambda_i t}. \tag{3.5}$$

where p_i is the probability of the environment drawing the hazard rate λ_i.

The suggested approach was expanded to a broader scenario where the environment's hazard rate is parameterized by a continuous probability distribution $p(\lambda)$. The resulting expected survival rate is obtained by integrating exponential hazard rates over the entire range of possible values as

$$\sigma(t) = \int_0^\infty p(\lambda) e^{-\lambda t} d\lambda. \tag{3.6}$$

Employing hazard rates sampled from a continuous probability distribution represents a more flexible strategy, enhancing the agent's adaptability in learning within environments with varying degrees of risk and uncertainty.

Sozou [3] departs from (3.6) to show that an exponential prior distribution of the form

$$p(\lambda) = \frac{1}{k} e^{-\lambda/k}, \tag{3.7}$$

satisfies,

$$\int_0^\infty p(\lambda) e^{-\lambda t} d\lambda = \frac{1}{1+kt}, \tag{3.8}$$

yielding the hyperbolic discounting function.

Sozou [3] also studied several other prior distributions for the underlying hazard rates and their impact on the resulting survival function, from which we highlight the gamma prior $p(\lambda) = \frac{(\lambda/k)^{c-1} exp(-\lambda/k)}{k\Gamma(c)}$ where Γ is the gamma function and c a positive constant, yielding the survival function

$$\sigma(t) = \frac{1}{(1+kt)^c}. \tag{3.9}$$

It is worth noting that the parameter c can change significantly the prior distribution function, leading to different survival functions. For instance, when $c = 1$ the gamma prior reduces to the exponential prior and its associated survival function is, consequently, the one matching the hyperbolic discounting model as in (3.2). On the other hand, when $c \to \infty$ the gamma prior converges to a Dirac delta function which models a constant known hazard rate c/k and yields the survival function associated with the conventional exponential discounting.

The comparison between the application of the several prior distributions has shown that the different priors lead to different non-exponential survival functions. However, Sozou [3] also compared the shape of the resulting survival functions and showed that significant changes in the prior distributions for the hazard rate lead to survival functions that, although different, remain relatively close to the hyperbolic discounting curve.

Sozou [3] also states that the studied prior distributions for the hazard rate can be updated in order to incorporate new information about the environment's hazard rate, such as the agent's survival time. This

update process can be performed by following a Bayesian approach, updating the prior distributions to better represent the environment's hazard rate.

Subsequently, Fedus et al. [2] introduced a similar strategy to adapt temporal difference methods, specifically Q-learning, by leveraging the flexibility provided by non-exponential discounting. The standard Q-learning algorithm proposed by Watkins and Dayan [14] is a RL algorithm that uses an exponential discounting factor, γ, to devalue future rewards. This factor is usually part of the problem definition and is set before the agent starts interacting with the environment. The discount factor imposes an expected time scale on the learning process and creates a time horizon beyond which the agent does not consider future rewards. To overcome this limitation, the authors proposed an alternative version of the algorithm that computes non-exponential discounting functions, such as the hyperbolic case, while applying conventional discounting techniques that retain the convergence guarantees of the original algorithm. More specifically, their work showed that distributing the learning process across different time horizons by learning distinct value functions concurrently, each one of them associated with a configuration featuring a different hazard rate, can surpass the performance of conventional value-based methods in environments with uncertain hazard rates.

To compute non-exponentially discounted Q-values, Fedus et al. [2] went through an approach similar to the one proposed by Sozou [3] to approximate the hyperbolic survival functions, as shown in (3.6). The authors showed that non-exponential discount functions can be approximated by integrating the exponential discount function over all possible time horizons, weighted by prior distributions of γ, $p(\gamma)$, equivalent to the priors of λ used by Sozou [3]. The expression used to approximate these discount functions is given by

$$d(t) = \int_0^\infty p(\gamma)\gamma^t d\gamma. \tag{3.10}$$

Consequently, non-exponential discounted Q-values can be computed as

$$\begin{aligned} Q_\pi^\Gamma(s,a) &= \int_0^1 p(\gamma)\mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t r_(s_t,a_t) \mid S_0 = s, A_0 = a\right] d\gamma \\ &= \int_0^1 p(\gamma) Q_\pi^\gamma(s,a) d\gamma, \end{aligned} \tag{3.11}$$

where

$$p(\gamma) = \frac{1}{k}\gamma^{1/k-1}, \quad k > 0. \tag{3.12}$$

Therefore, the authors applied the Q-learning algorithm with different discount factors and computed a set of tables of Q-values for each one of them. These discrete set of tables were then combined to approximate the integral in (3.11) and obtain the final Q-values for the hyperbolic discount function.

Over the years, the principle of approximating non-exponential discount functions, combining inde-

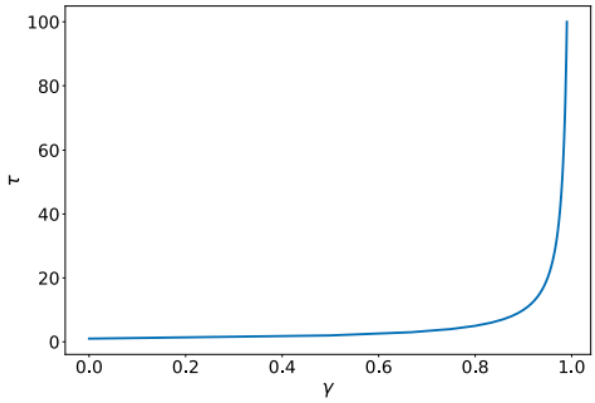

Figure 3.3: Relation between the discount factor γ and the expected time horizon τ (from Sherstan [4]).

pendent value functions associated with different time horizons, has been applied and discussed in other research works [13, 15, 16]. For instance, Kurth-Nelson et al. [13] presented a distributed system composed of several independent agents, learning independent representations of the environment and applying a different exponential discount factor. However, the entire system acts as a single agent, combining the different representations using a voting mechanism to select the action to be executed. Consequently, the overall system's behavior stays very close to the expected behavior of a single agent using a hyperbolic discount factor.

Developing value functions across multiple time scales for approximating non-exponentially discounted functions constitutes a highly convenient strategy. However, one of the main drawbacks of these methods is the need to learn from a potentially infinite set of time horizons since the exponential discount factors belong to a continuous range of values.

A simple alternative to tackle this challenge involves discretizing the range of potential discount factors. This approach opens the question of how to choose the set of discount factors to be used in the approximation, as well as the number of values to be considered. The most straightforward solution is to sample discount factors from a uniform distribution. However, this solution does not take into account the fact that value functions do not change linearly with the discount factor. This non-linear relation, illustrated in Figure 3.3, was derived by Sherstan et al. [5] and follows

$$\tau = \frac{1}{1-\gamma}, \tag{3.13}$$

where for a given discount factor, γ, the expected time horizon, τ, corresponds to the expected number of time steps until the agent dies in the environment or until the future rewards become irrelevant.

The relation shows that uniformly sampling discount factors would favor the learning of value functions associated with shorter time horizons since only a small portion of the sampled values would be associated with long timescales. In order to balance the learning of value functions across different time

horizons, Sherstan et al. [5] suggested sampling discount factors uniformly from both discount factor, γ, and timescale, τ, scales simultaneously to obtain a more balanced set of discount factors. Since both discount factor and time horizon scales are inversely related, while the sampling of discount factors from the γ scale favors the learning of value functions associated with shorter time horizons, sampling from the τ scale favors the learning of value functions associated with longer time horizons. Consequently, sampling from both scales simultaneously leads to a more balanced set of discount factors, promoting learning value functions across different time horizons.

Reinke et al. [16] addressed a similar issue regarding the choice of the cardinality of the set of discount factors to be used in the approximation, since a richer set of discount factors can lead to a more accurate approximation of the target function but also increases the computational cost of the learning process. Their framework, *Average Reward Independent Gamma Ensemble*, is composed of several independent Q-learning modules learning over different timescales. To address the mentioned issue, the authors tested their algorithm with a distinct number of modules and showed that the performance of the algorithm does not linearly increase with the number of modules. Instead, the performance stabilizes after a short number of modules. These results provide relevant insights regarding the choice of the cardinality of the set of discount factors to be used in this type of approximation since they do not rely on large sets to achieve the best performance, allowing the use of a smaller set of discount factors and reducing the computational cost of the learning process.

Hyperbolically Discounted Temporal Difference

Another line of research was explored by Alexander and Brown [17] who proposed an alternative adaptation to temporal difference methods to accommodate hyperbolic discounting. The authors suggested a new learning algorithm entitled *Hyperbolically Discounted Temporal Difference* which is based on the standard temporal difference learning algorithm but implements a recursive formulation of hyperbolic discounting. While in standard temporal difference learning methods, the update rule requires the computation of the difference between the value of the current state and the discounted value of the next state, in their approach, the error term is given by the difference between the value of the current state and the value of the next state discounted by a factor that depends on the value of the current state. More specifically, the temporal difference error term is given by an expression similar to the following

$$\delta_t = r(s_t, a_t) + \frac{1 - kV(s_t)}{m} V(s_{t+1}) - V(s_t), \tag{3.14}$$

where the estimated value of the next state $V(s_{t+1})$ is discounted by a factor $\frac{1-kV(s_t)}{m}$ that depends on the value of the current state $V(s_t)$. The discounting factor also depends on a positive value k that refers to the curvature of the hyperbolic discounting function and a positive scaling factor m dependent on the

magnitude of the rewards and prevents the discounting factor from varying with the rewards' magnitude.

Regarding the influence of rewards' magnitude on the discounting process, even though the work of Alexander and Brown [17] prevents the discount factor from scaling with the rewards' magnitude, other works suggest that in human behavior, the temporal discounting process is amount-dependent [12], resulting in lower discount factors for larger rewards and higher discounting rates for smaller rewards.

3.4 General discounting models

While hyperbolic discounting has gained importance for its ability to capture time inconsistency in human preferences, several lines of research have explored a wider range of discounting models. Although not as suitable for modeling human behavior, these approaches aim to provide a more versatile framework for modeling the temporal aspects of reward collection, accommodating a wider range of scenarios and patterns. By understanding and incorporating diverse discounting strategies, researchers addressed the limitations associated with specific discounting models and improved the generalizability of RL algorithms.

Aligned with this research line, General Value Functions (GVFs) were first introduced by Sutton et al. [18] as a general framework for estimating value functions. This paper introduced *Horde*, an architecture composed of a set of independent agents, each one of them responsible for dealing with a specific aspect of the environment, contributing to the overall knowledge of the system as a whole. Therefore, GVFs are well-suited for tasks where the learning agent needs to simultaneously consider multiple and potentially conflicting objectives.

Regarding the discounting approach, GVFs are characterized by allowing more flexible approaches to the discounting process. Instead of relying on a fixed discount factor, GVFs can feature a state-dependent discounting function that consists of a mapping that associates each state in the environment to a different discount factor [18]. This approach linking the discount factors to different states of the environment was led even further by White [19] who proposed a similar method that associates independent discount factors to different state-action pairs, achieving a transition-dependent discounting function.

These specific discounting methods are, naturally, more complex and not as well studied and understood as the standard fixed exponential discounting approach. However, they provide a more flexible framework that can obtain interesting results when applied to deal with scenarios involving risk and uncertainty where agents face the risk of abrupt episode termination at each timestep. Therefore, taking different actions at different states can lead to different outcomes that can involve more or less risk to the agent's continuity in the environment. Consequently, mapping a different discount factor to each transition or state can model this type of scenario more accurately.

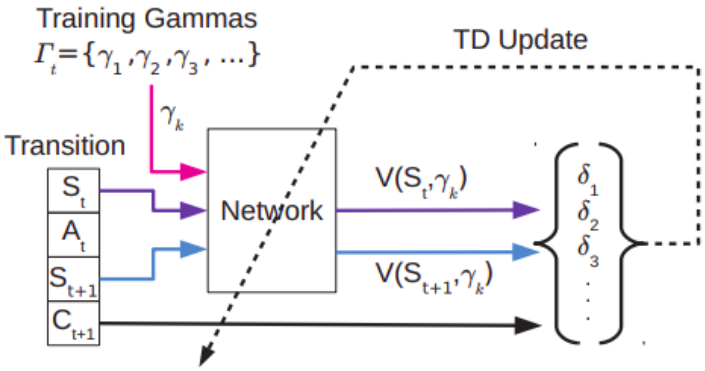

Figure 3.4: The Γ-net training process (from Sherstan et al. [5]).

Using as a basis the concept of general value functions, Sherstan et al. [5] presented Γ-*nets* (Figure 3.4), a framework for generalizing value function estimation over multiple timescales, taking advantage of the generalization power of neural networks. Their approach consists of a neural network whose function is to approximate the value function taking as input a set of discount factors. The training process involves sampling a different set of discount factors for each transition and computing the temporal difference errors for each one of them. The TD-errors are then used to update the weights of the neural network, allowing it to approximate value functions for a wide range of discount factors. Their approach was tested in a set of experiments including a set of Atari video games, obtaining promising results and having also a small cost in performance when compared to the standard fixed exponential discounting approach.

3.5 Adapting the discount factor

Choosing an appropriate discount factor for the learning process is a challenging task. As previously mentioned, the discount factor is usually a fixed parameter that is set before the agent starts interacting with the environment. While small discount factors lead to a more myopic behavior, prioritizing short-term rewards, large discount factors promote a more forward-thinking approach, emphasizing the accumulation of long-term rewards.

In continuing tasks with no terminal state, the choice of the discount factor is not a straightforward task process since it imposes a timescale on the learning process. Depending on the chosen discount factor, the maximized expected return at the imposed timescale may not correspond to the optimal behavior in the continuing task. For instance, in a scenario with sparse rewards, a short timescale may discard the possibility of obtaining a reward in the far future, leading to suboptimal choices. To overcome this limitation, it has been proved the existence of a critical discount factor, γ^*, dependent

on the environment's dynamics, that imposes a sufficiently large timescale on the learning process. Therefore, choosing a discount factor, γ, such that $\gamma^* < \gamma$ allows the learning of a policy, known as Blackwell-optimal policy [20], that maximizes the expected return in the long run. Since the critical discount factor is dependent on the environment's dynamics, it is not known a priori and needs to be estimated during the learning process. To do so, naively increasing the discount factor is not a viable solution since most algorithms suffer from instability problems when the discount factor is very close to 1 [21]. This issue was addressed by Tang et al. [22] who proposed a method to interpolate the value function associated with a specific discount factor from the value function for a lower discount factor, using Taylor expansions. As a consequence, one of the main advantages of their approximation method lies in the fact that it allows for avoiding the instability problems associated with high discount factors by computing the corresponding value function departing from a lower and stable discount factor.

From a different perspective, in tasks involving hazard and uncertainty, the choice of the discount factor is also a challenging task, particularly due to the fact that the hazard rate is unknown and stochastic. As outlined in Section 3.3, in these scenarios, agents may learn from a set of different discount factors, combining the resulting policies to obtain a final generic policy. However, a different line of research has been exploring different methods focusing on leveraging the agent's experience to adjust the discount factor throughout the learning process.

One concrete example of this type of approach is the work of Xu et al. [23] who took advantage of the generalization power of neural networks to tailor agents to diverse scenarios during real-time interactions with the environment. Their methodology consists of a *meta-gradient* algorithm that is able to adjust meta-parameters takien as input, solely deriving from sequences of interactions with the environment. Specifically, the algorithm, when applied to $TD(\lambda)$, can meta-learn both the discount factor γ and the temporal difference parameter λ. Moreover, it also supports the meta-parameter adaptation in a state-dependent setting, allowing variations in the discount factor and temporal difference parameter across different states.

Hafner et al. [24] proposed *Dreamer*, a model-based RL agent designed to handle complex visual control tasks that use high-dimensional pixel observations as input. Their approach involves learning a world model by observing a dataset of past experiences and encoding both observations and actions into a compressed latent space. In this compact latent space, the agent can plan and learn from fully imagined trajectories, which improves sample efficiency by reducing the number of required interactions with the environment. *Dreamer* can also be applied in early termination tasks, where the episode can end at any time step, such as when the agent achieves a specific goal. Consequently, every latent state in an imagined trajectory can potentially be a terminal state. Therefore, *Dreamer* associates a predicted discount factor to each latent state, which allows the agent to weigh down the future rewards in the imagined trajectory according to the probability of each latent state being a final state.

Recently, Kim et al. [25] suggested an update rule for adjusting the discount factor, tested on both on-policy and off-policy algorithms. To find an appropriate value for the discount factor, the authors' adaptive algorithm is composed of two parameters that constitute the limits of a range for possible discount factors. Since the purpose of this family of adaptive algorithms is to perform better in environments with associated uncertainty, their approach uses the differences between the expected and actual returns to adjust the bounds of the discount factor range during the training phase, successively reducing the range of possible discount factors until the optimal value is found. Their proposal was tested in a set of experiments involving on-policy and off-policy algorithms and performed better when compared to a simpler approach that incrementally increases the discount factor during the training phase.

3.6 Discussion

As discussed, over the years, several researchers have been exploring different approaches to adapt the learning process to the specificities of different environments and different tasks. In order to overcome the limitations of the standard exponential discounting model, researchers have proposed:

1. The hyperbolic discounting to adapt the learning process to the uncertainty of the environment and better approximate the discounting process applied by humans in choices involving risk;

2. Generic discounting models to provide a more versatile framework that can accommodate a wider range of scenarios and patterns by including, for instance, state-dependent discounting functions;

3. Learning over multiple timescales to split the learning process across different time horizons, allowing the agent to learn value functions associated with different discount factors simultaneously;

4. Adaptation mechanisms for the discount factor to obtain the best performance in continuing tasks, avoiding the instability problems associated with high discount factors.

Regarding the application of these methods in order to develop agents that can deal with hazardous environments, a considerable line of research has been exploring the adoption of hyperbolic discounting models to adapt the learning process to a closer approximation of the discounting process applied by humans in choices involving uncertainty. However, even though this discounting approach can model phenomenons such as preference reversals, the literature lacks clear comparisons between both hyperbolic and exponential discounting when applied in hazardous environments.

Following the reasoning of approximating the hyperbolic discounting model, several research works explored the idea of approximating the newer discounting functions by splitting the learning process across different time horizons [2, 3, 13, 15, 16]. This family of methods obtained promising results in experiments involving risk and uncertainty, showing that the learning process can be adapted to these

scenarios by learning value functions across different time horizons. However, in scenarios with a fixed and unknown hazard rate, learning over multiple horizons introduces an inherent error since it requires the agent to adapt and generalize its knowledge across potential time scales. Hence, their flexibility comes at a small cost in precision compared to employing the conventional exponential discounting model in a context where the hazard rate was known. Moreover, these methods ignore the fact that, after the training phase, agents can still collect valuable clues about the environment's hazard rate. Therefore, agents should be able to adapt their behavior to the environment's hazard rate after the training phase, leading to a more precise discounting process and being more robust to changes in the environment's dynamics. To overcome this limitation, adjusting the discount factor interactively can be a good alternative. However, the literature is still scarce in providing clear answers on how to adapt the discount factor in such a way that agents can make those adjustments in a dynamic way, during the interaction with the environment.

Given the research gaps identified in the literature, in this work, we contribute by exploring two distinct research directions. First, given the theoretical benefits of hyperbolic discounting in hazardous environments, we perform a systematic comparison between the hyperbolic and exponential discounting models in hazardous environments, tackling the lack of clear comparisons between these two discounting models in such scenarios. We explore the performance of agents employing both discounting models and enhance, with empirical results, the advantages of hyperbolic discounting in hazardous environments, especially under stochastic hazard rates. Second, as existing discount factor estimation methods from the literature are not directly tailored for hazardous environments, we introduce a novel approach that adjusts the discount factor interactively during the agent's interaction with the environment. This allows the agent to adapt to the hazard rate and remain robust to changes in underlying uncertainty conditions.

4

Exploring Differences Between Hyperbolic and Exponential Discounting

Contents

In this chapter, we present a systematic study of the main differences between hyperbolic and exponential discounting and the implications of these differences in contexts of uncertainty over the environment's hazard rate. To properly understand those differences, we develop a set of tests to compare the performance of agents using both discounting methods in different scenarios.

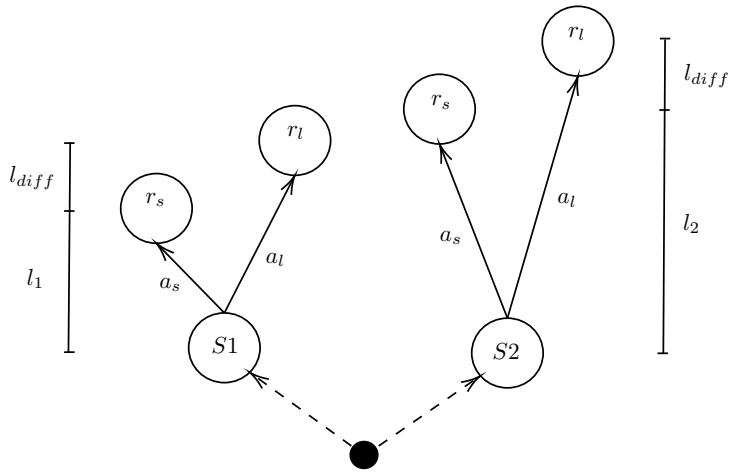

Figure 4.1: Environment inducing preference reversals. The agent starts non-deterministically in one of the states $S1$ or $S2$ and can choose between two actions: action a_s, leading to a smaller and more immediate reward, r_s; or action a_l, leading to a larger and more delayed reward, r_l. The rewards are distanced between them by a constant offset, l_{diff}. However, when the agent starts in state $S1$, the closest reward is available after l_1 steps while when the agent starts in state $S2$, the closest reward is available after l_2 steps.

4.1 Conceptual differences between discounting methods

As exposed in Chapter 3, hyperbolic discounting can be used as a tool to face uncertainty and hazardous environments, being a good alternative to conventional exponential discounting with a fixed discount factor. However, hyperbolic and exponential discounting present some significant conceptual differences that translate into different behaviors and computed policies. In this section, we explore the impact that the discounting method has on the learned policies and show that both discounting methods are complementary, capturing different sets of policies, neither of which fully encompassing the other.

The main difference between hyperbolic and exponential discounting lies in the shape of the discount function. On one side, the exponential discounting approach uses a constant discount factor that is applied uniformly to all future rewards and, on the other side, the hyperbolic discounting approach uses a non-constant discount factor that is applied differently to each future reward, greater for rewards that are closer in time and smaller for rewards that are further apart. This difference leads to a set of implications, including the aforementioned preference reversals.

To show the occurrence of the preference reversals and the consequent impact on the learned policies, we created a simple toy environment that allow us to analyze the behavior of both hyperbolically and exponentially discounted agents in a scenario that potentiates the reversal of the agent's preferences. As represented in Figure 4.1, the environment consists of a simple choice between two actions, a_s and a_l, each one leading to a different reward, one smaller and more immediate, r_s, and the other larger and more delayed, r_l. To test the occurrence of preference reversals, the agent is exposed to two possible scenarios, represented by the starting states $S1$ and $S2$, where the agent is non-deterministically placed

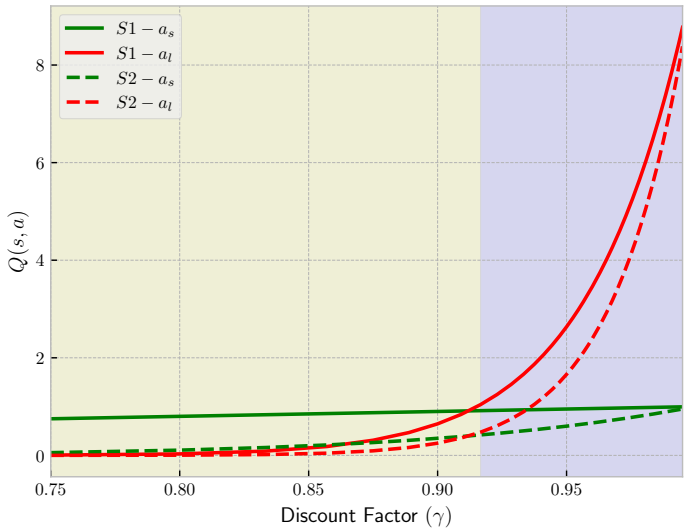

Figure 4.2: Q-values computed by exponentially discounted agents with different discount factors. The results show the Q-values computed by the agents corresponding to the states $S1$ and $S2$, given the two possible actions, a_s and a_l. The colored areas (blue and yellow) indicate ranges of discount factors for which the agent learns the same policy, i.e., all the discount factors within the same area lead to the same action selection.

at the beginning of each episode. In each state, the agent can choose between the two actions, a_s and a_l, leading to the rewards, r_s and r_l, respectively. Nonetheless, the rewards are available after different numbers of steps, depending on the initial state. In state $S1$, the agent can obtain the closest reward after l_1 steps, while in state $S2$, the closest reward is available after l_2 steps. The offset between the rewards, l_{diff}, is kept constant, allowing us to analyze the impact of the delay on the rewards in the discounting process and, consequently, in the learned policies.

Even though the environment can potentiate the occurrence of preference reversals, the rewards and delays need to be properly defined to induce such behavior. To achieve the desired experimental setup, the values assigned to the rewards and delays must satisfy

$$\frac{r_s}{1+l_1} > \frac{r_l}{1+l_1+l_{diff}} \quad and \quad \frac{r_s}{1+l_2} < \frac{r_l}{1+l_2+l_{diff}}, \tag{4.1}$$

where the first inequality represents the preference for the smaller reward when the agent starts in $S1$, where both rewards are closer, and the second inequality represents the preference for the larger reward when the agent starts in $S2$, where both rewards are further apart. The values for the rewards and delays that we use in the next experiments, satisfy the inequalities and are $r_s = 1$, $r_l = 10$, $l_{diff} = 25$, $l_1 = 1$ and $l_2 = 10$.

In order to show that agents using the standard exponential discounting and hyperbolic discounting learn different and complementary policies in the toy environment, in Figure 4.2, we start by analyzing

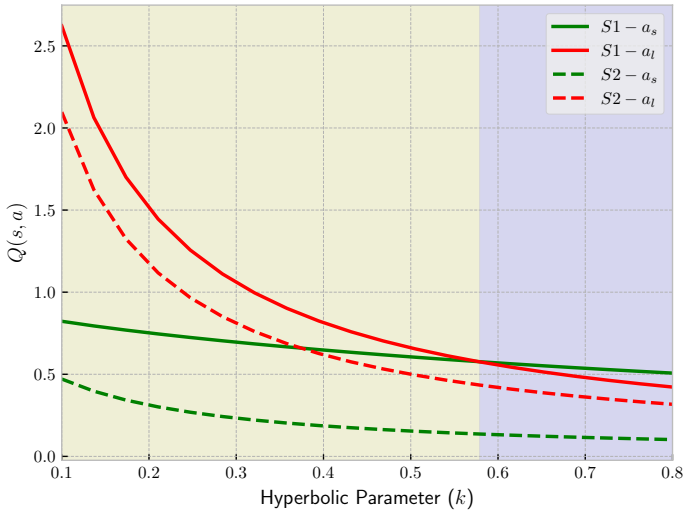

Figure 4.3: Q-values computed by hyperbolically discounted agents with different hyperbolic parameters. The results show the Q-values computed by the agents corresponding to the states $S1$ and $S2$, given the two possible actions, a_s and a_l. The colored areas (blue and yellow) indicate ranges of hyperbolic parameters for which the agent learns the same policy, i.e., all the hyperbolic parameters within the same area lead to the same action selection.

exponentially discounted agents. The results refer to the Q-values computed by agents using different discount factors. We can observe that the computed Q-values corresponding to the state $S1$, represented by solid lines, are higher for the smaller and more immediate reward, r_s, when the discount factor is lower, and the opposite occurs for the larger and more delayed reward, r_l. Referring to the Q-values corresponding to state $S2$, represented by dashed lines, we can observe a very similar behavior to the one observed in $S1$, being r_s preferred over r_l when the discount factor is lower and the opposite when the discount factor is higher.

Moreover, it is important to notice that, in both states, the preference for r_s occurs on the same range of discount factors, as well as the preference for r_l. This leads to the conclusion that, for the entire range of discount factors, there are two possible optimal policies, represented by the colored areas in yellow and blue. For the range of discount factors colored in yellow, the agent prefers r_s either at $S1$ or $S2$, while for the range of discount factors colored in blue, the agent prefers r_l in both states. All the conclusions drawn from the results obtained in this analysis are a direct consequence of the fact that the exponential discounting approach applies a constant discount factor to all future rewards, not considering the delay needed to obtain them, and consequently, not being able to represent the preference reversals.

We now analyze the impact of the hyperbolic parameter in the policies computed by hyperbolically discounted agents. In the results shown in Figure 4.3, we can observe the Q-values computed by agents using different hyperbolic parameters for both states, $S1$ and $S2$.

We conclude that for the range of the hyperbolic parameters, k, colored in yellow, the agent prefers

to select r_l in both states, assigning higher Q-values to the corresponding state-action pair. However, the policy computed for the range of hyperbolic parameters colored in blue results in a preference for r_s in $S1$ and r_l in $S2$. This behavior shows a preference for a different action in each one of the possible states, showing that the hyperbolically discounted agents are able to capture the preference reversals and policies that are not possible to be computed by the exponentially discounted agents.

In summary, the exponentially discounted agents, depending on the discount factor, are only able to encode two different policies, preferring r_s in both states or preferring r_l in both states as well. On the other hand, the hyperbolically discounted agents, depending on the hyperbolic parameter, can also encode two policies, preferring r_l in both states or preferring r_s in $S1$ and r_l in $S2$. Therefore, we can observe that the set of policies that can be encoded by each discounting approach is not contained in the set of optimal policies that can be encoded by the other approach, revealing that none of the approaches can be considered more expressive than the other. That demonstrates that both discounting methods are complementary and the choice between them should depend on the specificities of the problem and the desired behavior.

4.2 Discounting methods and uncertainty

Even though the results obtained in the previous section show that exponential and hyperbolic discounting can be used to serve different purposes, as discussed in Chapter 3, several studies have shown that hyperbolic discounting can be a good alternative to face uncertainty and hazardous environments. In this section, we investigate the implications of these differences in hazardous environments with several levels of uncertainty, from fixed to randomly sampled hazard rates.

4.2.1 Experimental environment

To properly understand the limitations of the conventional discounting methods in environments with uncertain hazard rates, we modeled as a hazardous MDP an adaptation of the *Pathworld* environment proposed by Fedus et. al [2]. This environment consists of a set of paths from which agents can choose to follow, not being able to change paths after one has already been selected. Each one of the paths has a different unique length but is also associated with a different reward that can only be obtained at the end of the path. In addition, during the interaction with the environment, the agent is also subject to an unknown hazard rate that determines the probability of the agent's death at each timestep. All these constraints make agents face a nontrivial decision at the beginning of each episode since longer paths are associated with higher rewards but also with a higher probability of death, which depends on the unknown hazard rate.

To properly define the environment, there is the need to define a set of parameters:

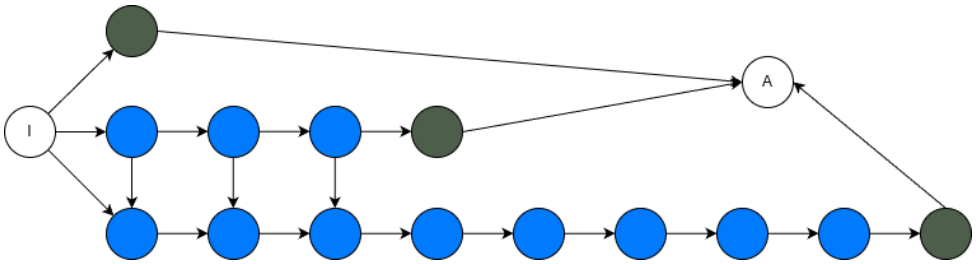

Figure 4.4: Dynamics of the adapted *Pathworld* environment featuring 3 paths (p_1, p_2 and p_3) with quadratic lengths (i.e., $l(p_i) = i^2$). The agents start their interaction with the environment at the initial state (marked with I), from which they can choose to follow any of the paths. At the intermediate states (in blue) agents can opt to continue following the current path or switch to another path with a larger length. Agents can only obtain rewards when achieving the end of the paths (in green), staying there until transitioning to the absorbing state (marked with an A). Since agents are subject to a hazard rate in each step, they can transition from any state to the absorbing state (marked with a A). For the sake of simplicity, these transitions are not represented in the figure.

- P - set of paths composing the environment, such that $P = \{p_1, p_2, ..., p_n\}$;
- $l(p_i)$ - length of the path p_i;
- $r(p_i)$ - reward associated with the path p_i;
- λ - hazard rate of the environment, such that the probability of the agent's death at each timestep is $1 - e^{-\lambda}$.

Since in the original *Pathworld* environment the agent is only allowed to choose paths at the beginning of each episode, we modify the original setting to allow the agent to switch paths at any timestep. This slight modification gives more flexibility to the agent and can increase the complexity of the environment, increasing also the number of strategies that the agent can use to maximize its reward.

In this modified version of the *Pathworld* environment, the state space refers to the position of the agent in the environment, as well as an initial state and an absorbing state that represents the end of the episode. Regarding the action space, for the sake of simplicity, the agent can choose to follow any of the paths at any timestep. However, we limit the agent only to be able to switch to paths with a higher length than the current one. The dynamics of the environment are illustrated in Figure 4.4.

4.2.2 Experimental results

In the following section, we present the results obtained in the experiments performed in the adaptation *Pathworld* environment, assessing the behavior of agents using hyperbolic and exponential discounting in hazardous environments with different levels of uncertainty.

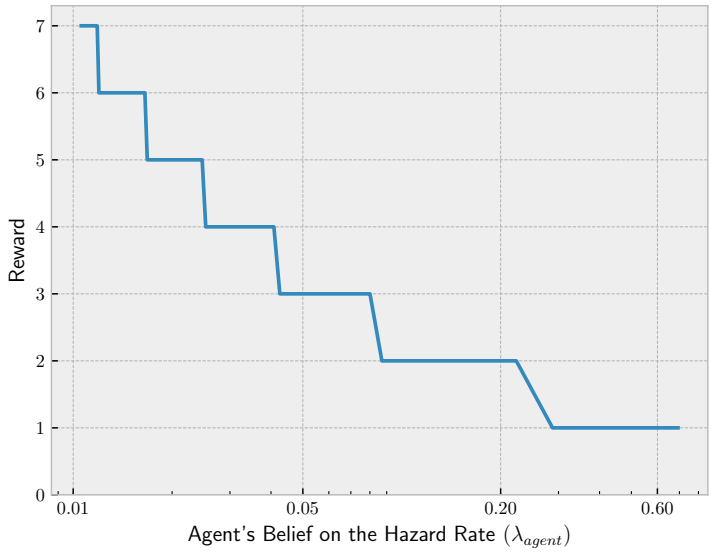

Figure 4.5: Reward obtained by the agent in a non-hazardous version of the *Pathworld* environment given different beliefs, λ_{agent}, about the environment's hazard rate.

No hazard rate

Before diving into the experiments involving uncertainty and agents featuring hyperbolic discounting it is important to understand the behavior of the exponentially discounted agents in a non-hazardous version of the *Pathworld* environment, modeled as a hazardous MDP with a hazard rate equal to zero.

In this scenario, the agent is deployed in an environment where it is not subject to any hazard rate. Therefore, the most rewarding plan to follow consists of choosing the path with the highest reward, since it is guaranteed that the agent will reach the end of the path. As the problem is modeled as an hazardous MDP, the problem definition does not provide a predefined discount factor, γ, that can be applied by the agent in the planning process. Therefore, since exponentially discounted agents require a discount factor to be defined, the agent's behavior is directly affected by the agent's belief about the environment's hazard rate, λ_{agent}, which can be directly mapped to the discount factor, γ, by the relation $\gamma = e^{-\lambda_{agent}}$. Given that, it is possible to study the effect an inaccurate belief about the environment's hazard rate has on the agent's performance in an environment without hazards. From now on, the experiments will be performed in versions of the *Pathworld* environment composed of 7 paths with lengths that grow quadratically and rewards that grow linearly with the index of the path i.e., $l(p_i) = i^2$ and $r(p_i) = i$.

In Figure 4.5, we display the results obtained by agents using the VI algorithm in a non-hazardous version of the *Pathworld* environment given different beliefs about the environment's hazard rate, λ_{agent}. We assume that the agent is not aware of the fact that the environment is non-hazardous. As can be observed, as expected, the agent's reward decreases as we consider higher beliefs about the envi-

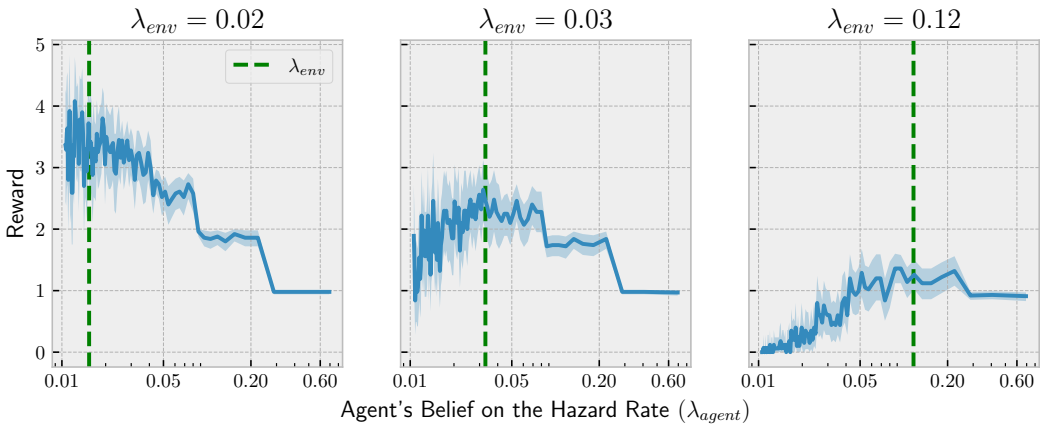

Figure 4.6: Average reward obtained by agents using an exponentially discounted VI algorithm given different beliefs about the environment's hazard rate in a hazardous version of the *Pathworld* environment. Each one of the plots refers to a fixed unknown hazard rate hazard rate, λ_{env}. The results are averaged over 100 episodes and the shaded area refers to a 99% confidence interval.

ronment's hazard rate, since higher beliefs imply lower expected time horizons and, consequently, the option for shorter paths. According to that, it is possible to observe that the agent is only capable obtain the maximum reward of 7 when the belief about the environment's hazard rate is low enough to consider all the paths in the environment (i.e. when $\lambda_{agent} \lesssim 0.0119 \Rightarrow \gamma \gtrsim 0.988$). This behavior shows that, in the context of hazardous MDPs, applying an arbitrary discount factor in the learning algorithm can lead to suboptimal behavior, being a crucial factor in the definition of the agent's behavior.

The policies obtained by these agents are precomputed and depend only on the agent's prior belief about the environment's hazard rate. Although agents with different beliefs can opt to reach the end of different paths, all of them stick to the same path from the moment they start the episode until the end. This kind of behavior is expected since taking an extra step switching paths increases the risk of not reaching the end of any of them.

Fixed hazard rate

Regarding the scenarios that can potentially benefit from the use of hyperbolic discounting, we start by analyzing the behavior of conventionally discounted agents in a hazardous version of the *Pathworld* environment in which the hazard rate, although unknown to the agent, stays constant throughout the episodes. To obtain a better understanding of the agent's behavior, in this version of the environment, we adopt a similar approach to the one used in the non-hazardous setting, running the VI algorithm with different beliefs about the environment's hazard rate, λ_{agent}. This approach allows us to analyze the impact of a mismatch between the agent's belief about the environment's hazard rate, λ_{agent}, and the actual hazard rate, λ_{env}, in the agent's performance. The obtained results are represented in Figure 4.6.

By analyzing the performance of the agents in each scenario, we conclude that a mismatch between λ_{agent} and λ_{env} can lead to a significant decrease in the agent's performance. In all of the test cases, including environments with hazard rates of 0.02, 0.03 and 0.12, the agents reached roughly the greatest average reward when the belief about the environment's hazard rate matched the actual hazard rate. Moreover, for the studied range of values for λ_{agent}, we can observe that the average reward grows until λ_{agent} reaches λ_{env}, and then, it rapidly starts to decrease. This kind of behavior is expected since, intuitively, when λ_{agent} is lower than λ_{env}, the agent tends to adopt a more conservative behavior that leads to the choice of shorter paths that are safer but also too conservative to maximize the collected rewards. On the other hand, when λ_{agent} is higher than λ_{env}, the agent tends to adopt a riskier behavior that leads to the choice of longer paths, that are more rewarding, but end up being too risky to maximize the agent's reward, since the agent is more likely to die before reaching the end of the path.

Randomly sampled hazard rate from a known distribution

Moving to the analysis of the scenarios where the agents are subject to an uncertain hazard rate that can change between episodes, we start by analyzing the behavior of the agents in a hazardous version of the *Pathworld* environment where the hazard rate is drawn from a known distribution. To this end, we introduce the hyperbolically discounted agents developed by Fedus et. al [2] and compare their performance with the performance of the exponentially discounted agents.

To properly compare the performance of both discounting approaches in such scenario, we adapt the environment to sample, at the beginning of each episode, the hazard rate from an exponential distribution such that $\lambda_{env} \sim \text{Exp}(k_{env})$, where k_{env} is the expected value of the distribution.

The plots in Figure 4.7 refer each to a different parameter of the hazard rate distribution, k_{env}, and show the average reward obtained by exponentially discounted agents using the VI algorithm computed over a range of beliefs about the environment's hazard rate, λ_{agent}. In each plot, it is also shown the average reward obtained by hyperbolically discounted agents computed assuming a correct prior about the random distribution, such that $k_{agent} = k_{env}$. The hyperbolically discounted agents were implemented using the same approach proposed by Fedus et. al [2] where k_{agent} was directly used as the hyperbolic parameter k in the approximation of the hyperbolically discounted Q-values as described in Section 3.3 of this document.

Observing the results from Figure 4.7, we conclude that, for all values of k_{env}, the hyperbolically discounted agents built given a hyperbolic parameter equal to k_{env} were able to match results of the best performant exponentially discounted agent over the tested range of λ_{agent}. However, the obtained results also show that the option for a hyperbolically discounted agent does not imply a significant gain of performance when compared to the option for the best exponentially discounted agent. To understand the causes of this finding we must consider that in the case of exponentially discounted agents, changing

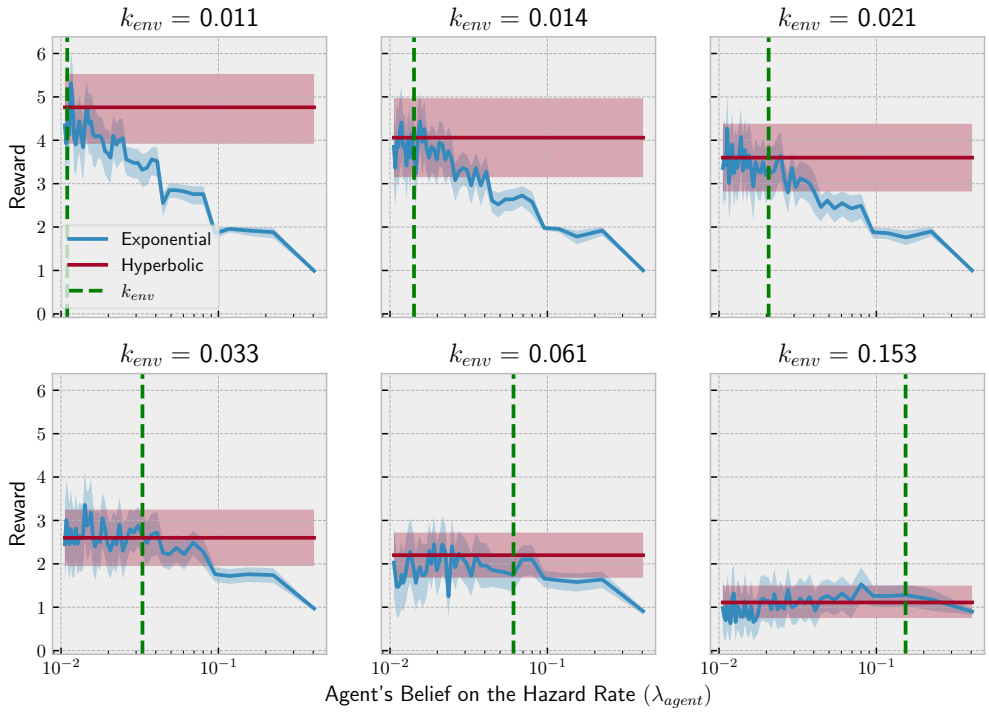

Figure 4.7: Average reward obtained by exponentially discounted agents using the VI algorithm over a range of priors about the environment's hazard rate, λ_{agent}, and hyperbolically discounted agents, as proposed by Fedus et. al [2], in a hazardous version of the *Pathworld* environment. The environment features an uncertain and episode-varying hazard rate drawn from an exponential distribution with an expected value, k_{env}. It is assumed that the hyperbolically discounted agents know k_{env}. All results are averaged over 100 episodes and the shaded area refers to a 99% confidence interval.

the belief about the environment's hazard rate, λ_{agent}, or, equivalently, the discount factor, γ, leads to a change in the agent's policy. In the specific case of the *Pathworld* environment, the optimal behavior consists of choosing a path at the beginning of the episode and sticking to it until the end. Given that, the hyperbolically discounted agents can, at most, match the performance of the best performant exponentially discounted agent, computing a policy that matches the one that sticks to the optimal path for the given hazard rate. Nevertheless, even though selecting the best exponentially discounted agent might be a good strategy, in more complex scenarios, it is not feasible to pre-compute policies for a high, possibly infinite, number of beliefs about the environment's hazard rate and select the best one. Adding to that, given an environment's hazard rate distribution with a known parameter, k_{env}, it is not trivial to analytically determine the best exponentially discounted agent, making the hyperbolically discounted agents a better option to deal with stochasticity in the environment's hazard rate.

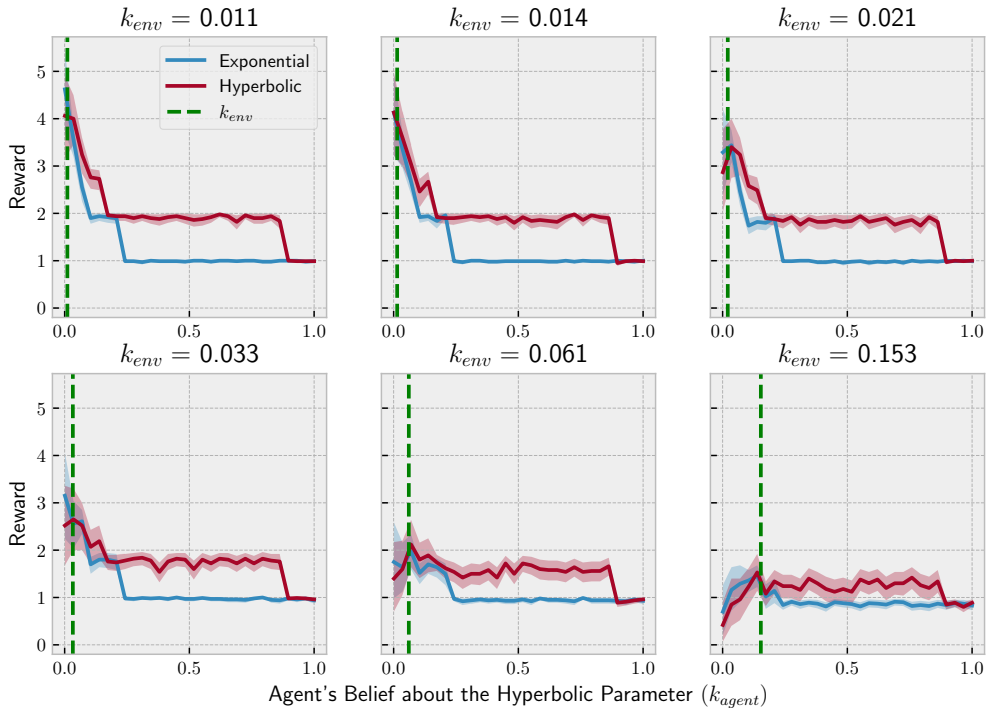

Figure 4.8: Average reward obtained by agents using an exponentially discounted VI algorithm and hyperbolically discounted agents in an environment with an uncertain and episode varying hazard rate drawn from an exponential distribution with an expected value, k_{env}, unknown to the agents. For each value of k_{agent}, the corresponding exponentially discounted agent applies a discount factor $\gamma = e^{-\lambda_{agent}}$ such that $\lambda_{agent} = k_{agent}$ and the hyperbolically discounted agent uses the hyperbolic parameter $k = k_{agent}$. The results are averaged over 100 episodes and the shaded area refers to a 99% confidence interval.

Randomly sampled hazard rate from an unknown distribution

We now move to the scenario where the agents are subject to an uncertain hazard rate drawn from a distribution with an unknown parameter k_{env}.

We analyze the agents' behavior in a similar approach to the one used in the previous scenario, comparing the performance of hyperbolically discounted agents with different hyperbolic parameters, k_{agent}, and exponentially discounted agents using different beliefs about the environment's hazard rate, λ_{agent}. To properly compare the two discounting approaches, we needed to connect the hyperbolic discounting parameter, k_{agent}, to a belief about the environment's hazard rate, λ_{agent}, which could be used by an exponentially discounting agent. A hyperbolic agent, characterized by the parameter k_{agent}, behaves as if it assumes a distribution of hazard rates with an expected value equal to k_{agent}. Therefore, we compare each hyperbolic agent with hyperbolic parameter, k_{agent}, to an exponentially discounting agent that uses a belief about the environment's hazard rate, λ_{agent}, equal to k_{agent}.

The plots in Figure 4.8 refer each to a different parameter of the hazard rate distribution, k_{env}, and show the average reward obtained by hyperbolically discounted agents using different priors, k_{agent},

about k_{env}, as well as the average reward obtained by exponentially discounted agents using different beliefs about the environment's hazard rate, λ_{agent}, each one directly mapped from the hyperbolic parameter, k_{agent}, from the corresponding hyperbolically discounted agent.

By observing the results, we can conclude that when the hyperbolic parameter is not known, the hyperbolically discounted agents proved to be more robust to the uncertainty in k_{env} than the exponentially discounted agents since when there is a mismatch between k_{agent} and k_{env}, the hyperbolically discounted agents are, in the vast majority of the cases, able to outperform or match the performance of the exponentially discounted agents. This behavior shows that the hyperbolic discounting approach is less sensitive to the different values for k_{agent} and better adapts to uncertainty.

4.3 Takeaways

In this chapter, we explored the advantages of applying hyperbolic discounting under hazardous environments and performed systematic comparisons between hyperbolic and widely used exponential discounting techniques.

As discussed in Section 4.1, exponential and hyperbolic discounting model distinct behaviors. Therefore, both discounting techniques are complementary, as they capture different sets of policies.

Additionally, in Section 4.2 we compare the performance of hyperbolically and exponentially discounted agents in hazardous environments with varying levels of uncertainty, from fixed to randomly sampled stochastic hazard rates. From the results, we conclude that hyperbolically discounted agents are more robust to uncertainty in the environment's hazard rate, especially when the environment features a stochastic hazard rate, drawn from either a known or unknown exponential distribution.

5

Adaptive Discounting Framework

Contents

Differently from the scenarios discussed in the previous chapter, where the agent faces stochastic hazard rates sampled from random distributions, in this chapter, we study the case where agents face unknown deterministic hazard rates. More specifically, we study agents that can deal with either stationary or non-stationary hazard rates. To address this issue, we propose a mechanism that allows the online adaptation of the discount factor in planning algorithms to improve the agent's performance in hazardous environments. Along the chapter, we present the main aspects of the proposed solution and describe the development process of the proposed solution as well as the implementation details.

5.1 General Overview

As previously expressed, our proposed solution is based on the development of an agent that incorporates an adaptive discounting framework that allows the agent to tune its estimate over time. We start

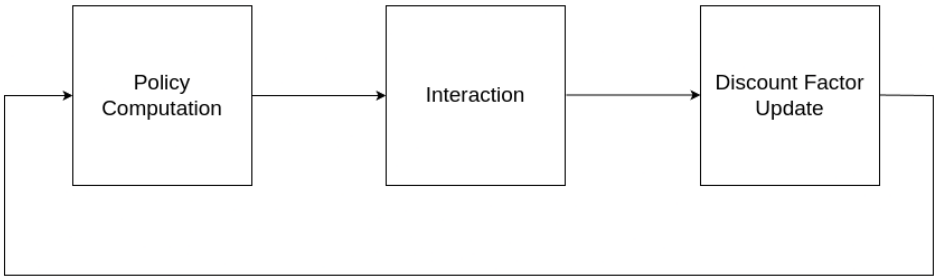

Figure 5.1: General overview of the proposed solution.

by giving a general overview of the proposed solution, presenting the main aspects of the mechanism and how it can be applied to planning algorithms, such as VI.

Our solution consists of an agent that is able to adapt its discount factor over time by interacting with the environment and performing incremental updates to the estimate between each episode of interaction. This inter-episodic update process is based on the information gathered by the agent by the end of each episode, more specifically, the duration of the episode. The episode's duration is used to estimate the most appropriate discount factor to be considered in the next episode. A general overview of the proposed solution is depicted in Figure 5.1.

The mechanism that serves as the core of the proposed agent can be split into three main phases. The first phase is the policy computation phase, where the agent selects, from a set of precomputed policies, the most appropriate policy to be used in the next interaction with the environment. It is important to notice that the actual computation of the policies occurs only once, before the interaction with the environment, allowing the agent to simply choose between the precomputed policies after that. After selecting a policy, the agent starts the interaction with the environment, following the selected policy during one full episode. By the end of the episode, the agent collects information about the episode's duration and uses this information as input to the remaining module of our solution. This module is responsible for the discount factor adaptation, where the agent updates its estimate of the discount factor based on the information gathered during the episode feeding this new estimate to the module capable of selecting the most appropriate policy for the next episode. This process is repeated for several episodes, allowing the agent to adapt its discount factor estimate over time and improve its performance incrementally. A detailed explanation of the policy computation and discount factor adaptation modules is given in the following section.

5.2 Implementation

In this section, we present the main aspects of the implementation of the proposed solution, from the conceptual design of the agent to the implementation details of each one of its modules.

5.2.1 Policy Computation

An agent capable of adapting its discount factor over time must be equipped with a mechanism to deal with the several changes in the discount factor estimate, computing the most appropriate policy for each estimated value. The process of computing the most appropriate policy for a given estimate of the discount factor can be done in two different ways.

On one side, the agent can compute the policy using the current estimate of the discount factor in an online way and use the freshly computed policy in the next moment of interaction. However, the online computation of the policy can be computationally expensive, especially when the agent is using complex algorithms that require long training times and computational resources. Moreover, a small change in the discount factor can have no consequences on the optimal policy computed by the agent, especially when the perturbation in the discount factor is small. This phenomenon can be observed in the studies performed in the adaptation of the *Pathworld* environment, where there are only a few optimal policies that can be computed by the agent over the whole range of discount factors. In this particular case, a small change in the estimate of the discount factor, performed by the end of an episode of interaction, can lead to the computation of the same optimal policy used in the previous episode, wasting computational resources and time for no gain of performance.

On the other side, assuming knowledge of the environment's transition and reward functions, the agent can compute the optimal policy for a range of discount factors before the interaction with the environment, storing the computed policies in such a way that the agent can select the most appropriate policy for the current estimate of the discount factor. This approach can be considered less accurate since the agent is not computing the optimal policy for the exact estimate of the discount factor, being restricted to the precomputed set of policies. However, this restriction translates into a gain of efficiency at test time, since the agent can simply select the most appropriate policy from the set of precomputed policies, without the need to compute the optimal policy online. Adding to that, choosing a good set of discount factors for which the agent will compute the optimal policy can mitigate the loss of accuracy in the process of selecting the most appropriate policy to use in each episode.

Posing the above considerations, to the end of this work, we follow the second approach, precomputing the optimal policy for a set of discount factors before the interaction with the environment. To do so, we applied the VI algorithm [9] to compute the needed policies and store them to be used in the selection process occurring between the episodes of interaction, every time the agent updates its estimate of the discount factor.

As a consequence of the precomputation of the optimal policies for a predefined set of discount factors, the selection of an appropriate set of discount factors poses a challenge in the implementation of the agent. This process is a crucial step in the implementation of the agent since it can directly affect the agent's performance.

In terms of the density of the set of discount factors, if the set is too sparse, the agent can be stuck in the same policy for several episodes. On the other hand, if the set is too dense, the agent can waste computational resources computing policies that are very similar to each other, leading to a waste of computational resources and training time.

In addition to the overall density of the set of discount factors (i.e., the total number of discount factors within the set), the spacing between consecutive discount factors is also a critical factor in the selection process. As discussed in Chapter 3, there is a direct relationship between the discount factor and the expected environment's hazard rate. Moreover, the discount factor can be seen as a parameter that imposes a time horizon to the agent, from which the agent stops to consider future rewards. The relation between the discount factor and its expected time horizon (3.13), studied by Sherstan et al. [4], was shown to be nonlinear, with the expected time horizon growing slowly as the discount factor stays low and growing sharply as the discount factor approaches 1. In this sense, choosing a simple linear spacing between the discount factors can lead to a set of discount factors that are not well distributed in terms of the expected time horizon, favoring the selection of policies with lower time horizons and penalizing the selection of policies with higher time horizons.

In order to mitigate the above-mentioned issues, our selection process is fundamentally grounded in the concept of the expected time horizon and the non-linear relation between the time horizons and the corresponding discount factors. Our approach is based on the idea of selecting a meaningful range of time horizons and computing the corresponding discount factors that are directly connected to the expected time horizon.

To concretize this idea, we choose a set of consecutive time horizons that vary from 2 to a maximum value, τ_{max}. The set of time horizons starts at 2 since considering a time horizon of 1, as shown in (3.13), would lead to a discount factor of 0, which disables the agent from considering future rewards. The maximum value of the set, τ_{max}, is a hyperparameter of the agent and can be set according to the complexity of the running environments or the specifications of the underlying task. More explicitly, in episodic tasks, this parameter can be adapted according to the length of the largest trajectories that the agent can perform during an episode. Regarding continuous tasks, where the agent can interact with the environment for a potentially infinite amount of time, the parameter can be set to a large value, imposing a large maximum time horizon on the agent, and allowing the agent to consider future rewards in the long run. In these specific cases of continuous tasks, the parameter must be set taking into account that, for very large values of τ_{max} (corresponding to discount factors close to 1), common RL and planning algorithms can have convergence issues. Adding to that, for large environments with expected large trajectories requiring a large maximum time horizon, computing one policy for each time horizon in the range $\{2, ..., \tau_{max}\}$ can be computationally expensive and most of the time unnecessary since close time horizons can lead to similar policies. To mitigate this issue, we added a spacing factor, δ_τ, that allows

the agent to compute policies for a subset of the time horizons in the range $\{2, ..., \tau_{max}\}$, uniformly spaced by δ_τ. After receiving both τ_{max} and δ_τ as inputs, our agent computes the set of policies for the selected time horizons using the VI algorithm, and stores them to be used in the selection process that occurs between the episodes of interaction. A pseudocode of the policy computation process is given in Algorithm A.1 from Appendix A.

The process that completes the module of policy computation lies in the translation of the current estimate of the discount factor to the policy that should be used in the next interaction phase. Since the estimate of the discount factor can vary continuously and the set of precomputed policies is restricted to a discrete set of discount factors, the agent must link its estimate to the most appropriate policy. In order to complete this step, our agent uses a simple heuristic that selects the policy computed for the discount factor that is the closest to the current estimate of the discount factor. This heuristic, although simple, can be really effective when combined with a good set of precomputed policies. A pseudocode explaining the heuristic and its integration in the whole module of policy computation is given in Algorithm A.2 from Appendix A.

5.2.2 Discount Factor Adaptation

The process of computing and selecting the most appropriate policy for a given estimate of the discount factor is the first step in the implementation of the proposed agent. Therefore, the core of our solution lies in the discount factor adaptation module. This module is responsible for updating the agent's estimate of the discount factor based on the information gathered during the interaction, allowing the agent to adapt its estimate based on the belief about the environment's hazard rate. In this section, we present the main aspects of the discount factor adaptation process starting with an adaptation method that allows the estimation to be effective in an environment with a fixed and unknown hazard rate and then extending the method to deal with environments where the hazard rate can suffer perturbations over time.

Gathering Information

Every approach to the development of an agent that can adapt its discount factor over time should be grounded in the information gathered by the agent during the interaction with the environment. In this sense, the first step in the discount factor adaptation process is the collection of the information that will be used to update the agent's estimate of the discount factor. This information can be any kind of data that can be used to infer the interaction between the agent and the environment, such as the duration of the episode, the number of steps taken by the agent, and the rewards collected during the episode, among others. In this work, since we are directly interested in the application of the adaptive mechanism in scenarios involving risk and uncertainty, we take the duration of the episode as the key information to be used in the discount factor adaptation process.

Since our agent is specifically designed to operate in hazardous environments, it is intended for use in the hazardous variant of the standard MDP model, known as hazardous MDPs. As previously outlined, in this model, the agent, by being subject to a hazard rate, λ, can die at each time step. The agent's death signals the end of the episode and can be practically modeled by introducing an absorbing state to which the agent is sent after dying. For our adaptation module, the duration of the episode corresponds to the number of steps taken before the agent dies. This information is used as input to the discount factor adaptation module, allowing the agent to update the estimate of the discount factor based on the observed episode duration. Formally, given an episode corresponding to a sequence

$$\{s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{T-1}, a_{T-1}, r_{T-1}, s_T\},$$

where s_T is the absorbing state, the agent stores the duration of the episode, T, and uses this information to update the estimate of the discount factor for the next episode.

Adapting to a Fixed Hazard Rate

Diving into the details of the core of the discount factor adaptation module, we start by presenting a method that allows the agent to adapt its estimate of the discount factor in an environment with a fixed, albeit unknown, hazard rate.

To ensure the adaptation mechanism is effective in environments with a fixed hazard rate and capable of performing the adaptation process online, with successive and incremental updates to the agent's estimates based on observed episode durations, we adopt a Bayesian approach to model the agent's belief about the environment's hazard rate. As discussed in Chapter 3, in order to apply the Bayes' theorem and obtain updated estimates of the hazard rate, we need to define two fundamental concepts: the prior distribution and the likelihood function. While the prior distribution represents the agent's initial belief about the environment's hazard rate, the likelihood function describes the probability of observing a given episode duration given a specific hazard rate.

We begin by selecting the likelihood function and choose to model the episode duration as a random variable following a geometric distribution. As previously detailed in Chapter 2, the geometric distribution is a discrete probability distribution that models the number of trials required to achieve the first success in a sequence of Bernoulli trials. Linking this definition to the hazard rate estimation problem, in the context of the agent's interaction with the environment, the episode duration is viewed as the number of trials before the agent's death occurs. Given this, the geometric distribution defined in (2.8) as

$$P(X = k) = (1 - p)^{k-1} \cdot p, \quad k = 1, 2, 3, ...,$$

is suitable for modeling the episode duration, where p corresponds to $1 - e^{-\lambda}$, the unknown probability

of death at each time step, and X is the random variable representing the episode duration. Following the selection of the geometric distribution as the likelihood function, we adopted the Beta distribution, defined in (2.9), as the prior distribution, given that it is a conjugate prior for the geometric distribution. Since this property makes the posterior distribution update a straightforward adjustment of the shape parameters, α and β, the update process, carried out after observing an episode duration T, with prior parameters α and β, is made following

$$\begin{cases} \alpha' = \alpha + 1 \\ \beta' = \beta + T \end{cases},$$

where α' and β' are the updated shape parameters of the Beta distribution, representing the agent's updated belief about the hazard rate. As shown in the update rule, the update process increments the parameter α by 1 with each new observation and updates β by adding the observed episode duration, T. This allows the agent to refine its estimate of the hazard rate based on the data collected during its interactions with the environment. The updated parameters, α' and β', are then used as the prior for the next update, enabling the agent's belief about the hazard rate to be continuously adjusted based on episode durations. As the number of observations increases, the posterior distribution becomes more concentrated around the true hazard rate.

The final step in the discount factor adaptation module is determining a discount factor that can be used to select the most suitable policy for the next episode. Since the posterior distribution follows a Beta distribution and models the probability of the agent "dying" at each time step, the expected value of this distribution can be used in the calculation of the discount factor. The properties of the Beta distribution allow its expected value to be easily computed from the parameters, α and β, as:

$$E[X] = \frac{\alpha}{\alpha + \beta}.$$

This value represents the expected probability of death at each time step and can be directly translated into the discount factor, γ, using the expression

$$\gamma = 1 - E[X] = 1 - \frac{\alpha}{\alpha + \beta}.$$

Thus, the discount factor, derived from the expected value of the posterior distribution, is used to select the most appropriate policy for the upcoming episode, as outlined in previous sections.

Variable Unknown Hazard Rate

Although the adaptation method discussed works well in environments with a fixed hazard rate, it faces significant adaptability challenges when the hazard rate is non-stationary. As the number of observations

increases, the posterior distribution becomes less responsive to new data, being overly influenced by past observations that may no longer accurately represent the current hazard rate.

To address this issue, we developed an extension to the fixed hazard rate adaptation method that allows the agent to easily adapt to changes in the environment's hazard rate. This extension, as outlined in Algorithm 5.1, introduces a scaling factor that divides the shape parameters of the posterior distribution by a constant value, $scalingFactor$, every time the number of observations reaches a multiple of this factor.

This process effectively reduces the influence of past observations on the posterior distribution, allowing the agent to adapt more quickly to changes in the hazard rate. By scaling down the shape parameters, the agent performs a "controlled reset" of its estimate that, even though retaining the same expected value, becomes more responsive to new data. As a result, the agent can adjust its estimate of the hazard rate more rapidly, ensuring that the posterior distribution remains responsive to new data and accurately reflects the current hazard rate at all times.

Algorithm 5.1: Adaptation to a variable hazard rate

Input: α, β: Prior distribution parameters, T: Episode duration, $scalingFactor$: Convergence threshold
Output: α', β': Posterior distribution parameters
begin
  $\alpha' \leftarrow \alpha + 1$
  $\beta' \leftarrow \beta + T$
  if $\alpha \mod scalingFactor = 0$ **then**
    $\alpha' \leftarrow \frac{\alpha'}{scalingFactor}$
    $\beta' \leftarrow \frac{\beta'}{scalingFactor}$
  return α', β'

6

Evaluation

Contents

In this chapter, we conduct a set of experiments to evaluate the proposed adaptation mechanism in both stationary and non-stationary hazard rate scenarios[1]. To assess the effectiveness of the proposed mechanism, we compare its performance against baseline strategies that operate under ideal conditions where the hazard rate is known beforehand.

6.1 Experimental Setup

To evaluate the proposed adaptation mechanism and observe its behavior under different conditions of risk and uncertainty, the experiments must be conducted in a simulated environment with a controlled hazard rate, introducing uncertainty in the system by exposing the agent to a probability of failure at

[1] The proposed adaptive discount factor framework and test environment were implemented using Python. Several libraries and frameworks were employed to build the agent and conduct the experiments, including *NumPy*, *Matplotlib*, and *Pygame*. The source code for the proposed solution and test environment can be accessed via the following GitHub repository: https://github.com/guilhermersalvador/adaptive-discount-factor.

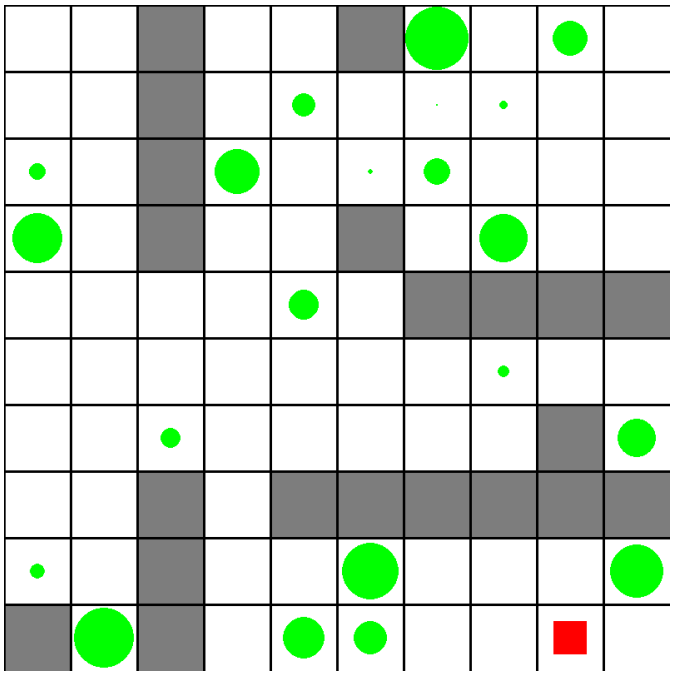

Figure 6.1: Custom-built grid environment. Visual representation of the agent, represented by a red square, in its start position and the several rewards available in the environment, represented by green circles, sized according to the reward value (larger circles represent higher rewards). The agent can navigate through all white cells in the grid, being restricted to moving to the walls represented by gray cells.

each time step. Moreover, the environment must be reasonably complex and sensitive to changes in the agents' time horizons and risk preferences, allowing us to observe the effects of the proposed adaptation mechanism on the agent's performance. To provide useful insights into the behavior of the proposed adaptation mechanism, we built a custom simulation environment with higher complexity than the adaptation of the *Pathworld* environment used in the experiments discussed in Chapter 4.

Therefore, we create a custom-built simulation environment that consists of a hazardous grid world where the agent must navigate through the world collecting multiple rewards, randomly placed at different distances from the starting point. The agent can collect multiple rewards in a single episode and must decide how far it is willing to go to collect the rewards and what the best path to take. The agent should consider that the risk of failure at each step can lead to a non-materialization of the rewards present in the planned path. A visual representation of the environment is shown in Figure 6.1 where it is represented by both the agent in its start position and the several rewards available in the environment.

To increase the complexity and expressiveness of our simulation environment we also introduce some additional features that make the environment more challenging and unpredictable. These features include the inclusion of walls that restrict the agent's movement and the introduction of stochasticity in the agent's actions, where the agent can move in a direction different from the one it intended inducing perturbations in the planned paths. Moreover, as a hazardous environment, the agent is exposed to a

probability of failure at each time step, which can lead to the agent's death and the end of the episode.

In summary, the custom-built simulation environment exhibits the following features and dynamics:

- State Space — The environment is represented as an $N \times N$ grid, where the agent can navigate freely. It is composed of $N^2 + 1$ states, N^2 grid cells, and an additional absorbing state that represents the termination of the episode.
- Action Space — The agent has four possible movements — up, down, left, and right.
- Reward Function — Rewards are randomly distributed across the grid cells, at the world's generation, and agents can collect them by being in the same cell as the rewards. The value assigned to each reward is unique and ranges from 1 to the number total number of rewards in the environment.
- Transition Dynamics — The agent can move freely within the grid but is restricted by walls and grid boundaries. If an action would lead the agent into a wall or outside the grid, the agent remains in the same cell. Additionally, action outcomes are subject to stochasticity, where the chosen action may be replaced by a random action with a certain probability. The agent also faces a failure probability at each time step, resulting from the environment's hazard rate, which can cause the agent to transition to the absorbing state.
- Hazard Rate — The environment's hazard rate is set at the beginning of each episode and remains constant throughout the episode. While stationary during an episode, the hazard rate can vary between episodes, either following a fixed value or oscillating according to a predefined pattern.

6.2 Experimental Results

After defining the simulation environment and the experimental setup, we conduct a series of experiments to evaluate the proposed adaptation mechanism in both fixed and oscillating hazard rate scenarios. The experiments are designed to assess the performance of the proposed mechanism and compare it against baseline strategies that operate as an ”oracle”, having full knowledge of the hazard rate.

6.2.1 Adaptation to a Fixed Hazard Rate

We first evaluate the agent's adaptation to a deterministic and stationary hazard rate. In this scenario, the agent is exposed to a constant hazard rate throughout several episodes of interaction, allowing us to observe the agent's behavior under different conditions of risk and uncertainty. In the next set of experiments, we assess some evaluation metrics such as the convergence speed of the proposed adaptation mechanism to the optimal discount factor and the agent's performance under different environments'

Figure 6.2: Adaptation to a fixed hazard rate. The plot shows the variation of the agent's belief about the hazard rate, λ_{agent}, throughout 200 episodes of interaction with a hazardous environment with a fixed hazard rate, $\lambda_{env} = 0.05$.

hazard rates. We perform the next experiments, in the grid environment where the obtained rewards are highly dependent on the environment's hazard rate. Therefore agents' performance is highly sensitive to changes in the expected time horizon and risk preferences. The experiments are conducted in a grid environment with 10x10 cells, featuring a 0.1 probability of action stochasticity and approximately 30% of the cells containing rewards.

We start by validating the proposed adaptation mechanism by assessing the convergence speed of the framework to the exact hazard rate of the environment and the accuracy of the agent's estimation. We conduct a set of experiments to observe the agent's belief about the hazard rate throughout the episodes of interaction with the environment. The results are shown in Figure 6.2 and, as we observe, the proposed adaptation mechanism initializes its belief about the hazard rate with a reasonably high value, which leads to a conservative behavior in the initial episodes. However, as the agent interacts with the environment and collects more information about the duration of each episode, the agent's belief about the hazard rate converges to the optimal value, 0.05, after only a few episodes. It is observable that from roughly the 20^{th} episode, the agent's belief about the hazard rate stabilizes around the optimal value starting to orbit around it but with no significant deviations. This behavior shows that, even though the initial estimate of the hazard rate is far from the true value, the agent can quickly adapt its belief to the optimal value.

After validating the proposed adaptation mechanism and observing its convergence to the true environment's hazard rate, we assess the effectiveness of the proposed mechanism in terms of the total

Table 6.1: Discount factors range for policy computation.

γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	γ_{10}
0.5	0.917	0.955	0.969	0.976	0.981	0.984	0.986	0.988	0.989

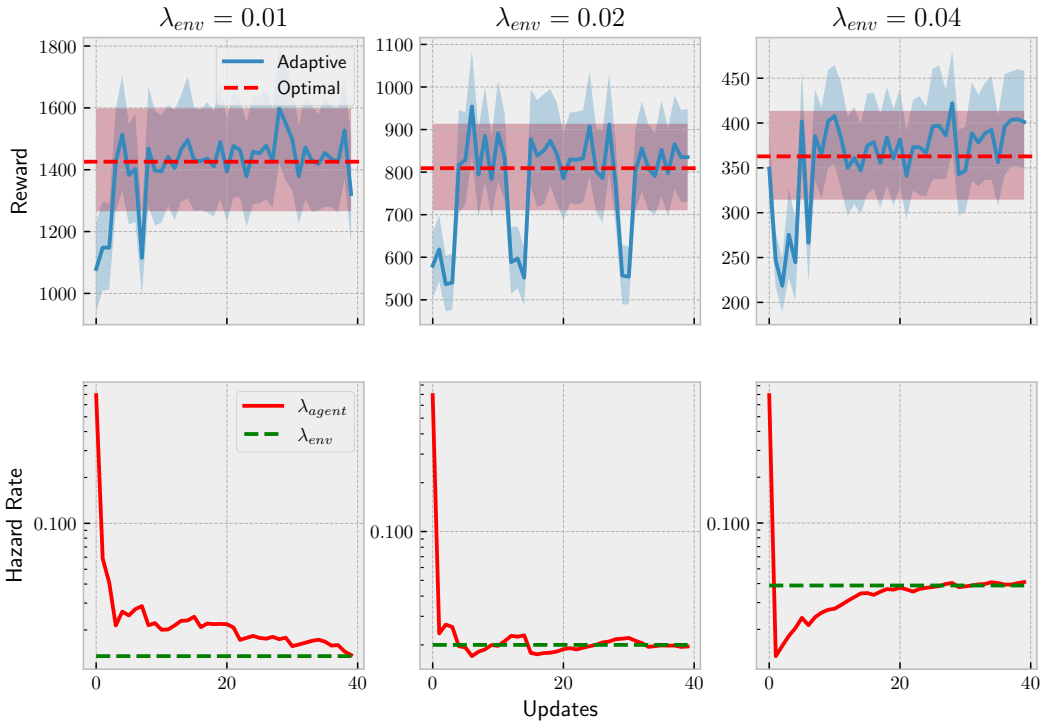

Figure 6.3: Comparison of agent's performance and belief about the hazard rate for different fixed hazard rates. The plot shows the total rewards collected by the agent between updates of its belief about the hazard rate, λ_{agent}, in different fixed hazard rates, λ_{env}. The results are averaged over 500 episodes and the shaded area corresponds to a 99% confidence interval.

rewards collected by the agent in the hazardous grid environment. As mentioned during the solution specification, we need to define a set of discount factors from which the agent precomputes the base policies. Following the methodology presented in Chapter 5, we define the set of discount factors corresponding to time horizons until 100 - the length of the largest trajectory that the agent can take in the environment, visiting all cells. To reduce the cardinality of the set of discount factors, we define a step of 10 between each time horizon, resulting in the set of 10 discount factors listed in Table 6.1.

By running the agent in the hazardous grid environment with the fixed hazard rate, we can analyze the agent's performance in terms of the total rewards collected in each episode, as well as the influence of the oscillations of the agent's belief about the hazard rate in its performance. The results of the experiments conducted to evaluate the agent's performance in different fixed hazard rates are shown in Figure 6.3 and clearly demonstrate the impact of the oscillations of the agent's belief about the hazard rate in its performance. Even though the experiments refer to different fixed hazard rates, the agent's

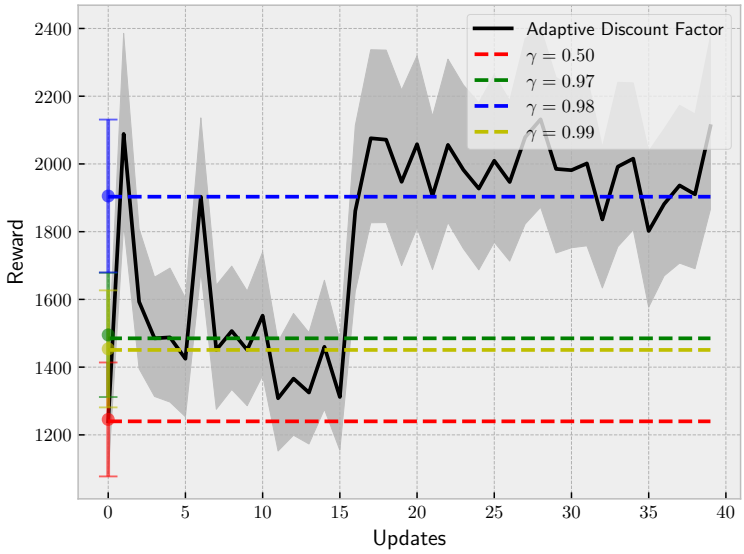

Figure 6.4: Performance comparison between an adaptive agent and agents applying policies that compose the set of precomputed policies of the adaptive agent in a fixed hazard rate scenario. The plot shows the average rewards collected by the adaptive discounting agent, over 40 estimate updates, and baseline agents that apply the policies that compose the set of precomputed policies of the adaptive agent. The agents interact with an environment with a fixed hazard rate, $\lambda_{env} = 0.04$, and the results are averaged over 500 episodes. The shaded area corresponds to a 99% confidence interval.

patterns of behavior are similar, with the only major difference in the magnitude of the total rewards collected, being lower as the hazard rate increases. For all the tested values, we observe that when the agent's belief about the hazard rate diverges from the true value the agent's performance suffers significant drops. However, as the agent's belief about the hazard rate converges to the true value, the agent's performance improves significantly and stabilizes, reaching its maximum performance after a few episodes as the agent's belief about the hazard rate stabilizes as well.

We also conduct a comparative analysis to evaluate the effectiveness of the proposed adaptation mechanism against the hypothetical ideal scenario where the agent has full knowledge of the hazard rate. To perform this comparison, we compare our adaptive agents against exponentially discounted agents that act as an "oracle" since they use the optimal discount factor (mapped from the true hazard rate) to compute the optimal policy. The obtained results, also represented in Figure 6.3, allow us to conclude that after the initial stabilization episodes, the adaptive agents are able to achieve similar performance to the "oracle" agents, showing that the proposed adaptation mechanism is not only efficient in finding the optimal discount factor but also in achieving similar performance to agents with full knowledge of the hazard rate.

Comparing our adaptive agents with 'oracle' agents provides valuable insights into the effectiveness of the proposed adaptation mechanism in hazardous environments. Nonetheless, exploring the benefits of the adaptation framework compared to the direct use of the precomputed policies that support our

Table 6.2: Mean squared error between the agent's estimation and the true hazard rate, given different values for the $scalingFactor$ hyperparameter. The results refer to the interaction of the agent with an oscillating hazard rate environment for 200 episodes.

Scaling Factor	Error
1	0.00652
5	0.00292
10	0.00239
20	0.00342
40	0.00559

adaptive agents, helps us to establish useful baselines for measuring the added value of the adaptation process. The results of this comparison are presented in Figure 6.4 and provide a direct comparison between the adaptive agent and four baseline agents, each applying policies selected from the adaptive agent's set of precomputed policies. For simplicity, only a subset of these precomputed policies was represented. As expected, after a few updates to the agent's belief about the hazard rate, the adaptive agent outperforms most baseline agents and matches the performance of the best one. This suggests that the proposed adaptation mechanism not only accurately estimates the hazard rate but also selects the most suitable policy for each belief, making correct use of the available policies.

6.2.2 Adaptation to an Oscillating Hazard Rate

After evaluating and validating the proposed adaptation mechanism in a fixed hazard rate scenario, we assess the extension of the mechanism to a still deterministic but non-stationary hazard rate scenario. We now consider the setting where the agent is exposed to a grid environment with the same dimensions and features as the previous section, but featuring a non-stationary hazard rate that follows a sinusoidal predefined pattern.

Apart from the selection of an appropriate set of an appropriate set of discount factors, which proceeds the same way as in the stationary hazard rate scenario, our agents follow a slightly different procedure to update their beliefs given the oscillations and the non-stationary nature of the hazard rate. In this scenario, the agent's belief about the hazard rate is updated at each episode, following the Algorithm 5.1. Therefore, the agent must be tuned by the hyperparameter $scalingFactor$ that controls the length of the agent's memory of past hazard rates. This hyperparameter can be tuned to control the agent's sensitivity to the oscillations of the hazard rate, allowing the agent to improve its performance in each kind of scenario. After testing some values for the $scalingFactor$ hyperparameter, we found that the best performance was achieved with a value of 10, being the selection process supported by the results shown in Table 6.2, referring to the mean squared error between the agent's estimation and the true hazard rate given different values for the $scalingFactor$ hyperparameter.

By choosing a value of 10 for the $scalingFactor$ hyperparameter, we induce a controlled "memory re-

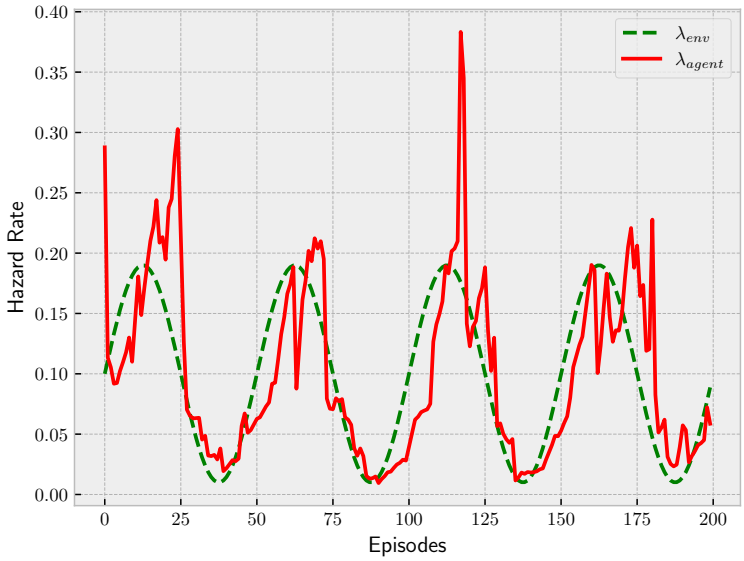

Figure 6.5: Adaptation to an oscillating hazard rate. The plot shows the variation of the agent's belief about the hazard rate, λ_{agent}, for 200 episodes of interaction with a hazardous environment with an oscillating hazard rate, $\lambda_{env}(t) = 0.1 + 0.09\sin(0.04\pi t)$, where t is the episode number.

set" in the agent's belief about the hazard rate that occurs every 10 episodes, allowing the agent to adapt to the oscillations of the hazard rate and improve its performance. Moreover, the obtained results validate the importance of the $scalingFactor$ hyperparameter in adapting the agent to the oscillating hazard and the inability of the standard adaptation mechanism evaluated in the fixed hazard rate scenario to deal with this new scenario. This conclusion comes from the fact that, when the $scalingFactor$ hyperparameter is set to 1 (equivalent to the standard adaptation mechanism), the error observed between the agent's estimation and the true hazard rate is significantly higher than every other tested values.

With both the evaluation environment and agents set, we conduct a series of experiments to evaluate both the adaptability of the proposed mechanism to the non-stationarity of the hazard rate and its end performance. We also compare our agents' final performance against those with prior knowledge of the hazard rate.

As depicted in Figure 6.5, we analyze the variations of the agent's belief about the hazard rate given the oscillations of the true hazard rate of the environment. From the observed behavior we extract that the agent's belief about the hazard rate is able to adapt properly to the continuous oscillation of the true hazard rate. Moreover, we observe the impact of the $scalingFactor$ hyperparameter in some of the spikes observed in the agent's belief about the hazard rate when the true hazard rate approaches its maximum value. These spikes are a result of the memory reset induced by the $scalingFactor$ hyperparameter. This effect is explained by the fact that, after observing a sequence of episodes with an increasing tendency, the agent's belief tends to follow the same pattern, leading to the aforementioned

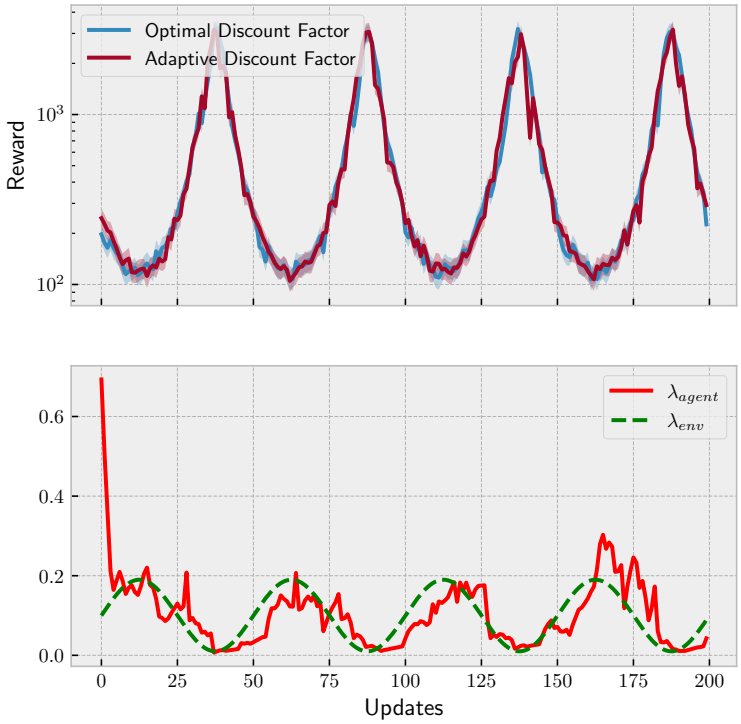

Figure 6.6: Comparison of agent's performance and belief about the hazard rate in an oscillating hazard rate scenario. The plot shows the total rewards collected by the agent between updates of its belief about the hazard rate, λ_{agent}, in an oscillating hazard rate environment, $\lambda_{env}(t) = 0.1 + 0.09\sin(0.04\pi t)$, where t is the episode number. It is also represented a baseline agent that uses the optimal discount factor mapped from the true hazard rate at each episode. The results are averaged over 500 episodes and the shaded area corresponds to a 99% confidence interval.

spikes. However, when the true hazard rate starts to decrease after reaching its maximum value, the capacity to stop considering the far past episodes is essential to capture the new tendency in a short period.

We also perform experiments to evaluate the agent's performance and compare it with the performance of the "oracle" agents that have full knowledge of the hazard rate at each episode, being the results shown in Figure 6.6. As seen, we observe that the agent's performance is significantly affected by the oscillations of the hazard rate, as it was expected. However, both performances of our proposed method and the "oracle" agents are very similar, showing that the proposed adaptation mechanism can adapt to the oscillating hazard rate and use policies with matching performance to the optimal ones.

As discussed earlier, comparing our adaptation framework to the direct use of the precomputed policies helps validating the added value of the adaptation process. Thus, we evaluate the adaptive agent against some baseline agents using the precomputed policies that support the adaptive agent. The results are shown in Figure 6.7 and suggest that, even though the obtained rewards are highly influenced by the oscillations of the hazard rate, the adaptive agent is able to achieve, at any time,

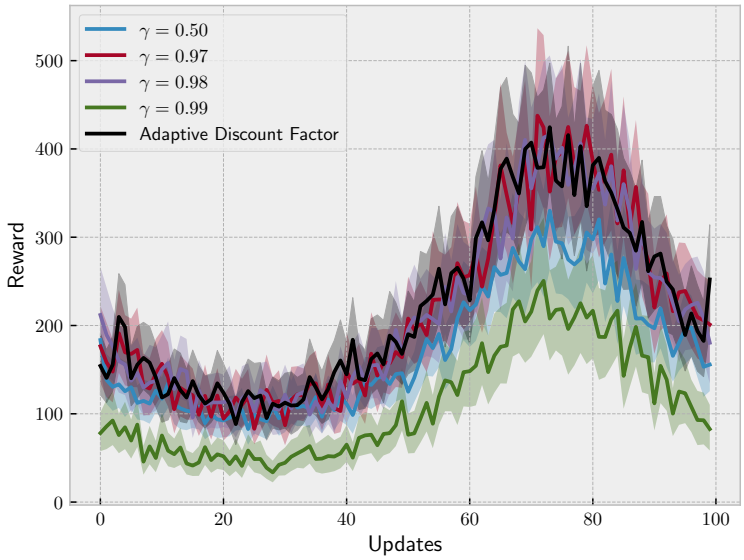

Figure 6.7: Performance comparison between an adaptive agent and agents applying policies that compose the set of precomputed policies of the adaptive agent in a non-stationary hazard rate scenario. The plot shows the average rewards collected by the adaptive discounting agent, over 100 estimate updates, and baseline agents that apply the policies that compose the set of precomputed policies of the adaptive agent. The agents interact with an environment with an oscillating hazard rate, $\lambda_{env}(t) = 0.1 + 0.45\sin(0.02\pi t)$, and the results are averaged over 500 episodes. The shaded area corresponds to a 99% confidence interval.

similar performance to the best baseline agent, clearly outperforming some of the other baseline agents. This result reinforces the effectiveness of the proposed adaptation mechanism in dealing with the non-stationarity of the hazard rate.

6.2.3 Takeaways

The experiments presented in this chapter offer valuable insights into the effectiveness of the proposed adaptation mechanism in hazardous environments. The results demonstrate that the mechanism successfully adapts to both stationary and non-stationary hazard rates. Specifically, our solution accurately estimates the hazard rate of the environments and selects the most appropriate policy for each belief, outperforming baseline agents using precomputed policies and achieving performance comparable to "oracle" agents with full knowledge of hazard dynamics. Although our agents were tested only in the hazardous grid environment, the promising results suggest that this approach may generalize to other environments with similar characteristics, such as the sensitivity to the agent's time horizon and risk preferences.

7

Conclusion

Contents

In this chapter, we summarize the key findings and contributions of the work presented in this thesis and discuss potential directions for future research, aiming to provide better insights into the application of RL and planning techniques in hazardous environments and further development of the proposed adaptive discount factor framework.

7.1 Summary and Final Remarks

In this thesis, we study and develop agents capable of dealing with the challenges posed by hazardous planning tasks. Such tasks are characterized by the presence of uncertainty and risk that expose the agents to potential hazards that terminate the interaction process prematurely. Hazardous MDPs provide a framework to model these tasks, not including a discount factor as part of the problem definition. This way, the application of the standard and most used discounting mechanism, exponential discounting, may lead to suboptimal behavior since the discount factor applied by the agents may not be properly

adapted to the environment's uncertainty dynamics. Therefore we address the problem of planning in hazardous environments, investigating the research question that motivated our work: *How might the discounting process be adapted to improve the performance of agents in hazardous environments?* To address this question, we give two main contributions. Firstly, we compare the performance of both hyperbolic and exponential discounting mechanisms and study solutions to deal with different settings of uncertainty in hazardous environments, from the presence of a single stationary hazard rate to the presence of non-stationary hazard rates, either deterministically or stochastically changing over time. Secondly, we focus on developing an adaptive discount factor framework to deal with environments featuring deterministic stationary and non-stationary hazard rates proposing a solution that is robust to different settings of uncertainty.

We study the impact of the discount factor on the performance of agents in hazardous environments and compare the performance of the exponential and hyperbolic discounting mechanisms in different settings of uncertainty. Our experimental results show the limitations of the standard exponential discounting, featuring a fixed discount factor may have in hazardous environments, specifically when applied in hazardous MDPs. Therefore, we empirically demonstrate that the application of exponential discounting may lead to suboptimal behavior if the discount factor is not properly adapted to the environment's hazard rates. Moreover, we also find alternatives to the exponential discounting mechanism, such as the hyperbolic discounting, that revealed to be more robust to the uncertainty dynamics of hazardous environments being a better option when the agents face stochastic hazard rates sampled from unknown random distributions.

We also focus on the scenarios where the hazard rates are deterministic, possibly stationary or non-stationary. To mitigate the limitations of the exponential discounting mechanism, we propose a novel adaptive discount factor framework that is able to perform the adaptation of the discount factor to the environment's hazard rates after each episode of interaction, taking advantage on the episodes' duration to estimate the hazard rates. Our solution shows as a promising approach to improve the agents' performance in hazardous environments, featuring deterministic stationary and non-stationary hazard rates. Our experimental results show that the adaptive discount factor framework is able to improve the agents' performance over time, being also able to match the performance of strong baselines, such as fully informed agents that act under the knowledge of the environment's hazard rates.

In summary, it is important to highlight that planning in hazardous environments is a challenging task, and the application of standard RL or planning techniques may not be straightforward, or even lead to suboptimal behavior. Therefore, the development of novel methods to deal with the uncertainty and risk present in hazardous environments is crucial to improve the agents' performance. The discounting process, being intrinsically related to the concept of risk and uncertainty, plays a crucial role in the agents' decision-making process, and the study and exploration of different discounting mechanisms

and their adaptation to the environment's uncertainty dynamics are essential to improve the agents' performance in hazardous environments.

7.2 Future Work

We now elaborate on potential directions for future work that may be interesting to explore regarding the decision-making process in hazardous environments.

In this work, we only focus on planning in hazardous environments and do not explore the application of the proposed adaptive discount factor framework in the context of RL. Therefore an interesting direction for future work would be to extend our adaptive discount factor framework to the RL setting, where agents do not fully know the environment's dynamics, such as transition probabilities and the reward function. In this case, the direct application of our framework *as is* may not be straightforward, as it was designed for traditional planning settings. Thus, exploring how to adapt the framework to both model-free and model-based RL algorithms would be a valuable area of research.

Regarding the details of the adaptive discount factor framework, introducing more flexibility in the adaptation process would be beneficial, particularly in scenarios with non-stationary hazard rates. Currently, the framework requires a predefined hyperparameter to control the influence of past observations on the hazard rate estimate. Depending on the environment's hazard rate dynamics and the variability in hazard rates, this hyperparameter can slightly affect agent performance. A potential improvement would be to eliminate the need for this hyperparameter by developing a model that estimates hazard rates more flexibly, leveraging the entire history of observations.

Finally, since our adaptive discount factor framework is designed to perform inter-episodic updates of the discount factor, it would be interesting to explore the development of an online adaptation mechanism that adjusts the discount factor during the episodes. This would allow agents to adapt the discount factor in real-time, taking advantage of the information gathered during the interaction process to update the followed policies more quickly. This approach could possibly cover a wider range of uncertainty dynamics, particularly in environments with non-stationary hazard rates that fluctuate within the same episode or depend on the environment's states.

Bibliography

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

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

[3] P. D. Sozou, "On hyperbolic discounting and uncertain hazard rates," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 265, no. 1409, pp. 2015–2020, 1998.

[4] C. Sherstan, "Representation and general value functions," Ph.D. dissertation, University of Alberta, 2020.

[5] C. Sherstan, S. Dohare, J. MacGlashan, J. Günther, and P. M. Pilarski, "Gamma-nets: Generalizing value estimation over timescale," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5717–5725.

[6] L. Green, E. B. Fisher, S. Perlow, and L. Sherman, "Preference reversal and self control: Choice as a function of reward amount and delay." *Behaviour Analysis Letters*, 1981.

[7] H. Rachlin and L. Green, "Commitment, choice and self-control 1," *Journal of the experimental analysis of behavior*, vol. 17, no. 1, pp. 15–22, 1972.

[8] H. Rachlin, A. Raineri, and D. Cross, "Subjective probability and delay," *Journal of the experimental analysis of behavior*, vol. 55, no. 2, pp. 233–244, 1991.

[9] R. Bellman, "A markovian decision process," *Journal of mathematics and mechanics*, pp. 679–684, 1957.

[10] D. Fink, "A compendium of conjugate priors," *See http://www. people. cornell. edu/-pages/df36/CONJINTRnew% 20TEX. pdf*, vol. 46, 1997.

[11] L. Green and J. Myerson, "Exponential versus hyperbolic discounting of delayed outcomes: Risk and waiting time," *American Zoologist*, vol. 36, no. 4, pp. 496–505, 1996.

[12] L. Green, N. Fristoe, and J. Myerson, “Temporal discounting and preference reversals in choice between delayed outcomes,” *Psychonomic Bulletin & Review*, vol. 1, pp. 383–389, 1994.

[13] Z. Kurth-Nelson and A. D. Redish, “Temporal-difference reinforcement learning with distributed representations,” *PLoS One*, vol. 4, no. 10, p. e7362, 2009.

[14] C. J. Watkins and P. Dayan, “Q-learning,” *Machine learning*, vol. 8, pp. 279–292, 1992.

[15] J. Romoff, P. Henderson, A. Touati, E. Brunskill, J. Pineau, and Y. Ollivier, “Separating value functions across time-scales,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 5468–5477.

[16] C. Reinke, E. Uchibe, and K. Doya, “Average reward optimization with multiple discounting reinforcement learners,” in *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part I 24*. Springer, 2017, pp. 789–800.

[17] W. H. Alexander and J. W. Brown, “Hyperbolically discounted temporal difference learning,” *Neural computation*, vol. 22, no. 6, pp. 1511–1527, 2010.

[18] R. S. Sutton, J. Modayil, M. Delp, T. Degris, P. M. Pilarski, A. White, and D. Precup, “Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction,” in *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, 2011, pp. 761–768.

[19] M. White, “Unifying task specification in reinforcement learning,” in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 3742–3750.

[20] D. Blackwell, “Discrete dynamic programming,” *The Annals of Mathematical Statistics*, pp. 719–726, 1962.

[21] A. Naik, R. Shariff, N. Yasui, H. Yao, and R. S. Sutton, “Discounted reinforcement learning is not an optimization problem,” *arXiv preprint arXiv:1910.02140*, 2019.

[22] Y. Tang, M. Rowland, R. Munos, and M. Valko, “Taylor expansion of discount factors,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 130–10 140.

[23] Z. Xu, H. P. van Hasselt, and D. Silver, “Meta-gradient reinforcement learning,” *Advances in neural information processing systems*, vol. 31, 2018.

[24] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, “Dream to control: Learning behaviors by latent imagination,” *arXiv preprint arXiv:1912.01603*, 2019.

[25] M. Kim, J.-S. Kim, M.-S. Choi, and J.-H. Park, “Adaptive discount factor for deep reinforcement learning in continuing tasks with uncertainty,” *Sensors*, vol. 22, no. 19, p. 7266, 2022.

Algorithms Pseudocode

In this appendix, we provide the pseudocode for the algorithms that compose the modules of the adaptive discounting framework. The Algorithm A.1, describes the computation of the set of policies for different time horizons, while the Algorithm A.2, describes the policy selection process based on the current estimate of the discount factor.

Algorithm A.1: Computation of the set of policies

Input: τ_{max}: Maximum Time Horizon, δ_τ: Spacing Factor
Output: Π: Set of computed policies for different time horizons
begin
  $\Pi \leftarrow \emptyset$
  for $\tau \leftarrow 2$ **to** τ_{max} **by** δ_τ **do**
    $\gamma \leftarrow 1 - \frac{1}{\tau}$
    $\pi_\gamma \leftarrow$ ValueIteration(γ)
    $\Pi \leftarrow \Pi \cup \{(\gamma, \pi_\gamma)\}$
  return Π

Algorithm A.2: Policy selection

Input: Π: Set of precomputed policies, γ_{est}: Current estimate of the discount factor
Output: π: Selected policy
begin
  $\pi \leftarrow \emptyset$
  $minDist \leftarrow \infty$
  for each $(\gamma, \pi_\gamma) \in \Pi$ **do**
    $dist \leftarrow |\gamma - \gamma_{est}|$
    if $dist < minDist$ **then**
      $minDist \leftarrow dist$
      $\pi \leftarrow \pi_\gamma$
  return π