A Black Swan Hypothesis in Markov Decision Process via Irrationality

Hyunin Lee HYUNIN@BERKELEY.EDU
University of California, Berkeley
Berkeley, CA 94720-1776, USA

David Abel DMABEL@DEEPMIND.COM
Google Deepmind
London, EC4A 3TW, United Kingdom

Ming Jin JINMING@VT.EDU
Virginia Tech
Blacksburg, VA 24061, USA

Javad Lavaei LAVAEI@BERKELEY.EDU
University of California, Berkeley
Berkeley, CA 94720-1776, USA

Somayeh Sojoudi SOJOUDI@BERKELEY.EDU
University of California, Berkeley
Berkeley, CA 94720-1776, USA

Abstract

Black swan events are statistically rare occurrences that carry extremely high risks. A typical view of defining black swan events is heavily assumed to originate from an unpredictable time-varying environments; however, the community lacks a comprehensive definition of black swan events. To this end, this paper challenges that the standard view is *incomplete* and claims that high-risk, statistically rare events can also occur in unchanging environments due to human misperception of their value and likelihood, which we call as *spatial black swan event*. We first carefully categorize black swan events, focusing on spatial black swan events, and mathematically formalize the definition of black swan events. We hope these definitions can pave the way for the development of algorithms to prevent such events by rationally correcting human perception.

Keywords: Risk, Reinforcement learning, Irrationality, Cumulative prospect theorem.

1 Introduction

Life is the cumulative effect of a handful of significant shocks.

The Black Swan: The Impact of the Highly Improbable, Nassim Nicholas Taleb

The reason behind the Lehman Brothers bankruptcy, an unexpected event with extremely negative impacts on the world economy, remains controversial. However, a strong explanation points to the irrationality of human decision-making. The firm declared bankruptcy within 72 hours without any precursor (McDonald and Robinson (2009)), and the only factor that changed during those three days was investors' *faith* in the company (Housel (2023); Mawutor (2014); Fleming and Sarkar (2014)). *Faith* was intrinsically believed by investors as an axiom. They made optimal decisions (being rational) based on this faith, which turned out to be suboptimal (being irrational) once the faith was revealed to be false during those 72 hours . Referring to the unexpected bankruptcy event mentioned previously, we call such a rare and high-risk event, often rationalized retrospectively with the benefit of hindsight, a *black swan* (Taleb (2010)).

Black swan events remain one of the unsolved problems in machine learning safety (Hendrycks et al. (2021)) , and this work focuses on the origin of black swan events from the perspective of *human perception*, our main messages being Hypothesis 2 and Definition 7, rather than on how to design robust algorithms against them . We expect that providing a novel perspective to understand black swan events can offer new insights for designing safe machine learning algorithms . Supported by extensive documentation of black swan events, such as the dissolution of the Soviet Union , the terrorist attacks of September 11, 2001, and the Brexit vote (Taleb (2010)), we focus on specific types of black swans that occur even in a stationary environment . We refer to these as *spatial* black swans. Based on the above example , we deduce that certain types of black swan occur due to *misperception* in the way humans perceive the world (Hypothesis 2). Executing an optimal policy based on misperception inevitably causes the agent to encounter unavoidable risks, which we define as spatial black swans (Definition 7) . From a broad perspective, our work proposes the following informal hypothesis regarding the origin of black swan events and provides an informal definition of spatial black swans built upon this hypothesis as follows.

Hypothesis 1 (Black swan origin (informal)). *Black swan events occur due to temporal and spatial misperception, and spatial black swans occur due to human misperception of the real world.*

Definition 1 (Spatial black swan (informal))**.** *A spatial black swan event is a state and action pair that humans perceive as an infeasible event and assess its reward in a pessimistic manner.*

Paper structure. Starting from the informal hypothesis and definition of spatial black swans as mentioned above (Hypothesis 1, Definition 1), we structure the paper to provide concrete foundations for our main message: Hypothesis 2, and Definition 7 as follows. First, in Section 2, we define spatial and temporal black swans, then focus on *spatial black swans* for the rest of the paper. Section 3 provides evidence for the soundness of Hypothesis 1. Specifically, in Subsection 3.1, we emphasize the necessity of a novel perspective to understand black swans by demonstrating that the existing decision-making frameworks under risk are insufficient, and in Subsection 3.2, we explain how misperception can be related to black swans. Section 4 introduces Cumulative Prospect Theorem (CPT), a well-known theorem for explaining irrational human behavior in the real world, to mathematically formalize the term *misperception* in Hypothesis 1 and Definition 1. In Section 5, we introduce three types of Markov Decision Processes: Universe MDP, Human MDP, and Human-Estimation MDP, which we denote as $\mathcal{M}$, $\mathcal{M}^\dagger$, and $\widehat{\mathcal{M}}^\dagger$, respectively. Before presenting the concrete definition and hypothesis of spatial black swans, we provide case studies in Section 6 to illustrate how human misperception can lead to suboptimal policy (Theorems 1, 2, and 3). Subsequently, in Section 7, we present our main message, proposing the spatial black swan hypothesis (Hypothesis 2) and a definition of spatial black swan events (Definition 7) utilizing $\mathcal{M}$, $\mathcal{M}^\dagger$, $\widehat{\mathcal{M}}^\dagger$, and CPT. It is worth noting that Definition 7 employs notations from $\mathcal{M}$, $\mathcal{M}^\dagger$, and $\widehat{\mathcal{M}}^\dagger$ and is mathematically characterized by concepts from CPT. Finally, in Section 8, we explore the properties of spatial black swan events, particularly how their presence establishes a lower bound on achieving true policy performance (Theorem 4) and affects the timing of black swan event occurrences (Theorem 5).

Notations. The sets of natural, real, nonnegative, and nonpositive real numbers are denoted by $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}_{\geq 0}$, and $\mathbb{R}_{\leq 0}$ respectively. For a finite set Z, the notation $|Z|$ represents its cardinality, and $\Delta(Z)$ denotes the probability simplex on Z. Given $X, Y \in \mathbb{N}$ with $X < Y$, we define $[X] \coloneqq \{1, 2, \ldots, X\}$, the closed interval $[X, Y] \coloneqq \{X, X+1, \ldots, Y\}$. For $x \in \mathbb{R}_+$, the floor function $\lfloor x \rfloor$ is defined as $\max\{n \in \mathbb{N} \cup \{0\} \mid n \leq x\}$.

Markov Decision Process. We consider a finite horizon non-stationary Markov Decision Process (MDP) denoted as $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \{P_t\}_{t=0}^{T}, \{R_t\}_{t=0}^{T}, \gamma \rangle$, where $\mathcal{S}$ represents the state space, $\mathcal{A}$ denotes the action space, $P_t : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability function at time t, $R_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function as time t, γ is the discount factor, and $T \in \mathbb{N}$ is the horizon length. We denote a policy as $\pi \in \Pi$, where $\Pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a set of policies, and its performance as $J(\pi) = \mathbb{E}_{\pi, \mathcal{M}}[G]$ where $G = \sum_{t=0}^{T-1} R_t(s_t, a_t)$ is a scalar return. We denote T length trajectory from $\mathcal{M}$ with policy π as $\{s_0, a_0, r_0, s_1, a_1, r_1, \cdots, s_T\}$. Note that based on the non-stationary finite horizon MDP , we restrict our analysis to a stationary MDP to elaborate on *spatial* black swans in Sections 4, 5, 7, and 8, but fully utilized non-stationarity for the case study in Section 6, and black swan clarification in Section 2.

Problem setting. Our problem setting is that humans perceive the world $\mathcal{M}$ as $\mathcal{M}^\dagger$ and collect misperceived trajectory data $\{s_0^\dagger, a_0, r_0^\dagger, s_1^\dagger, a_1, r_1^\dagger, \cdots, s_T^\dagger\}$, where misperception occurs on the state and reward through a function $g: \mathcal{S} \to \mathcal{S}$ and a function $u: \mathbb{R} \to \mathbb{R}$, respectively. This means that the state and reward are distorted as $s_t^\dagger = g(s_t)$ and $r_t^\dagger = u(r_t)$ for all $t \in [T]$. Humans then estimate the reward and visitation probability (or transition probability) of $\mathcal{M}^\dagger$ from the misperceived trajectory data to form their estimation model $\widehat{\mathcal{M}}^\dagger$.

Main message and supporting evidence. Our main message (Hypothesis 2, Definition 7) starts with the observation that even though humans can decrease the *estimation* gap to make $\widehat{\mathcal{M}}^\dagger$ converge to $\mathcal{M}^\dagger$ by rolling out a longer horizon T or better learning algorithms to reconstruct visitation probability and reward function. However, the uncertainty of estimating $\mathcal{M}$ has a lower bound due to the *perception* gap between $\mathcal{M}$ and $\mathcal{M}^\dagger$. In this sense, we convey that spatial black swans can occur due to the perception gap, even if the agent has zero estimation error and computes its optimal policy from $\widehat{\mathcal{M}}^\dagger$.

Our main message is supported by our primary theorem, Theorem 4, which provides two key insights. First, it demonstrates that the policy performance gap between $\mathcal{M}$ and $\widehat{\mathcal{M}}^\dagger$ has a lower bound due to the perception gap between $\mathcal{M}$ and $\mathcal{M}^\dagger$, even though the estimation error between $\mathcal{M}^\dagger$ and $\widehat{\mathcal{M}}^\dagger$ asymptotically converges to zero as the horizon length T increases. Second, Theorem 4 quantifies this lower bound in terms of three factors: the number of spatial black swans, the minimum probability of spatial black swans, and the magnitude of misperception between $\mathcal{M}$ and $\mathcal{M}^\dagger$. It also highlights that the effective number of spatial black swans (number × probability) is more significant than considering sole probability or the number of occurrences for future algorithm design. Additionally, Theorem 5 provides a hitting time of spatial black swans.

Contributions. This work offers three pivotal contributions:

- A novel perspective on black swan events: We propose that these events arise from human misperception of space and time. This idea, introduced as Hypothesis 1, is further developed into Hypothesis 2 in Section 7, building on foundations laid in Sections 3, 4, and 5.

- A novel definition of spatial black swans: We introduce and define a specific type of black swan event (initially presented in Definition 1 and later formalized as Definition 7 in Section 7). This definition builds upon our core hypothesis 2, providing a new framework for understanding these rare but impactful events.

- Theoretical analysis of spatial black swans: We conduct a rigorous examination of how these events affect optimal policy performance (Theorem 4) and their occurrence probability over time (Theorem 5). This analysis is supported by case studies (Theorems 1, 2, and 3) demonstrating the impact of human misperception on decision-making in various scenarios.

2 Spatial and temporal blackswans

Based on Hypothesis 1, this prompts us to investigate the concept of *misperception*. Initially, we must clearly define what constitutes *perception*. According to the definition of an agent by Barandiaran et al. (2009) and Orseau and Ring (2012), an agent views its environment through the lens of so-called *spatio-temporal* dimensions. Consequently, if a black swan event arises from misperception, the conceptual framework of perception leads us to question whether the misperception originates from spatial or temporal dimensions. This first leads us to define the black swan event dimension as follows.

Definition 2 (Black swan dimension)**.** *In a Markov Decision Process, a dimension of a black swan event is defined on $\mathcal{S} \times \mathcal{A} \times [T]$ with $\mathcal{S}$ representing the state space and $\mathcal{A}$ representing the action space and T is time (horizon length). We collect black swan events as a set $\mathcal{B}$ ($\mathcal{B}$ will be more elaborated on Hypothesis 2).*

In Definition 2, unit time refers to any variable that represents its time heterogeneity such as discrete step or discrete episode (Lee et al. (2024) or discrete real-time (Abel et al. (2024); Dong et al. (2022)) in general Markov Decision Process setting.

Assumption 1. *For fixed time $t' \in [T]$, black swan events at time t', i.e., the set $\{(s,a) | (s,a,t = t') \in \mathcal{B}\}$ is invariant set under static agent's perception of MDP at time t'.*

Assumption 1 ensures that if the environment and the agent's perception of the environment are fixed at time t, then the black swan events are fully determined. This assumption is also used to classify black swans in Proposition 1, Example 1, and Remark 1.

The definition 2 leads us to first classify whether the black swan comes from a misconception within the space (s,a), termed *spatial mispercpetion*, or a misperception along the unit time (t), termed *temporal misperception*. We define *temporal misperception* as the inherent inaccuracies in time-series data prediction algorithms from agent, exemplified by black swans such as COVID-19 or earthquakes, and define *temporal black swans* as those originating from temporal misperception. This type of misperception inevitably results from the non-stationarity of the environment, which impacts the algorithms' ability to predict future events accurately. In contrast, we define *spatial misperception* as a misperception that occurs in a stationary environment, often due to incorrect assessments or misunderstandings of the spatial aspects of the data. We define *spatial black swans* as those originating from spatial misperception, such as the Russia-Ukraine war, the Lehman Brothers bankruptcy, the Brexit vote or the September 11, 2001 terrorist attacks. Intuitively, temporal black swans occur due to the non-stationarity of the environment, while spatial black swans occur even in stationary environments. Based on the above description, we provide a definition of spatial and temporal black swans as follows.

Definition 3 (Spatial black swans). *For a given MDP $\mathcal{M}$ for a time interval $[T]$, suppose (s, a, t_{bs}) is a black swan where $t_{bs} \in [T]$. If (s, a, t) is a black swan event for $\forall t \in [T]$, then we define (s, a, t_{bs}) as a spatial black swan.*

Definition 4 (Temporal black swans). *For a given MDP $\mathcal{M}$ for a time interval $[T]$, suppose (s, a, t_{bs}) is a black swan where $t_{bs} \in [T]$. If there exists $t \in [T]$ such that (s, a, t) is not a black swan event, then we define (s, a, t_{bs}) as a temporal black swan.*

Based on Definitions 3, 4 and Assumption 1, we can always identify a unit time interval that classifies any black swan events as a spatial black swan as following Proposition 1 and Example 1.

Proposition 1. *If (s, a, t_{bs}) is a black swan event, then there exists a time interval $[t'_1, t'_2] \subseteq [T]$ that classifies (s, a, t_{bs}) as a spatial black swan within $[t_1, t_2]$.*

We provide an intuitive interpretation for Proposition 1 as the following example.

Example 1. *Suppose (s, a, t_{bs}) is a black swan event.*

- *case 1.* $\mathcal{M}$ *is a non-stationary MDP where P_t, R_t changes for every unit time, i.e. $P_t \neq P_{t+1}, R_t \neq R_{t+1}$. If $t_1 = t_2 = t_{bs}$, then (s, a, t_{bs}) is a spatial black swan. If $t_1 \neq t_2$ where $t_{bs} \in [t_1, t_2]$, then (s, a, t_{bs}) could not be identified as spatial black swan or temporal black swan.*
- *case 2.* $\mathcal{M}$ *is a piecewise non-stationary MDP where P_t, R_t changes for $\lfloor T/k \rfloor$ times, i.e. $P_t = P_{t+1}, R_t = R_{t+1}$ for $t \in [kj, kj + (k-1)], j = 0, 1, \cdots, \lfloor T/k \rfloor$. If $t_{bs} \in [kj_{bs}, kj_{bs} + (k-1)]$, then (s, a, t_{bs}) is a spatial black swans when $t_1 = kj_{bs}, t_2 = kj_{bs} + (k-1)$.*
- *case 3.* $\mathcal{M}$ *is stationary MDP where $P_t = P_{t+1}, R_t = R_{t+1}$ for $\forall t \in [T-1]$, then (s, a, t_{bs}) is a spatial black swan regardless of interval $[t_1, t_2]$.*

We formulate case 3 of Example 1 as the following Remark.

Remark 1. *If $\mathcal{M}$ is stationary, then any black swan events (s, a, t) are spatial black swans, denoted as (s, a).*

This work investigates Remark 1. For the remainder of the paper, we focus on spatial black swans by concretizing Hypothesis 1 and Definition 1 in the context of spatial black swans. Based on Definition 3 and Remark 1 within the framework of Markov Decision Processes, we provide an informal proposition of spatial black swans by adapting the preliminary definition proposed by Taleb (2010).

Proposition 2 (Spatial black swan (informal)). *For given MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$. We define an event σ as a state and action $(s, a) \in \mathcal{S} \times \mathcal{A}$. If σ is a spatial black swan event, then it satisfies that*

1. *σ is a high-risk event.*
2. *σ is a rare event.*
3. *After σ is first observed at time t, it is rationalized by hindsight, as if it could have been expected.*

Note that the first and second properties of Proposition 2 characterize the appearance and nature of such events, while the third property elucidates strategies for avoiding black swans, specifically addressing their post-event properties. Therefore, we elaborate on the first and second properties of Proposition 2 through three different MDPs ($\mathcal{M}$, $\mathcal{M}^\dagger$, $\widehat{\mathcal{M}}^\dagger$) introduced in Section 5 and CPT, which will be introduced in Section 4, then complete it as Definition 7.

3 Necessity of a new perspective to understand black swans and evidence for Hypothesis 1

In this section, we focus not only on addressing the necessity of a new perspective to understand black swan events but also on providing evidence for the proposed perspective of black swan origin (Hypothesis 1). This is concretized by examining the following two questions. First, in Subsection 3.1, we discuss the insufficiency of existing decision-making rules under risk by exploring related works, which support the need for a new perspective to understand black swans. Specifically, we address *why existing safe reinforcement learning strategies for solving Markov Decision Processes are insufficient to handle black swan events?*. If this premise is validated, then in Subsection 3.2, we elaborate on the motivation and related works that support our informal hypothesis of black swan origin (Hypothesis 1). Specifically, we explore *how irrationality relates to misperception and how irrationality could bring about black swan events.*

3.1 Decision Making Under Risk

Based on the comprehensive survey on safe reinforcement learning in Garcıa and Fernández (2015), the algorithms can be classified into threefold: worst case criterion, risk-sensitive criterion and constraint criterion. We elaborate on why the existence of black swans in the environment renders these three approaches insufficient.

Worst case criterion. Learning algorithms of the worst case criterion focus on devising a control policy that maximizes policy performance under the least favorable scenario encountered during the learning process, defined as $\max_{\pi\in\Pi}\min_{w\in\mathcal{W}} J(\pi; w)$, where $\mathcal{W}$ represents the set of uncertainties. This criterion can be categorized based on whether $\mathcal{W}$ is defined in the environment or in the estimation of the model. The presence of black swan events in the worst case, where $\mathcal{W}$ represents aleatoric uncertainty of the environment (Heger (1994); Coraluppi (1997); Coraluppi and Marcus (1999, 2000)), results in overly conservative, and thus potentially ineffective, policies. This occurs because the significant impact of black swan events inflates the size of $\mathcal{W}$, even though such events are rare. In practical terms, this could manifest itself as abstaining from any economic activity (π), such as not investing in stocks or not depositing a check against future potential bankruptcies ($\min_{w\in\mathcal{W}} J(\pi;w)$) in order to maximize its income ($\max_{\pi\in\Pi}(\cdot)$), or maintaining constant health precautions such as wearing mask or maintaining distance with groups (π) to prepare for a possible pandemic ($\cdot = \min_{w\in\mathcal{W}} J(\pi;w)$) in order to maintain its health ($\max_{\pi\in\Pi}$). Similarly, when $\mathcal{W}$ encompasses the uncertainty of the model parameter Bagnell et al. (2001); Iyengar (2005); Nilim and El Ghaoui (2005); Wiesemann et al. (2013); Xu and Mannor (2010) - as seen in robust MDP or distributionally robust MDP - this aligns closely with our black swan hypothesis, where misperception of the world model is similar to uncertainty in model estimation. However, the need to accommodate black swan events requires enlarging the possible set of models ($|\mathcal{W}|$), leading to extremely conservative policies. This can be likened to performing an overly pessimistic portfolio optimization (π), where every bank is assumed to have a minimal but possible risk of bankruptcy ($\min_{win\mathcal{W}} J(\pi;w)$), thus influencing asset allocation strategies ($\max_{\pi\in\Pi}\min_{w\in\mathcal{W}} J(\pi;w)$) to be extremely conservative in asset investing.

Risk sensitive criterion. Risk-sensitive algorithms strike a balance between maximizing reinforcement and mitigating risk events by incorporating a sensitivity factor $\beta < 0$ (Howard and Matheson (1972); Chung and Sobel (1987); Patek (2001)). These algorithms optimize an alternative value function $J(\pi) = \beta^{-1}\log\mathbb{E}_\pi[\exp^{\beta G}]$, where β controls the desired level of risk. However, it is recognized that associating risk with the variance of the return is practical, as in $J(\pi) = \beta^{-1}\log\mathbb{E}_\pi[\exp^{\beta G}] = \max_{\pi\in\Pi}\mathbb{E}_\pi[G] + \frac{\beta}{2}\mathrm{var}(G) + \mathcal{O}(\beta^2)$, and the existence of black swan events does not significantly affect the returns of variance ($\mathrm{var}(G)$) due to their rare nature. It should be noted that risk-sensitive approaches are not well suited for handling black swan events, as the same policy performance with small variance can entail substantial risks (Geibel and Wysotzki (2005)). More generally, the objective of the exponential utility function is one example of risk-sensitive learning based on a trade-off between return and risk, i.e., $\max_{\pi\in\Pi}(\mathbb{E}_\pi[G] - \beta w)$ (Zhang et al. (2018)), where w is replaced by $\mathrm{Var}(G)$. This approach is known in the literature as the variance-penalized criterion (Gosavi (2009)), the expected value-variance criterion (Taha (2007); Heger (1994)), and the expected-value-minus-variance criterion (Geibel and Wysotzki (2005)). However, a fundamental limitation of using return variance as a risk measure is that it does not account for the fat tails of the distribution (Huisman et al. (1998); Bradley and Taqqu (2003); Bubeck et al. (2013); Agrawal et al. (2021)). Consequently, risk can be underestimated due to the oversight of low probability but highly severe events (black swans).

Furthermore, a critical question arises regarding whether the log-exponential function belongs to *appropriate utility function class* for defining *real-world risk*. Risk-sensitive MDPs have been shown to be equivalent to robust MDPs that focus on maximizing the worst-case criterion, indicating that the log-exponential utility function may not be beneficial in the presence of black swans (Osogami (2012); Moldovan and Abbeel (2012); Leqi et al. (2019)). This issue was first raised by Leqi et al. (2019) and led to the proposal of a more realistic risk definition called 'Human-aligned risk', which also incorporates human misperception akin to our informal black swan hypothesis (Hypothesis 1).

Constrained Criterion. The constrained criterion is applied in the literature to constrained Markov processes where the goal is to maximize the expected return while maintaining other types of expected utilities below certain thresholds. This can be formulated as $\max_{\pi \in \Pi} \mathbb{E}_{\pi}[G]$ subject to N multiple constraints $h_i(G) \leq \alpha_i$, for $i \in [N]$, where $h_i : \mathbb{R} \to \mathbb{R}$ is a function of return G (Geibel (2006)). Typical constraints include ensuring the expectation of return exceeds a specific minimum threshold (α), such as $\mathbb{E}[G] \geq \alpha$, or softening these hard constraints by allowing a permissible probability of violation (ϵ), such as $\mathbb{P}(\mathbb{E}[G] \geq \alpha) \geq 1 - \epsilon$, known as chance-constraint (Delage and Mannor (2010); Ponda et al. (2013)). Constraints might also limit the return variance, such as $\mathrm{Var}(G) \leq \alpha$ (Di Castro et al. (2012)). However, the presence of black swans highlights one of the challenges with the Constrained Criterion, specifically the appropriate selection of α. The presence of black swans necessitates a lower α, which in turn leads to more conservative policies. Furthermore, a black swan event is determined at least by the environment's state and its action, rather than its full return. Therefore, constraints should be redefined over more fine-grained inputs—not merely returns, but in terms of state and action—which leads to our definition of black swan dimensions (Definition 2).

3.2 How irrationality relates with spatial black swans.

Before starting Subsection 3.2, we clarify that the term *irrationality* is used here to denote rational behavior based on a false belief. In this subsection, we first review existing work on the four rational axioms and then claim how two of these axioms should be modified to account for *irrationality* in human decision-making.

Rationality in decision making. In the foundation of decision theory, rationality is understood as internal consistency (Sugden (1991); Savage (1972)). A prerequisite for achieving rationality in decision making is the ability to compare outcomes, denoted as set $\mathcal{O}$, through a *preference* relation in a *rational* manner. In Neumann and Morgenstern (1944), it is demonstrated that preferences, combined with *rationality axioms* and probabilities for possible outcomes, denoted as p_i which is a probability of outcome $o_i \in \mathcal{O}$, imply the existence of utility values for those outcomes that express a preference relation as the expectation of a scalar-valued function of outcomes. Define the choice (or lotteries) as set $\mathcal{C}$, which is a combination of selecting total n outcomes, that is, $\sum_{i=1}^{n} p_i o_i$. The essential rationality axioms are as follows.

1. Completeness: Given two choices, either one is preferred over the other or they are considered equally preferable.

2. Transitivity: If A is preferred to B and B is preferred to C, then A must be preferred to C.

3. Independence: If A is preferred to B, and a event probability $p \in [0,1]$, then $pA + (1-p)C$ should be preferred to $pB + (1-p)C$.

4. Continuity: If A is preferred to B and B is preferred to C, there exists a event probability $p \in [0,1]$ such that B is considered equally preferable to $pA + (1-p)C$.

Expanding on these axioms, Sunehag and Hutter (2015) extends rational choice theory to encompass the full reinforcement learning problem, further axiomatizing the concept in Sunehag and Hutter (2011) to establish a rational reinforcement learning framework that facilitates optimism, crucial for systematic explorative behavior. Subsequent studies focusing on defining rationality in reinforcement learning, such as Shakerinava and Ravanbakhsh (2022); Bowling et al. (2023), concentrate on the axioms of assigning utilities to all finite trajectories of a Markov Decision Process. Specifically, Shakerinava and Ravanbakhsh (2022); Bowling et al. (2023) clarify the reward hypothesis Sutton that underpins the design of rational agents by introducing an additional axiom to existing rationality axioms. Furthermore, Pitis (2024) explores the design of multi-objective rational agents, and Carr et al. (2024) explores and defines rational feedback in Large Language Models (LLMs) by investigating the existence of optimal policies within a framework of learning from rational preference feedback (LRPF).

Irrationality due to subjective probability. The definition of irrationality and its origins has been extensively investigated through case studies in various fields such as psychology, education, and particularly economics. Simon (1993) defined irrationality as being poorly adapted to human goals, diverging from the norm of human's object, influenced by emotional or psychological factors in decision-making. Subsequently, Martino et al. (2006); Gilovich et al. (2002) further concretized what exactly these *emotional or psychological factors* entail by describing them as information loss during human perception of the real-world. More specifically, Martino et al. (2006) pointed out that in a world filled with symbolic artifacts, where optimal decision-making often requires skills of abstraction and decontextualization, such mechanisms may render human choices irrational. Further studies, such as Opaluch and Segerson (1989), scrutinize more deeply and classify the *irrationality* of human behavior into five factors: subjective probability, regret/disappointment, reference points, complexity, and ambivalence.

In this paper, we focus on the *subjective probability* factor to elucidate the relationship between irrationality and spatial black swans. Opaluch and Segerson (1989) explores subjective probabilities as an early modification to the expected utility model from Neummann and Morgenstern (1944), focusing on decision makers who rely on *personal beliefs* about probabilities rather than objective truths. This minor conceptual shift can lead to significant behavioral changes due to the imperfect information and processing abilities of individuals. Especially, Opaluch and Segerson (1989) highlights the difficulty in accurately estimating the probability of *rare events* - such as black swans - which often leads to critical errors in judgment. These errors occur because rare events provide insufficient data for accurate probability estimation or are misunderstood due to their infrequency, leading to perceptions that such events are either less likely or virtually impossible. This misperception is exemplified in various scenarios, such as:

1. An individual working in a dangerous job who has never personally observed an accident may underestimate the probability of an accident occurring Drakopoulos and Theodossiou (2016); Pandit et al. (2019).

2. Media coverage of events such as plane crashes may cause an overestimation of the probability of a crash, since the public is aware of all crashes but not of all safe trips Wahlberg and Sjoberg (2000); Vasterman et al. (2005); van der Meer et al. (2022).

3. The popularity of purchasing lottery tickets may be explainable in terms of people's inability to comprehend the true probability of winning, influenced instead by news accounts of 'real' people who win multi-million dollar prizes (Rogers (1998); Wheeler and Wheeler (2007); BetterUp (2022)).

Therefore, in this work, we mathematically define spatial black swan events by *subjective probability* in Sections 5 and 7. We further assert that among the four rational axioms, independence and continuity should be modified with a subjective probability function $w : [0,1] \to [0,1]$ to fully address human-decision making, which will be elaborated as *human MDP* in Subsection 5.2. The function w could be interpreted as humans' *implicit probability belief* (Simon (1993)). We specify the property of function w at Assumption 3 in Section 5 inspired by some case studies of human behavior from Kahneman and Tversky (2013).

1. *Subjective* independence: If A is preferred to B and an event probability $p \in [0,1]$, then humans perceive that $w(p)A + (1 - w(p))C$ should be preferred to $w(p)B + (1 - w(p))C$, while in *real-world*, $pA + (1-p)C$ should be preferred to $pB + (1-p)C$.

2. *Subjective* continuity: If A is preferred to B and B is preferred to C, there exists an event probability $p \in [0,1]$ such that B is considered equally preferable to $pA+(1-p)C$ in the real-world. However, humans perceive that there exists $B' \neq B$ that is equally preferable to $w(p)A + (1 - w(p))C$.

We elaborate on how subjective independence and subjective continuity could yield different optimal policies in the human-perceived world (which will be termed as Human MDP) and in the real-world (which will be termed as Universe MDP), as demonstrated through Examples 2 and 3, respectively. In particular, we further describe the emergence of spatial black swan events due to subjective probability in Example 2. First, Example 2 shows an example of modified 3^{rd} rational axiom.

Example 2 (Spatial black swan due to underestimation of low event probability)**.** *Suppose that humans can access the true utility function. The utilities of the outcomes* $a, b, c \in \mathcal{O}$ *are given as* $u(a) = -1, u(b) = -1000, u(c) = 1000,$ *where* b *is an extremely risky outcome.*

Suppose that the choices A, B, and C return outcomes a, b, and c with 100% certainty, respectively. Then, humans choose two compound choices L and M where L returns 0.001% of A and 99.999% of C, and compound choice M returns 0.001% of B and 99.999% of C. Through several trials, due to the extreme rarity of occurrence of choices A or B, regardless of choosing L or M, humans perceive that L and M return the same utility value of 1000, which can be interpreted as $w(0.001\%) = 0\%$ and $w(99.999\%) = 100\%$, and decide their optimal policy is random sampling among choices $\{L, M\}$. However, over time, humans have a probability of encountering choice B with a probability of 0.0005%, which is a black swan.

The modified 4^{th} rational axiom, irrational continuity, does not intuitively have a direct relationship with the emergence of black swans, but it provides how preference could be reversed and leads to an irrational decision.

Example 3 (How subjective probability leads to irrational risky decision)**.** *Continuing from Example 2, suppose that the compound choice N returns B with 50.05% and C with 49.95%. Then, in the real world, choice N and choice A return the same utility as $u(b) \cdot p(b) + u(c) \cdot p(c) = -1000 \cdot 0.5005 + 1000 \cdot 0.4995 = -1$. However, if humans perceive that choice N returns B with 50.01% and C with 49.99%, they consider choice N to return the value $u(b) \cdot p(b) + u(c) \cdot p(c) = -1000 \cdot 0.5001 + 1000 \cdot 0.4999 = -0.2$, and thus determine their optimal policy as choosing the choice N.*

In reality, choosing between choice N and A does not matter since both provide the same utility value, but choosing choice N is actually more risky than choice A since choice N contains the high-risk outcome B.

Note that the result B in Example 3 is not a black swan event, since it is not a rare event. However, through these examples 2 and 3, it is clear that human misperception (subjective probabilities) could yield different optimal policies in the real-world.

4 Cumulative prospect theorem

In this section, we provide a preliminary overview of Cumulative Prospect Theory (CPT), which offers a framework for understanding human decision making under risk and uncertainty. We utilize the principles of CPT to elaborate the concept of misperception and incorporate it into the MDP (Section 5) and further to define spatial black swan events (Section 7).

For a random variable X, let p_i where $i = 1, \dots, K$ denote the indices for the probability of incurring a gain or loss x_i for each $i = 1, \dots, K$. Given a utility function u and a weighting function w, the *Prospect Theory (PT)* value is defined as $V(X) = \sum_{i=1}^{K} u(x_i) w(p_i)$, and the *Cumulative Prospect Theory (CPT)* value is defined as $V(X) = \sum_{i=1}^{K} u(x_i) \left(w(\sum_{j=1}^{i} p_j) - w(\sum_{j=1}^{i-1} p_j) \right)$. Contrary to expected utility theory, which models decisions that perfectly rational agents would make (Rabin (2013)), i.e. $V(x) = \sum_{i=1}^{K} x_i p_i$, PT seeks to describe the actual behavior of humans, attempting to encompass their irrational decision-making processes. Specifically, PT introduces the concept of *probability distortion*, where individuals overestimate the likelihood of rare events and underestimate the likelihood of moderate to highly probable events (Figure 1b). *Value distortion* refers to the way individuals assess gains and losses (x-axis of Figure 1a), often valuing losses more heavily than equivalent gains, which reflects a behavior known as loss aversion (Figure 1a) Kahneman and Tversky (2013); Fennema and Wakker (1997). How PT explains human decision-making is well-described in the following example.

Example 4 (Insurance policies)**.** *Consider an example where the probability of an insured risk is* 1%*, the potential loss is* $1,000$*, and the insurance premium is* 15*. According to CPT, most would opt to pay the* 15 *premium to avoid the larger loss.*

Example 4 illustrates how a seemingly straightforward decision can be analyzed as a sequential decision-making problem within a two-step Markov Decision Process framework, where $\mathcal{S} = \{s_{base}, s_{premium}, s_{risk}\}$ and $\mathcal{A} = \{a_p, a_{np}\}$. The states $s_{base}, s_{premium}$, and s_{risk} represent receiving a loss of $0, -15$, and -1000, respectively. The actions a_p and a_{np} denote paying and not paying the premium, respectively. At time $t = 0$, humans make a choice between a_p and a_{np} based on a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ stating at initial state $s_0 = s_{base}$. Choosing a_p results in a guaranteed transition to state $s_1 = s_{preimium}$. Otherwise, choosing a_{np} potentially leads to state s_{base} with a reward of 0 with a 0.99 probability, or to state s_{risk} with a reward of -1000 with a 0.01 probability. According to the expected utility theorem, which assumes rationality, the estimated value of choosing a_p is calculated as $V(s_{base}) = 1 \times r(s_{premium}) = 1 \times (-15) = -15$, and the value of choosing a_{np} as $V(s_{base}) = 0.99 \times r(s_{base}) + 0.01 \times r(s_{risk}) = 0.99 \times 0 + 0.01 \times (-1000) = -10$. Rationality would lead humans to prefer a_{np} since its expected cost is lower than that of a_p, i.e. $a_{np} = \arg\max_{a \in \mathcal{A}} V(s_{base})$. However, this choice is counterintuitive and often does not align with real-world human decision making.

CPT uses a similar measure as PT, except that the w is a function of cumulative probabilities. The concept involves using an S-shaped utility function, which adheres to the diminishing sensitivity property. If we set the weighting function w or utility function u to be the identity function, then we retrieve the classical expected utility mode (Rabin (2013)).

5 Agent-Environment intersects as perception

Thus far, we have informally introduced the black swan hypothesis (Hypothesis 1) and spatial black swan definition (Definition 1) in Section 1 and elaborated on its necessity (Subsection 3.1) and supporting evidence (Subsection 3.2) in Section 3. Then, we introduce one existing work to concretize misperception, CPT, in Section 4. In Section 5, we utilize CPT to elaborate the Hypothesis 1 by introducing universe, human, and human-estimation MDP.

5.1 Misperception is information loss

Based on Hypothesis 1, this prompts us to investigate the concept of *misperception*. Initially, we must clearly define what constitutes *perception*. In *The Quest for a Common Model of the Intelligent Decision Maker*, Sutton defines perception as one of four principal components of agents, stating: "The perception component processes the stream of observations and actions to produce the subjective state, a summary of the agent-world interaction so far that is useful for selecting action (the reactive policy), for predicting future reward (the value function), and for predicting future subjective states (the transition model)" Sutton (2022). This definition leads us to consider misperception as the *information loss* occurring when processing observations into the subjective state, such that the reward and transition model are not equivalent to those from the environment. The interpretation of misperception as *information loss during processing* is somewhat ambiguous, depending on how the boundary between the agent and the environment is defined. The concept of a boundary between the agent and environment was first proposed by Turing as a 'skin of an onion' Turing (2009), and later, Jiang (2019) suggested that algorithms are not boundary-invariant.

Therefore, we propose a new agent-environment framework that incorporates the notion that *misperception is the information loss from an agent's processing*. This framework positions perception at the intersection between the agent and the environment. We provide a detailed description of our agent-environment framework in Figure 2 and further elaborate in the following subsection.

5.2 Universe, Human, and Human-Estimation MDP

Restricting the Markov Decision Process introduced in Section 3 in a stationary environment, we consider a single episode finite horizon stationary Markov Decision Process (MDP) denoted as $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma, T \rangle$, where $\mathcal{S}$ represents the state space, $\mathcal{A}$ denotes the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability function, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function, γ is the discount factor, and $T \geq 0$ is a Horizon. We define $\mathcal{M}$ as the **universe MDP** and operate under the assumption that the abstraction from the universe to the model $\mathcal{M}$ is lossless, preserving all relevant information. Given a policy π, the agent collects data $\{s_0, a_0, r_0, s_1, \cdots\}$ as it interacts with the environment at discrete time steps t. The process starts from a fixed initial state s_0, and we define the value function as follows:

$$V(s_0) \coloneqq \mathbb{E}\left[\sum_{t=0}^{T-1} R(s_t, a_t) \mid \pi, P\right] \tag{1}$$

To fully leverage Hypothesis 1, and 'perception as information loss from agent's processing' from subsection 5.1, we define the **human MDP** $\mathcal{M}^\dagger = \langle \mathcal{S}^\dagger, \mathcal{A}^\dagger, P^\dagger, R^\dagger, \gamma \rangle$ where the agent experiences distortions in cumulative distribution of normalized visitation probability $P^\pi(s,a) \coloneqq \frac{1-\gamma^T}{1-\gamma} \sum_{t=0}^{T-1} \gamma^t \mathbb{P}((s_t, a_t) = (s,a) | s_0, \pi, P)$ and reward function R by functions

w^+, w^-, u^+, w^- (function characteristics are explained in Assumptions 2 and 3). Note that $\mathbb{P}(s, a|s_0, \pi, P)$ is the probability of visiting (s, a) at time t by policy π starting from s_0.

One internal assumption in the human MDP is that its state and action spaces are the same as those of the universe MDP, i.e. $\mathcal{S}^\dagger = \mathcal{S}$ and $\mathcal{A}^\dagger = \mathcal{A}$. This assumption is significant, as insufficient exploration might result in a limited understanding of the entire state space by the human (agent), and the discrepancy between the human MDP and the universe MDP could be considerable, especially in a large discrete state and action space. However, we can augment the human MDP to match with universe MDP by using the following method.

Remark 2. *If humans cannot perceive some state $s \in \mathcal{S}$, we can augment the human state space with $\left(S^\dagger\right)' = S^\dagger \cup \{s\}$, define $R^\dagger(s, a) = R(s, a)$, and set $P^\dagger(s' \mid s, a) = P(s' \mid s, a)$ while $P(s \mid s', a) = 0$ for all $s \in \mathcal{S}^\dagger, a \in \mathcal{A}^\dagger$. Thus, the new state s does not affect decision-making in a human MDP since the probability that the trajectory visits the state s is zero.*

From now on, we denote human MDP $\mathcal{M}^\dagger = \langle \mathcal{S}, \mathcal{A}, P^\dagger, R^\dagger, \gamma \rangle$ For a finite state and action space, we can define the order statistics of $P^\pi(s, a)$ for $\forall (s, a)$ and define the cumulative distribution. For short notation, we use $\int P^\pi(s, a)$ as its cumulative distribution. These distortions are defined by the relationships:

$$\int P^{\dagger,\pi}(s, a) = \begin{cases} w^+(\int P^\pi(s, a)) \text{ if } R(s, a) \geq 0 \\ w^-(\int P^\pi(s, a)) \text{ if } R(s, a) < 0 \end{cases} \quad , \forall (s, a) \in \mathcal{S} \times \mathcal{A} \tag{2}$$

$$R^\dagger(s, a) = \begin{cases} u^+(R(s, a)) \text{ if } R(s, a) \geq 0 \\ u^-(R(s, a)) \text{ if } R(s, a) < 0 \end{cases} \quad , \forall (s, a) \in \mathcal{S} \times \mathcal{A} \tag{3}$$

We also define the value function of human MDP as follows.

$$V^\dagger(s_0) \coloneqq \mathbb{E}\left[\sum_{t=0}^{T-1} R^\dagger(s_t, a_t) \mid \pi, P^\dagger\right] \tag{4}$$

In Equation (2), note that misperceptions are applied to the visitation probability P^π rather than the transition probability P itself. This approach is more reasonable, as humans distort the probability of *events*, and we have defined the dimension of spatial black swan *events* as (s, a) (Definition 2). We explore case studies where humans misperceive P itself in Section 6. However, we address the following definition and lemma to bridge the gap between these two types of misperceptions, implying that theoretical analyses on one are also interchangeably applicable to the other.

Definition 5 (Biased and perceived reward and visitation)**.** *For given constant $\kappa_r, \kappa_d \in \mathbb{R}_+$, if $\max_{(s,a)} |R^\dagger(s, a) - \widehat{R}^\dagger(s, a)| \leq \kappa_r$ holds, then $\widehat{R}^\dagger(s, a)$ is κ_r-biased reward. Also, if $\max_{(s,a)} |P^{\pi,\dagger}(s, a) - \widehat{P}^{\pi,\dagger}(s, a)| \leq \kappa_d$ holds, then $\widehat{P}^{\pi,\dagger}(s, a)$ is κ_d-biased visitation probability.*

Also, for given constant $\epsilon_r, \epsilon_d \in \mathbb{R}_+$, if $\max_{(s,a)} |R(s,a) - R^\dagger(s,a)| < \epsilon_r$ holds, then $R^\dagger(s,a)$ is ϵ_r-perceived reward. Also, if $\max_{(s,a)} |P^\pi(s,a) - P^{\pi,\dagger}(s,a)| < \epsilon_d$ holds, then $P^{\dagger,\pi}(s,a)$ is ϵ_d-perceived visitation probability.

Lemma 1. *If $\max_{s,a} \|P(\cdot|s,a) - P^\dagger(\cdot|s,a)\|_1 \leq \frac{(1-\gamma)^2}{\gamma}\epsilon_d$ where $\epsilon_d > 0$, then the agent can guarantee ϵ_d-perceived visitation probability (See Definition 5 for the definition of ϵ_d-perceived).*

Then, the human MDP $\mathcal{M}^\dagger$ is fully characterized by the utility functions u^+, u^- and the weight functions w^+, w^- derived from the universe MDP $\mathcal{M}$. We proceed to delineate the properties of these utility and weight functions as described by (Kahneman and Tversky (2013)), under the following assumptions:

Assumption 2 (**Utility function**)**.** *A function $u^+ : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a non-decreasing concave function that satisfies $\lim_{h\to 0+}(u^+)'(h) \leq 1$. A function $u^- : \mathbb{R}_{\leq 0} \to \mathbb{R}_{\leq 0}$ is a nondecreasing convex function that satisfies $\lim_{h\to 0-}(u^-)'(h) > 1$.*

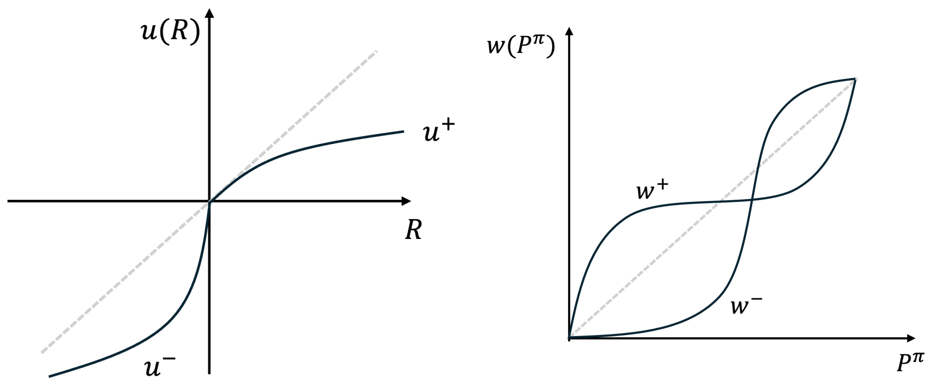
Assumption 3 (**Weight function**)**.** *Let $w^+, w^- : [0,1] \to [0,1]$ be a differentiable function, then those satisfy*

1. *$w^+(0) = 0, w^+(1) = 1$ and $w^-(0) = 0, w^-(1) = 1$.*
2. *There exists $a, b \in (0,1)$ such that $w^+(a) = a, w^-(b) = b$.*
3. *$(w^+)'(x)$ is monotonically decreasing on $x \in [0,a)$ and monotonically increasing on $x \in (a,1]$. Meanwhile, $(w^-)'(x)$ is monotonically increasing on $x \in [0,b)$ and monotonically decreasing on $x \in (b,1]$.*

Figure 1 illustrates the geometric properties of the utility and weight functions, as defined under Assumptions 2 and 3. Note that the assumptions concerning the weight functions (Assumption 3) stipulate that w^+ and w^- are Lipschitz continuous, with constants $(w^+)'(a)$ and $(w^-)'(b)$, as derived from the mean value theorem. We denote these constants as L^+ and L^-, respectively.

Based on the perceptions of the agent, it executes a trajectory using the policy π within the human MDP, $\mathcal{M}^\dagger$, and *estimates* the perceived reward $R^\dagger(s,a)$ and visitation probability $P^{\dagger,\pi}(s,a)$ as $\widehat{R}^\dagger(s,a)$ and $\widehat{P}^{\dagger,\pi}(s,a)$, respectively, from its trajectory. The estimation error bounds are influenced by the complexity of the MDP, including tabular MDPs, linear MDPs, and low-rank MDPs. This configuration is defined as the **human-estimation MDP** $\widehat{\mathcal{M}}^\dagger = \langle \mathcal{S}, \mathcal{A}, \widehat{P}^\dagger, \widehat{R}^\dagger, \gamma \rangle$. In a similar way of value function definition in universe MDP as follows.

$$\widehat{V}^\dagger(s_0) := \mathbb{E}\left[\sum_{t=0}^{T-1} \widehat{R}^\dagger(s_t, a_t) \mid \pi, \widehat{P}^\dagger\right].$$



(a) Utility function u (b) Weight function w

Figure 1: Utility and weight functions. A line with a gray color denotes $y = x$.

$$\overbrace{\mathcal{M} \xLeftrightarrow[\epsilon_r,\epsilon_d]{\text{perception}} \mathcal{M}^\dagger}^{\text{Environment}} \underbrace{\xLeftrightarrow[\kappa_r,\kappa_d]{\text{estimation}} \widehat{\mathcal{M}}^\dagger}_{\text{Agent}}$$

Figure 2: Agent-environment intersects with perception

Based on the definition of Universe MDP, Human MDP, and Human-Estimation MDP, we define their gap as follows, and its relationship could be described as Figure 2.

Since CPT models irrational decision-making, our work begins by defining modified value functions in an infinite-horizon Markov Decision Process (MDP) inspired by CPT. As a preliminary step, we start by defining the CPT-value function in a discrete state and action space.

First, in a discrete state and action space, the value function (Equation (1)) could be expressed as an inner product of reward function R and normalized occupancy measure P^π as follows,

$$V(s_0) = \frac{1-\gamma^H}{1-\gamma} \sum_{s,a \in \mathcal{S}\times\mathcal{A}} R(s,a) P^\pi(s,a) \tag{5}$$

Based on Definition 5, the CPT distorts the reward and its visitation probability as follows,

$$V^\dagger(s_0) = \frac{1-\gamma^H}{1-\gamma} \sum_{s,a \in \mathcal{S}\times\mathcal{A}} u(R(s,a)) \frac{d}{dsda} w\left(\int P^\pi(s,a) \right). \tag{6}$$

where † denotes the value was distorted due to misperception. As one property of CPT is that human perception exhibits distinct distortions of events based on whether the associated rewards are positive or negative, we divide the functions $u(R(s,a))$ and $w(P^\pi(s,a))$ into $u^-(R(s,a)), w^-(P^\pi(s,a))$ where $R(s,a) < 0$, and $u^+(R(s,a)), w^+(P^\pi(s,a))$ where $R(s,a) \geq 0$. Assume that the rewards from all state-action pairs $R(s,a)$ are ordered as $R_{[1]} \leq \cdots \leq R_{[l]} \leq 0 \leq R_{[l+1]} \leq \cdots \leq R_{[|\mathcal{S}||\mathcal{A}|]}$, and the visitation probability as $P^\pi_{[1]} \leq P^\pi_{[2]} \leq \cdots \leq P^\pi_{[|\mathcal{S}||\mathcal{A}|]}$. Then, CPT-value function is defined as follows:

$$\begin{aligned} V^\dagger(s_0) &= \sum_{i=1}^{|\mathcal{S}||\mathcal{A}|} u(R_{[i]}) \left(w\left(\sum_{j=1}^{i} P^\pi_{[j]} \right) - w\left(\sum_{j=1}^{i-1} P^\pi_{[j]} \right) \right) \\ &= \sum_{i=1}^{l} u^-(R_{[i]}) \left(w^-\left(\sum_{j=1}^{i} P^\pi_{[j]} \right) - w^-\left(\sum_{j=1}^{i-1} P^\pi_{[j]} \right) \right) \\ &\quad + \sum_{i=l+1}^{|\mathcal{S}||\mathcal{A}|} u^+(R_{[i]}) \left(w^+\left(\sum_{j=i}^{|\mathcal{S}||\mathcal{A}|} P^\pi_{[j]} \right) - w^+\left(\sum_{j=i+1}^{|\mathcal{S}||\mathcal{A}|} P^\pi_{[j]} \right) \right) \end{aligned} \tag{7}$$

If we define the reward as the random variable X, then we can regard its instance as $R_{[i]}$ and its probability as $P^\pi_{[i]}$ where the probability is dependent on the policy π. Suppose that reward function $R : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is one to one function. Then the probability $R^{-1} \circ P^\pi : \mathbb{R} \to [0,1]$ denotes the probability of reward and we denote it as $\mathbb{P}_r$. Then, for a reward random variable $\mathcal{R} \sim \mathbb{P}_r$, expanding the how CPT- applied value function look like in Equation (4), we can define the value function based on continuous state and actions space as follows.

$$V^\dagger(s_0) = \int_0^\infty w^+\left(\mathbb{P}_r(u^+(\mathcal{R}) > r)\right) dr - \int_0^\infty w^-\left(\mathbb{P}_r(u^-(\mathcal{R}) > r)\right) dr \tag{8}$$

We use the fact that for real-value function g, it holds that $\mathbb{E}[g(\mathcal{R})] = \int_0^\infty \Pr(g(\mathcal{R}) > r))dr$. In this sense, we define the black swan event in the continuous state and action space.

5.3 Problem setting

Based on three different MDPs, $\mathcal{M}, \mathcal{M}^\dagger, \widehat{\mathcal{M}}^\dagger$, we consider the following problem setting. The agent rolls out in a single episode with a finite horizon T. If the agent has an unbiased perception, then the agent collects a trajectory $\{s_0, a_0, r_0, s_1, a_1, \cdots, s_{T-1}, a_{T-1}, s_T\}$. However, the agent perceives $\mathcal{M}$ as $\mathcal{M}^\dagger$ and now observe distorted state and reward as $\{g(s_0), a_0, u(r_0), g(s_1), a_1, \cdots, g(s_{T-1}), a_{T-1}, g(s_T)\}$ where function $g : \mathcal{S} \to \mathcal{S}$ distorts the state. Now, we can claim the following

Lemma 2. *Suppose the $\mathcal{M}$ is given. Then, for any function w that satisfies the Assumption 3, one can always find the function $g : \mathcal{S} \to \mathcal{S}$ that satisfies the following equation.*

$$w(\int P^\pi(s,a)) = \int P^\pi(g(s),a)$$

Note that Lemma 2 allows us to design the problem such that the agent distorts the visitation probability by receiving a distorted state induced by the function g. Lemma 2 justifies how the agent distorts the visitation probability from the trajectory data.

5.4 Utilizing CPT for black swan

We note that existing work on incorporating cumulative prospect theory (CPT) into reinforcement learning, such as (Prashanth et al. (2016); Jie et al. (2018); Danis et al. (2023)), primarily focus on estimating the CPT-based value function and optimizing it to derive an optimal policy. Specifically, (Prashanth et al. (2016); Jie et al. (2018)) demonstrate how to estimate the CPT value function using the Simultaneous Perturbation Stochastic Approximation method and how to compute its gradient for policy optimization algorithms. Additionally, (Shen et al. (2014); Ratliff and Mazumdar (2019)) proposed a novel Q-learning algorithm that applies a utility function to Temporal Difference (TD) errors and demonstrated its convergence. However, these studies (Prashanth et al. (2016); Jie et al. (2018); Danis et al. (2023); Shen et al. (2014); Ratliff and Mazumdar (2019)) do not focus on learning the utility and weight functions, u and w, but rather assume these as simple functions and focus on how to *estimate* these functions.

However, this study aims to elucidate the mechanisms by which black swan events arise from the discrepancies between $\mathcal{M}^\dagger$ and $\mathcal{M}$, despite the agent having perfect estimation, i.e., $\kappa_r = 0, \kappa_p = 0$. As future work, concentrating on devising strategies to *reweight* the functions u_+, u_-, and w to mitigate the divergence between the Human MDP $\mathcal{M}^\dagger$ and the universe MDP $\mathcal{M}$ is suggested as a way to achieve antifragility.

6 Case study: how optimal decision deviates under irrationality

So far, we have informally introduced the black swan hypothesis (Hypothesis 1) and spatial black swans (Definition 1) in Section 1, then elaborated on its necessity (Subsection 3.1) and provided evidence (Subsection 3.2) in Section 3. Subsequently, in Section 4, we introduced the Cumulative Prospect Theorem (CPT) to model human irrationality. In Section 5, we incorporated CPT into the Markov Decision Process (Subsection 5.2), viewed through the lens of existing work on defining perception (Subsection 5.1) by introducing universe, human, and human-estimation MDPs (Figure 2).

Given our main hypothesis that black swan events occur due to human misperception of the real world, we further investigate whether optimal policy also deviates due to misperceptions of *value or probability.* This is critical as overestimating or underestimating all values of probability does not necessarily lead to changes in optimal policy. For example, revisiting Example 4, overestimating or underestimating $\{r(s_{base}), r(s_{premium}), r(s_{risk})\}$ to the same magnitude does not alter the optimal policy for a human to choose a_p. More fundamentally, a crucial question that this paper addresses is how misperception influences the deviation of optimal policy. Therefore, in Section 6, we first present some case studies in MDPs with small complexity to demonstrate how optimal policy varies under misperception.

Before proceeding, we need to establish the core event of *subjective probability.* For example, while agents (humans) might distort transition probabilities and perceive them subjectively at a low level which is an intuitive way to model misperception. It is essential to recall that in the Markov Decision Process, the quality of an event is revealed through rewards defined over specific states and actions. This suggests that it is more reasonable to define the minimal object (event) as the state and action, and conduct modeling as distortion on the visitation probability of state and action rather than on the transition probability itself. Since this approach has not been investigated in existing work, we examine both cases in Section 6 and Section 7. Specifically, in Section 6, we assume distortion of transition probability within a non-stationary Markov decision process and investigate how optimal policy deviates due to misperception of transition probability. Then, in Section 7, we explore the distortion of the visitation probability within a stationary Markov decision process.

6.1 Problem setup for Section 6

In this section, we consider discrete state and action stationary Markov Decision Process. Build upon value function (Equation 4), we define value function and state value function of time t as

$$\begin{aligned} V_t^{\pi,\dagger}(s) &:= \mathbb{E}\left[\sum_{t'=t}^{T-1} \gamma^{T-1-t'} u(R(s_{t'}, a_{t'})) \mid P^\dagger, \pi, s_t = s\right] \\ Q_t^{\pi,\dagger}(s,a) &:= \mathbb{E}\left[\sum_{t'=t}^{T-1} \gamma^{T-1-t'} u(R(s_{t'}, a_{t'})) \mid P^\dagger, \pi, s_t = s, a_t = a\right] \end{aligned} \tag{9}$$

and define the optimal policy as time t as $\pi_t^{\star,\dagger} = \arg\max_\pi V_t^{\pi,\dagger}$. Then the following bellman equation holds,

$$\begin{aligned} V_t^{\pi,\dagger}(s) &= \sum_{a\in\mathcal{A}} \pi(a|s) Q_t^{\pi,\dagger}(s,a) \\ Q_t^{\pi,\dagger}(s,a) &= u(R(s,a)) + \sum_{i\in|\mathcal{S}|} \gamma \left(w\left(\sum_{j=1}^{i} P(s'_{[i]} \mid s,a) \right) - w\left(\sum_{j=1}^{i-1} P(s'_{[i]} \mid s,a) \right) \right) V_{t+1}(s_{[i]}) \end{aligned}$$

where $P(s_{[1]}|s,a) \leq P(s_{[2]}|s,a) \leq \cdots \leq P(s_{[|S|]}|s,a)$ holds. In addition, we assume that $\forall t \in [T-1], \forall s \in \mathcal{S}$ and $\forall a \in \mathcal{A}, R(s_t, a_t) = 0$. We only consider the reward function at the final stage $R(s_T, a_T)$. Although this assumption may appear unconventional, it aligns with standard practices in reinforcement learning, especially when focusing on terminal rewards. Especially, for each trajectory $\tau \in \mathcal{S}'$, where $\tau = (s_0, a_0, \ldots, s_H, a_H)$ represents a T-step decision sequence, set the terminal reward as $r'_T(\tau) := \sum_{t\in[T]} r_t(s_t, a_t)$, and $r'_t(\tau) = 0$ for all $t \in [T-1]$

6.2 Case 1. Contextual bandit ($T = 1$)

We begin with a simple case where the decision horizon is $T = 1$, commonly referred to as a contextual bandit (Lattimore and Szepesvári (2020)). Surprisingly, in this setting, the human optimal policy coincides with the real-world optimal policy. This is somewhat counterintuitive, as several significant examples (Examples 2, 3, and 4) suggest that human decision-making often exhibits irrationality.

Theorem 1 (One-step Human Optimal Policy)**.** *If $T = 1$, then the optimal policy from the universe MDP aligns with the optimal policy of the human MDP, i.e. $\pi^\star = \pi^{\star,\dagger}$.*

An important insight from Theorem 1 is that when decisions are not sequential, the typical distortions in human perception do not affect the alignment with the optimal policy of the real-world, suggesting that human irrationality is less influential in a single-step decision-making setting. This is further explained in the following remark.

Remark 3. *Continuing from Example 4, the order of perceived rewards does not change as $u^-(r(s_{loss})) < u^-(r(s_{premium})) < u^-(r(s_{base}))$ since u^- is a nondecreasing convex function.*

6.3 Case 2. $|\mathcal{S}| = 2$ when $T > 1$

Now, let us consider the simplest case where $H > 1$ where $|\mathcal{S}| = 2$. Surprisingly, we find results similar to those presented in Section

Theorem 2 (Multi-step human optimal policy)**.** *If $|\mathcal{S}| = 2$, then the optimal policy from the universe MDP also aligns with the optimal policy of the human MDP, that is, $\pi_t^{\star,\dagger} = \pi_t^\star$ for $\forall t \in [T]$.*

The proof of Theorem 2 is based mainly on the assumption that $|\mathcal{S}| = 2$, notably using the property that $w(1) = 1$. However, this technique cannot be applied directly if $|\mathcal{S}| \geq 3$. For example, in Example 4, where $|\mathcal{S}| = 3$, it is demonstrated that human decision making results in a suboptimal policy in the real-world. This outcome may seem counterintuitive. However, a heuristic analysis suggests that having only two states in the state space indicates that the actions do not introduce a varied randomness. Essentially, if randomness is introduced by any action, it would likely affect both states s_0 and s_1 if the action provides a nondeterministic next state. Thus, every action provides the same next-state set, which makes a mere comparison between two states. This observation implies that in scenarios with a smaller state space, being irrational (believing on false belief) does not affect to deviate from optimal policy. We can also interpret this result as if $|\mathcal{S}|$ is small, sequential decision-making problems are *easy* so that humans can always provide the optimal action in the real world.

6.4 Case 3: $|S| = 3$ with unbiased reward perception

We consider the hypothetical scenario where $u(r) = r$, indicating that humans have an unbiased perception of their reward.

Theorem 3 (Two step optimal decision when $|\mathcal{S}| = 3$)**.** *Given any state space $\mathcal{S}$ where $|\mathcal{S}| = 3$ and a decision horizon $T = 2$, there exists a transition probability P and reward R of the real world such that the optimal policy of human MDP differs from that of universe MDP.*

Theorems 1, 2, and 3 illustrate that the discrepancy between human and optimal decision making increases as the complexity of the environment increases ($\mathcal{S}$) or as the decision sequence (T) lengthens for any w function.

7 A definition of spatial black swans

Now, we introduce the definition of black swan inspired by CPT.

7.1 Black swan hypothesis

First, based on the newly proposed human model $\mathcal{M}^\dagger$, we concretize the informal hypothesis 1 as follows.

Hypothesis 2 (Spatial black swan origin)**.** *For a given universe MDP $\mathcal{M}$, the agent initially perceives the universe as $\mathcal{M}^{\dagger}$. Then it estimates $\mathcal{M}^{\dagger}$ as $\widehat{\mathcal{M}}^{\dagger}$. Subsequently, the agent rolls out trajectories in $\widehat{\mathcal{M}}^{\dagger}$ to learn its optimal policy, $\widehat{\pi}^{\dagger,\star}$. Then, all spatial black swan events (s,a) are functions of $\mathcal{M}$, $\mathcal{M}^{\dagger}$, and $\pi^{\dagger,\star}$.*

We define all spatial black swan events as a set $\mathcal{B}(\mathcal{M},\mathcal{M}^{\dagger},\pi^{\dagger,\star})$. It is important to note that $\mathcal{B}$ is a function of the agent's perception $\mathcal{M}^{\dagger}$, rather than its estimation $\widehat{\mathcal{M}}^{\dagger}$, since estimation is what the agent can improve over time. Specifically, all state and action events from $\mathcal{B}(\mathcal{M},\widehat{\mathcal{M}}^{\dagger},\widehat{\pi}^{\dagger,\star})$ can be regarded as broad 'risks' that encompass spatial black swan events. Among these risks, there are some that the agent can 'be averse to' and others that it 'cannot be averse to' in the process of improving model estimation $\widehat{\mathcal{M}}^{\dagger}$ to the human model $\mathcal{M}^{\dagger}$. The crux of spatial black swan events lies in their unexpectedness, which occurs even when the agent has prepared the optimal policy with zero estimation error. The model estimation process discussed in this paper aligns with existing work on imperfect model-based reinforcement learning (Jiang (2018); Gheshlaghi Azar et al. (2013); Fard and Pineau (2010); Agarwal et al. (2020); Lecarpentier and Rachelson (2019)). For future work, we leave room to define the 'risks' due to irrational human behavior that goes beyond spatial black swans. Specifically, events from $\mathcal{B}(\mathcal{M},\widehat{\mathcal{M}}^{\dagger},\widehat{\pi}^{\dagger,\star}) \setminus \mathcal{B}(\mathcal{M},\mathcal{M}^{\dagger},\pi^{\dagger,\star})$ are risk events that can be avoided as time goes by, and events from $\mathcal{B}(\mathcal{M},\mathcal{M}^{\dagger},\pi^{\dagger,\star})$ are non-avoidable risk, i.e. spatial black swans. This conceptual framework can be outlined by the following equation:

$$\lim_{\kappa_r,\kappa_d\to 0} \underbrace{\mathcal{B}(\mathcal{M},\widehat{\mathcal{M}}^{\dagger},\widehat{\pi}^{\dagger,\star})}_{\text{Risks}} = \underbrace{\mathcal{B}(\mathcal{M},\mathcal{M}^{\dagger},\pi^{\dagger,\star})}_{\text{Spatial black swans}}. \tag{10}$$

Equation (10) is supported by Definition 5, where $\kappa_r,\kappa_d \to 0$ leads to $|R(s,a)-\widehat{R}^{\dagger}(s,a)| \to 0$ and $|P^{\pi}(s,a) - \widehat{P}^{\pi,\dagger}(s,a)| \to 0$. This implies that $\widehat{\mathcal{M}}^{\dagger}$ converges to $\mathcal{M}^{\dagger}$ by the definitions of the Human MDP and the Human-Estimation MDP, and consequently, $\widehat{\pi}^{\dagger,\star}$ converges to $\pi^{\dagger,\star}$.

7.2 A definition of spatial black swans

As a preliminary step, we define spatial black swan events within a discrete state and action space. Note that we continue the problem setting that is discussed in deriving Equation (7) in Subsection 5.2.

Definition 6 (**Spatial black swan - discrete state and action space**)**.** *Among the indices of order statistics, if index $i < l$ meets the following criteria:*

1. $R_{[i]} - u^{-}(R_{[i]}) < -C_{bs}$.
2. $w^{-}\left(\sum_{j=1}^{i} P^{\pi}_{[j]}\right) = w^{-}\left(\sum_{j=1}^{i-1} P^{\pi}_{[j]}\right)$*, yet* $0 < P^{\pi}_{[i]} < \epsilon_{bs}$.

where $C_{bs} \gg 0$ and $\epsilon_{bs} > 0$ is given a constant, we define i as a spatial black swan event.

The 1st property of Definition 6 identifies a 'high-risk event' through the function gap between the universe reward R from $\mathcal{M}$ and the perceived reward $u^-(R)$ from $\mathcal{M}^\dagger$. Specifically, if the agent perceives R in a pessimistic way such that $R \ll u^-(R) < 0$, this is considered a high-risk event resulting from misperception. The 2nd property of Definition 6 also characterizes a 'rare event' through a lens of misperception. Technically, it describes that a spatial black swan event feasibly occurs in the universe ($0 < P^\pi_{[i]}$), but the agent perceives it as nonoccurring ($w^-\left(\sum_{j=1}^{i} P^\pi_{[j]}\right) = w^-\left(\sum_{j=1}^{i-1} P^\pi_{[j]}\right)$), i.e. an infeasible event.

In addition, we denote two parameters, C_{bs} and ϵ_{bs}, to represent the extent of distortion in the reward and the cumulative probability of visitation. Intuitively, the magnitude of C_{bs} and ϵ_{bs} is related to the extent of the misperception gap between $\mathcal{M}$ and $\mathcal{M}^\dagger$,i.e. ϵ_r, ϵ_p. We elaborate on this conjecture in Theorem 4 of Section 8. We now extend the definition of black swan events from discrete state and action spaces to continuous spaces as follows.

Definition 7 (**Spatial black swan**). *Given w^-, u^-, if the state-action pair (s,a) satisfies the following conditions:*

1. $R(s,a) - u^-(R(s,a)) < -C_{bs}$.
2. $\left.\frac{d(w^-(F(r)))}{dr}\right|_{r=R(s,a)} = 0$ *but* $0 < \left.\frac{dF(r)}{dr}\right|_{r=R(s,a)} < \epsilon_{bs}$.

where $F(r) := \int_{-\infty}^{r} d\mathbb{P}_r$ represents the cumulative distribution function of $\mathbb{P}_r$, then we define that state and action pair as a black swan event.

If the weight function $w^-(x)$ is differentiable, then the 2nd property of spatial black swan events can be further elaborated as follows: *The derivative* $\left.\frac{dw^-(x)}{dx}\right|_{x=F(R(s,a))} \cdot \mathbb{P}_r(r = R(s,a)) = 0$*, yet the probability density distribution remains non-zero and bounded, specifically* $0 < \mathbb{P}_r(r = R(s,a)) < \epsilon_{bs}$. Consequently, this implies $\left.\frac{dw^-(x)}{dx}\right|_{x=F(R(s,a))} = 0$. This refined second property will be utilized to extend the definition of spatial black swan events to Proposition 3 in Section 8, facilitating further analysis (see Figure 3b).

Also, for all functions w^-, u^- that prevent the existence of spatial black swan events, i.e., $\mathcal{B} = \varnothing$, we refer to such environments as safe perception, denoted by $w^-_\star, u^-_\star$. In contrast, environments characterized by w^-, u^- that inherently encompass spatial black swan events, $\mathcal{B} \neq \varnothing$, are termed risk perception environments. Specifically, if an agent perceives the world through $w^-_\star, u^-_\star$, then spatial black swan events are absent; however, it is crucial to recognize that $w^-_\star, u^-_\star$ are not unique functions (see Figure 3b).

8 Theoretical analysis of spatial black swans

Thus far, we have elaborated on the definition of spatial black swan events (Definition 7) from its informal definition (Definition 1) through the lens of our black swan hypothesis

(Hypothesis 2). In this section, we provide theoretical analysis that quantifies the impact of spatial black swan events. Specifically, we demonstrate how the existence of spatial black swans establishes a lower bound on policy performance (Theorem 4), and assesses the probability of encountering spatial black swan events over time (Theorem 5). It is important to note that Theorem 5 serves as a critical precursor to characterizing the third property of the informal proposition (Proposition 2), since the timing of learning to improve agent's perception is triggered by the statement 'after a black swan is first observed'. Subsequently, we posit that perception improvement learning exhibits antifragile behavior in agents.

8.1 Problem setting

As a preliminary step, we first establish the agent learning setting and provide some assumptions for the theoretical analysis. We utilize the problem setting elaborated in Subsection 5.3. For the theoretical analysis, we assume the following assumptions:

Assumption 4 (**Bounded reward**). *The R is bounded as $R \in [-R_{\max}, R_{\max}]$ where $R_{\max} > 0$*

Assumption 5 (**Relatively strong convexity**). *With Assumption 2, $u^-_\star(r) \le u^-(r)$ holds for $r < 0$.*

Before starting the theoretical analysis, the formal definition of spatial black swan events based on state and action pairs (Definition 7) imposes restrictions on further analysis due to the *openness* of $(s,a) \in \mathcal{B}$. Specifically, Definition 7 regarding the support of $\mathcal{S} \times \mathcal{A}$ does not ensure the existence of a closed subset $\mathcal{C} \subseteq \mathcal{S} \times \mathcal{A}$ such that $\forall (s,a) \in \mathcal{C} \implies (s,a) \in \mathcal{B}$. Therefore, we propose modifications to the support of the spatial black swan event definition concerning the reward. To this end, we further assume that the reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is bijective. Consequently, we introduce the following proposition of a spatial black swan event, an alternative to Definition 7, which is defined over reward values in closed intervals.

Proposition 3 (**Spatial black swan**). *Under Assumptions 1,2, and 3, if w^-, u^- satisfies spatial black swan definition (Definition 7), then $\forall r \in [-R_{\max}, -R_{bs}]$ satisfies*

1. $r - u^-(r) < -C_{bs}$
2. $w^-(F(r)) = 0$ *but* $0 < F(r) < \epsilon_{bs}$

where $R_{bs} - u^-(R_{bs}) = -C_{bs}$ holds.

The Proposition 3 is well elaborated in Figure 3. Proposition 3 facilitates the Definition 7 within the closed set $[-R_{\max}, -R_{bs}]$. This approach to defining spatial black swan events based on the support of the reward is more intuitive to understand the black swan origin hypothesis (Hypothesis 2). That is, Proposition 3 shows how decreasing the misperception gap ($\mathcal{M}^\dagger \to \mathcal{M}$) relates to reducing $|\mathcal{B}|$. This is because a reduction in the misperception gap, where $u^-(r) \to u_\star^-(r)$, leads to a decrease in the range of spatial black swan rewards, moving from $-R_{bs}$ to $-R_{\max}$, thus reducing the frequency of spatial black swan occurrences (see Figure 3).

Within the above problem setting, the agent's goal is to estimate the value function under safe perception $\mathcal{M}$ as follows:

$$V(s_0) = \int_0^\infty w^+ \left(\mathbb{P}_r(u^+(X) > r)\right) dr - \int_0^\infty w_\star^- \left(\mathbb{P}_r(u_\star^-(X) > r)\right) dr \tag{11}$$

However, the agent possesses its own perceptions $\mathcal{M}^\dagger$, for which we assume the risk perception is represented as:

$$V^\dagger(s_0) = \int_0^\infty w^+ \left(\mathbb{P}_r(u^+(X) > r)\right) dr - \int_0^\infty w^- \left(\mathbb{P}_r(u^-(X) > r)\right) dr \tag{12}$$

As time goes by, the agent's goal is learning the weight functions and utility functions such as $w^- \to w_\star^-$ and $u^- \to u_\star^-$. Then, by the single trajectory data up to time t, i.e. $\{s_i, a_i, u(r_i), s_{i+1}\}_{i=0}^t$ where the reward value itself and its sampling distribution are distorted due to the functions u and w, respectively. Let $r_i, i = 1, .., t$ denote n samples of the reward random variable X. We define the empirical distribution function for $u^+(X)$ and $u^-(X)$ as follows

$$\hat{F}_t^+(r) = \frac{1}{t}\sum_{i=1}^n \mathbf{1}_{(u^+(r_i) \le r)}, \quad \text{and} \quad \hat{F}_t^-(r) = \frac{1}{t}\sum_{i=1}^n \mathbf{1}_{(u^-(r_i) \le r)}$$

. using the EDFs, the CPT value up to time t can be estimated as follows,

$$\widehat{V}_t^\dagger(s_0) = \int_0^\infty w^+ \left(1 - \hat{F}_t^+(r)\right) dr - \int_0^\infty w^- \left(1 - \hat{F}_t^-(r)\right) dr \tag{13}$$

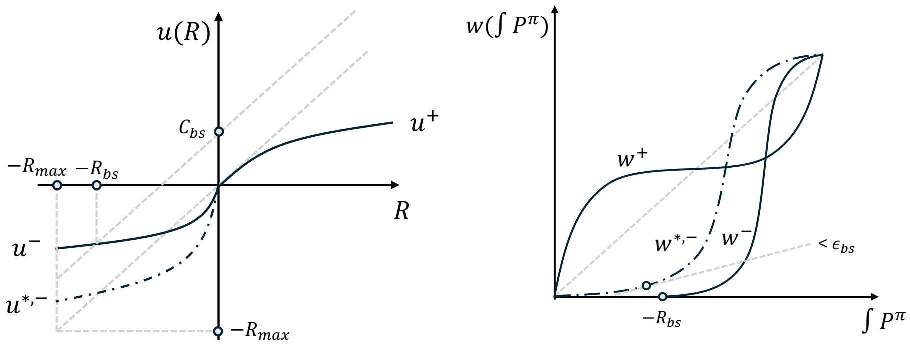
Again, we note that the gap between $\mathcal{M}$ and $\mathcal{M}^\dagger$ is defined over a gap between (u^-, w^-) and $(u_\star^-, w_\star^-)$ that is proportional to the existence of spatial black swan events.

8.2 Theortical analysis

Based on the perspective of the blackswan event as a reward (Proposition 3), the natural question would be how does the perception gap create a gap between the value function?

Theorem 4 (**Convergence of estimation value but lower bound on perceived value gap**). *Under Assumptions 2, 3, and 4, the asymptotic convergence of the value function estimation holds as follows,*

$$\widehat{V}_t^\dagger(s_0) \to V_t^\dagger(s_0) \quad a.s. \quad as \quad t \to \infty. \tag{14}$$



(a) Utility function u (b) Weight function w

Figure 3: Utility and weight functions with spatial black swans.

where upper bound of the estimation error (sample complexity) holds as:

$$P(|\widehat{V}_t^\dagger - V_t^\dagger| < \epsilon) \leq 1 - 4e^{-t\frac{\epsilon^2}{2c^2}}. \tag{15}$$

where $c = \max\{|L^+ u^+(R_{\max})|, |L^- u^-(-R_{\max})|\}$*.*

However, under specific conditions on $\epsilon_{bs}, \epsilon_{bs}^{\min}, R_{bs}$*, the lower bound of human value function gap is expressed as:*

$$|V_t^\dagger(s_0) - V_t(s_0)| = \Omega\left(\frac{\left((R_{\max} - R_{bs})\,\epsilon_{bs}^{\min} - R_{bs}\epsilon_{bs}\right)(R_{\max} - R_{bs})C_{bs}}{R_{\max}^2}\right) \tag{16}$$

due to misperception.

Insights of Theorem 4. The message of Theorem 4 establishes a support for our hypothesis 2 and Defintion 7. Firstly, Equation (15) establishes that the estimation error converges to zero as the agent rolls out a longer trajectory. However, Equation (16) offers the insight that the value function gap between $\mathcal{M}$ and $\mathcal{M}^\dagger$ has a lower bound, regardless of the trajectory length. It is straightforward to verify in Equation (16) that if $u^-(x) \to u_\star^-(x)$ and $w^-(x) \to w_\star^-(x)$, which implies $R_{bs} \to R_{\max}$ and $\epsilon_{bs} \to 0$, then the lower bound also converges to zero. Furthermore, Equation (16) supports our intuition that a higher distortion in reward perception (a large C_{bs} value) and a greater number of spatial black swan events (large $(R_{\max} - R_{bs})$) and a large minimum probability of spatial black swan event occurrence (large $\epsilon_{bs}^{\min}$) lead to a higher lower bound. It is important to note that the gap $R_{\max} - R_{bs}$ is associated with the number of spatial black swans, since we have assumed that the reward function maps uniquely, therefore there exists a unique (s, a) such that $R(s, a) = r$, where r is in the range $[-R_{\max}, R_{\max} + R_{bs}]$. Therefore, through Theorem 4, we conclude that even though the agent has a perfect estimation of what it has perceived, there still exists a lower bound to obtaining the value function of the universe, which increases as the parameter effect of spatial black swan events becomes more dominant.

Now, based on Theorem 4, the next natural question is *how to decrease the lower bound*, that is, how the agent learns to nudge $u_{\star}^{-} \to u_{\star}^{-}$ and $w_{\star}^{-} \to w_{\star}^{-}$. Before we ask when the agent encounters the black swan, so the perception correction would happen. Since black swan events are defined over probability, this question could be more concretized as *If the agent takes a step $h > h_{bs}$, the probability of encountering black swan events are at least δ_{bs}.* This question also further enables to naturally learn how to obtain antifragile behavior, which makes δ_{bs} to decrease as every time the agent updates its perception. We first start our analysis based on the non-zero one-step reachability assumption.

Theorem 5 (Spatial black swan hitting time). *Assume $\mathbb{P}_{\pi^\star}(s'|s) > 0$ for any two states $s, s' \in \mathcal{S}$, signifying that the one-step state reachability is non-zero, and consider that one step corresponds to a unit time. Then, if the agent takes t steps such that*

$$t \geq \frac{\log\left(\frac{\delta}{p_{\min}}\right)}{\log(1 - p_{\max})} + 1,$$

where $p_{\min} = \frac{R_{\max} - R_{bs}}{2R_{\max}} \epsilon_{bs}^{\min}$ and $p_{\max} = \frac{R_{\max} - R_{bs}}{2R_{\max}} \epsilon_{bs}$, it will encounter spatial black swan events at least with probability $\delta \in [0, 1]$.

9 Conclusion

In this paper, we propose a new perspective to understand the black swan events by utilizing human misperception and CPT. We have divided black swans into temporal and spatial black swans and focus on spatial black swans, where misperception occurs in the state and action space. To define spatial black swans, we propose three different MDPs: universe MDP ($\mathcal{M}$), human MDP ($\mathcal{M}^\dagger$), and human-estimation MDP ($\widehat{\mathcal{M}}^\dagger$) and also introduce the well-known irrational human behavior theorem, CPT. The main message is Hypothesis 2 and Definition 7. They highlight that despite humans being able to reduce the estimation gap, leading $\widehat{\mathcal{M}}^\dagger$ to converge to $\mathcal{M}^\dagger$ through longer horizons or better algorithms, there remains an inherent uncertainty due to the perception gap between $\mathcal{M}$ and $\mathcal{M}^\dagger$. This implies that spatial black swans can still occur because of this perception gap, even when the agent has zero estimation error and computes its optimal policy from $\widehat{\mathcal{M}}^\dagger$.

Acknowledgments and Disclosure of Funding

We thank Jason Jangho Choi for providing instructive insights on the early draft, and Donghao Ying for discussions on mathematical notation. The authors also thank Theophane Weber and Csaba Szepesvari and the rest of the Google DeepMind team who reviewed the paper.

Appendix A. Proofs

We first start from the following Lemma 3 to prove Lemma 1.

Lemma 3 (Bounding visitation probability of step h when well-perceived transition holds)**.** *If for all (s,a) holds ϵ_d-perceived transition probability, then we have*

$$\max_{\pi}\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|\mathbb{P}_h^{\pi}(s,a)-\mathbb{P}_h^{\pi,\dagger}(s,a)\right|\right)\le h\epsilon_p$$

that holds for all $h\in\mathbb{N}$

Proof of Lemma 3. Proof by induction. We use short notation for $P(s_h = s \mid s_{h-1} = s', a_{h-1} = a')$ as $P_h(s \mid s', a')$ and $P^{\dagger}(s_h = s \mid s_{h-1} = s', a_{h-1} = a')$ as $P_h^{\dagger}(s \mid s', a')$. By the definition of rational transition probability the statement holds at $h = 1$ for any policy π. Now, suppose the statement holds for $h-1$ for any policy π. Then, we have

$$\begin{aligned}
&\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|\mathbb{P}_h^{\pi}(s,a)-\mathbb{P}_h^{\pi,\dagger}(s,a)\right| \\
&\quad=\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\Bigg|\pi(a_h=a\mid s_h=s)\sum_{s',a'}\left(P_h(s\mid s',a')\mathbb{P}_{h-1}^{\pi}(s',a')\right) \\
&\qquad-\pi(a_h=a\mid s_h=s)\sum_{s',a'}\left(P_h^{\dagger}(s\mid s',a')\mathbb{P}_{h-1}^{\pi,\dagger}(s',a')\right)\Bigg| \\
&\quad\le\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\pi(a_h=a\mid s_h=s)\left|\sum_{s',a'}\left(P_h(s\mid s',a')\mathbb{P}_{h-1}^{\pi}(s',a')\right)-\sum_{s',a'}\left(P_h^{\dagger}(s\mid s',a')\mathbb{P}_{h-1}^{\pi,\dagger}(s',a')\right)\right| \\
&\quad=\sum_{s\in\mathcal{S}}\left|\sum_{s',a'}\left(P_h(s\mid s',a')\mathbb{P}_{h-1}^{\pi}(s',a')\right)-\sum_{s',a'}\left(P_h^{\dagger}(s\mid s',a')\mathbb{P}_{h-1}^{\pi,\dagger}(s',a')\right)\right| \\
&\quad=\sum_{s\in\mathcal{S}}\left|\sum_{s',a'}\left(P_h-P_h^{\dagger}\right)\mathbb{P}_{h-1}^{\pi}(s',a')+\sum_{s',a'}P_h^{\dagger}(s\mid s',a')\left(\mathbb{P}_{h-1}^{\pi}(s',a')-\mathbb{P}_{h-1}^{\pi,\dagger}(s',a')\right)\right| \\
&\quad\le\sum_{s',a'}\left|\sum_{s\in\mathcal{S}}\left(P_h-P_h^{\dagger}\right)\mathbb{P}_{h-1}^{\pi}(s',a')\right|+\sum_{s',a'}\left|\sum_{s\in\mathcal{S}}P_h^{\dagger}(s\mid s',a')\left(\mathbb{P}_{h-1}^{\pi}(s',a')-\mathbb{P}_{h-1}^{\pi,\dagger}(s',a')\right)\right| \\
&\quad\le\epsilon_p\sum_{s',a'}\mathbb{P}_{h-1}^{\pi}(s',a')+1\cdot(h-1)\epsilon_p \\
&\quad=\epsilon_p\cdot 1+(h-1)\epsilon_p \\
&\quad\le h\epsilon_p
\end{aligned}$$

The all of above inequalities holds for all π. Therefore, the statement holds for all $h\in\mathbb{N}$. ☐

Now, we state the proof of Lemma 1.

Proof of Lemma 1. Lemma 1 is almost a corollary that stems from Lemma 3. By the definition of visitation probability, we have

$$\begin{aligned}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|P^{\pi}(s,a)-P^{\pi,\dagger}(s,a)\right| &= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|\sum_{h=0}^{\infty}\gamma^h\left(\mathbb{P}_h^{\pi}(s,a)-\mathbb{P}_h^{\dagger,\pi}(s,a)\right)\right| \\ &\le \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{h=0}^{\infty}\gamma^h\left|\left(\mathbb{P}_h^{\pi}(s,a)-\mathbb{P}_h^{\dagger,\pi}(s,a)\right)\right| \\ &= \sum_{h=0}^{\infty}\gamma^h\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|\left(\mathbb{P}_h^{\pi}(s,a)-\mathbb{P}_h^{\dagger,\pi}(s,a)\right)\right| \\ &\le \sum_{h=0}^{\infty}\gamma^h h\frac{(1-\gamma)^2}{\gamma}\epsilon_p\end{aligned}$$

Let $S=\sum_{h=0}^{\infty}\gamma^h h$, then $\gamma S=\sum_{h=0}^{\infty}\gamma^{h+1}h=\sum_{h=1}^{\infty}\gamma^h(h-1)$. Then by subtracing those two equations, we have $(1-\gamma)S=\sum_{h=1}^{\infty}\gamma^h=\frac{\gamma}{1-\gamma}$. Therefore we have $S=\frac{\gamma}{(1-\gamma)^2}$. Finally, we have the following inequality

$$\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|P^{\pi}(s,a)-P^{\pi,\dagger}(s,a)\right|\le\frac{\gamma}{(1-\gamma)^2}\cdot\frac{(1-\gamma)^2}{\gamma}\epsilon_p=\epsilon_p$$

□

Proof of Lemma 2. First, note that we have assumed the image of the function R is closed and dense as $[-R_{\max},R_{\max}]$. Then, in the progress of projecting all (s,a) into the reward, we define the probability of reward as $\mathbb{P}(\mathcal{R}=r)=\sum_{\forall(s,a)\in\mathcal{S}\times\mathcal{A}}d^{\pi}(s,a)\mathbf{1}[R(s,a)=r]$. we use short notation for $\mathbb{P}(\mathcal{R}=r)$ as $\mathbb{P}_{\mathcal{R}}$. Now, since $d^{\pi}(s,a)$ is the visitation probability of visiting (s,a), then this could be converted to $\mathbb{P}(\mathcal{R}=r)$ by $d^{\pi}(\mathcal{R}=R^{-1}(s,a))$ where R^{-1} is many to one function.

Now, since $\mathbb{R}$ is the many to one function, we can define independent block the $\mathcal{S},\mathcal{A}$ as the set $Z(r):=\{(s,a)\in\mathcal{S}\times\mathcal{A}|R(s,a)=r\}$. Note that if $r_1\neq r_2$, then $Z(r_1)\cap Z(r_2)=0$. Then, if g satisfies the set Z to in be permutation- invariant. Namely, if $R(s_1,a)=R(s_2,a)$, then $R(g(s_1))=R(g(s_2),a)$ holds then there exists a one-to-one mapping function $h:[-R_{\max},R_{\max}]\to[-R_{\max},R_{\max}]$ such that

$$R(s,a)=h(R(g(s),a))$$

holds. The proof can be divided into two folds. The existence of such function and its a one-to-one mapping function is exists. We first prove the existence of such function h. This is because for any state and action s,a, suppose its reward value as r. Then suppose $g(s)=s'$. Then since image of function R is closed and dense, there exists $r'\in[-R_{\max},R_{\max}]$ such that $R(s',a)=r'$ holds. Then, one can say the function $r=h(r')$ exists. Now, we prove the one-to-one mapping property. suppose for two state and action pair (s_1,a_1) and (s_2,a_2)

and let $s_1' = g(s_1')$ and $s_2' = g(s_2')$. Now, suppose $R(s_1', a) \neq R(s_2', a)$ holds. Then, due to the property of g, then it should also satisfies $R(s_1, a) \neq R(s_2, a)$. Therefore, this concludes that h is the one-to-one mapping, and the following holds

$$\begin{aligned} d^\pi(R(g(s), a) = r) &= d^\pi(h(R(g(s), a)) = h(r)) \\ &= d^\pi(R(s, a) = h(r)) \\ &= \mathbb{P}(\mathcal{R} = h(r)) \end{aligned}$$

holds. we denote $\mathbb{P}(\mathcal{R} = h(r))$ as $\mathbb{P}_{h(\mathcal{R})}$. Then, let's define two different functions h^+ and h^- such that we want to claim that

$$w^-\left(\int_{-R_{max}}^{r} d\mathbb{P}_{\mathcal{R}}\right) = \int_{-R_{max}}^{r} d\mathbb{P}_{h^-(\mathcal{R})}, \quad \text{and} \quad w^+\left(\int_{-R_{max}}^{r} d\mathbb{P}_{\mathcal{R}}\right) = \int_{-R_{max}}^{r} d\mathbb{P}_{h^+(\mathcal{R})} \tag{17}$$

holds for any w^-, w^+. Since the proof for either is similar, we prove the case for the existence of h^- under w^- distortion.

Now, recall that for $0 < x < b$, $w^-(x) < x$ holds and for $b < x < 1$, $w^-(x) > x$ holds and $w^-(x)$ is monotically increasing function. Define $r_b \in [-R_{\max}, 0]$ such that $b := \int_{-R_{\max}}^{r_b} d\mathbb{P}_{\mathcal{R}}$ holds, and for notation simplicity we deonte $F^-(r) = \int_{-R_{\max}}^{r_b} d\mathbb{P}_{\mathcal{R}}$. Then, one can say $-R_{\max} < r < r_b$, $w(F(r)) < F(r)$ holds and. Then we can always find a unique ratio $0 < \gamma(r) < 1$ that depends on r such that $w^-(F(r)) = \int_{-R_{\max}}^{\gamma(r)r} d\mathbb{P}_r$ holds where

$$\gamma(r) = \frac{w^-(F(r))}{r}.$$

This leads to set $h(r) = \gamma(r)r = w^-(F(r))$ that satisfies (17) and also one-to-one mapping. In the same manner, we can also identify $h(r) = \gamma(r)r = w^-(F(r))$ where $r_b < r < 0$ holds for $\gamma(r) > 1$. Then, this completes that the function $h : r \to w^-(F(r))$ satisfies a one-to-one function and Equation (17). This completes the proof. □ □

Proof of Theorem 1. By the definition of optimal policy and the value function definition at the time $T = 1$, we have the optimal policy at time 0 as follows.

$$\begin{aligned} \pi^\star &= \arg\max_{\pi} V_0(s) \\ &= \arg\max_{a \in \mathcal{A}} Q_0(s, a) \\ &= \arg\max_{a \in \mathcal{A}} R(s, a) \\ \pi^{\star,\dagger} &= \arg\max_{a \in \mathcal{A}} V_0^\dagger(s) \\ &= \arg\max_{a \in \mathcal{A}} Q_0^\dagger(s, a) \\ &= \arg\max_{a \in \mathcal{A}} u(R(s, a)) \end{aligned}$$

for any fixed $s \in \mathcal{S}$, let's assume a^* is the argument that maximizes the $R(s,a)$. Since u is the non-decreasing convex function, a^* is still the same argument that maximizes the $u(R(s,a))$. Therefore, $\pi^\star = \pi^{\star,\dagger}$ holds. □ □

Proof of Theorem 2. We prove by backward induction. First by theorem 1, $\pi_T^\star = \pi_T^{\star,\dagger}$ holds. Now suppose that $\pi_{t'+1}^\star = \pi_{t'+1}^{\star,\dagger}$ holds for all $t' = t+1,\cdots,T$. Now, we prove the statement holds for t. To prove $\pi_t^\star = \pi_t^{\star,\dagger}$, it is sufficient to show if $Q_t^{\pi^\star}(s,a) \geq Q_t^{\pi}(s,a')$, then $Q_t^{\dagger,\pi^\star}(s,a) \geq Q_t^{\dagger,\pi^\star}(s,a')$ also holds for any actions $a, a' \in \mathcal{A}$. First, the gap $Q_t^{\pi^\star}(s,a) - Q_t^{\pi^\star}(s,a)$ could be expressed as

$$
\begin{aligned}
Q_t^{\pi}(s,a) - Q_t^{\pi}(s,a) &= R_t(s,a) - R_t(s,a') + \left\{\left(P(s_1|s,a) - P(s_2|s,a')\right)\left(V_{t+1}^{\pi^\star}(s_1) - V_{t+1}^{\pi^\star}(s_2)\right)\right\} \\
&= \left(P(s_1|s,a) - P(s_2|s,a')\right)\left(V_{t+1}^{\pi^\star}(s_1) - V_{t+1}^{\pi^\star}(s_2)\right)
\end{aligned}
$$

and $Q_t^{\dagger,\pi^\star}(s,a) - Q_t^{\dagger,\pi^\star}(s,a)$ as

$$
\begin{aligned}
Q_t^{\dagger,\pi^\star}(s,a) - Q_t^{\dagger,\pi^\star}(s,a) &= R_t^\dagger(s,a) - R_t^\dagger(s,a') + \left\{\left(P^\dagger(s_1|s,a) - P^\dagger(s_2|s,a')\right)\left(V_{t+1}^{\pi^\star}(s_1) - V_{t+1}^{\pi^\star}(s_2)\right)\right\} \\
&= \left(P^\dagger(s_1|s,a) - P^\dagger(s_2|s,a')\right)\left(V_{t+1}^{\dagger,\pi^\star}(s_1) - V_{t+1}^{\dagger,\pi^\star}(s_2)\right) \\
&= \left(w(P^\dagger(s_1|s,a)) - w(P^\dagger(s_2|s,a'))\right)\left(V_{t+1}^{\dagger,\pi^\star}(s_1) - V_{t+1}^{\dagger,\pi^\star}(s_2)\right)
\end{aligned}
$$

the reward during $t \in [1, T-1]$ is zero by our problem formulation assumption in section 6.1. Now, without loss of generality, we assume $V_{t+1}^{\pi^\star}(s_1) > V_{t+1}^{\pi^\star}(s_2)$. Then, due to our assumption that $\pi_{t'}^\star = \pi_{t'}^{\star,\dagger}$ holds for $t' = t+1,\cdots,T$, we also have $V_{t+1}^{\dagger,\pi^\star}(s_1) > V_{t+1}^{\dagger,\pi^\star}(s_2)$. Also, noticing that weight function w is also increasing function, then $P(s_1|s,a) > P(s_2|s,a)$ also guarantees $w(P(s_1|s,a)) > w(P(s_2|s,a))$ holds. Therefore, we can claim if $Q_t^{\pi}(s,a) - Q_t^{\pi}(s,a) > 0$ holds, then $Q_t^{\dagger,\pi^\star}(s,a) - Q_t^{\dagger,\pi^\star}(s,a) > 0$ also holds. Then, this leads to claim that $\arg\max Q_t^{\pi}(s,a) = \arg\max Q_t^{\dagger\pi}(s,a)$, which implies $\pi_t^\star = \pi_t^{\star,\dagger}$. This completes the proof. □ □

Proof of Theorem 3. Assume that Theorem 3 does not hold. Given $T = 2$, we have $V_2^{\pi}(s) = \max_{a\in\mathcal{A}} R_2(s,a) = R_2(s)$ for each state s. At time $t = 1$, assume $R_2(s_1) \leq R_2(s_2) \leq R_2(s_3)$. The condition $Q_1^{\dagger,\pi}(s,a_1) \geq Q_1^{\dagger,\pi}(s,a_2)$ is then expressed as:

$$
\begin{aligned}
& w\left(P\left(s_1 \mid s, a_1\right)\right) r_2\left(s_1\right) + \left(w\left(P\left(s_2 \mid s, a_1\right) + P\left(s_1 \mid s, a_1\right)\right) - w\left(P\left(s_1 \mid s, a_1\right)\right)\right) R_2\left(s_2\right) \\
& \quad + \left(1 - w\left(P\left(s_2 \mid s, a_1\right) + P\left(s_1 \mid s, a_1\right)\right)\right) R_3\left(s_3\right) \\
& \geq w\left(P\left(s_1 \mid s, a_2\right)\right) R_2\left(s_1\right) + \left(w\left(P\left(s_2 \mid s, a_2\right) + P\left(s_1 \mid s, a_2\right)\right) - w\left(P\left(s_1 \mid s, a_2\right)\right)\right) R_2\left(s_2\right) \\
& \quad + \left(1 - w\left(P\left(s_2 \mid s, a_2\right) + P\left(s_1 \mid s, a_2\right)\right)\right) R_3\left(s_3\right)
\end{aligned}
$$

which simplifies to:

$$
\begin{aligned}
&\left(w\left(P\left(s_1 \mid s,a_1\right)\right)-w\left(P\left(s_1 \mid s,a_2\right)\right)\right)\left(R_2\left(s_1\right)-R_3\left(s_3\right)\right)\\
&\quad+\left(\left(w\left(P\left(s_2 \mid s,a_1\right)+P\left(s_1 \mid s,a_1\right)\right)-w\left(P\left(s_1 \mid s,a_1\right)\right)\right)\right.\\
&\quad\left.-\left(w\left(P\left(s_2 \mid s,a_2\right)+P\left(s_1 \mid s,a_2\right)\right)-w\left(P\left(s_1 \mid s,a_2\right)\right)\right)\right)\left(R_2\left(s_2\right)-R_3\left(s_3\right)\right)\geq 0
\end{aligned}
$$

For the non-distorted case, the analogous expression is:

$$
\begin{aligned}
&\left(P\left(s_1 \mid s,a_1\right)-P\left(s_1 \mid s,a_2\right)\right)\left(R_2\left(s_1\right)-R_3\left(s_3\right)\right)\\
&\quad+\left(P\left(s_2 \mid s,a_1\right)-P\left(s_2 \mid s,a_2\right)\right)\left(R_2\left(s_2\right)-R_3\left(s_3\right)\right)\geq 0
\end{aligned}
$$

For arbitrary reward functions, R_2, the equality of the two cases under any weighting function w leads to:

$$
\begin{aligned}
&\frac{w\left(P\left(s_1 \mid s,a_1\right)\right)-w\left(P\left(s_1 \mid s,a_2\right)\right)}{w\left(P\left(s_2 \mid s,a_1\right)+P\left(s_1 \mid s,a_1\right)\right)-w\left(P\left(s_1 \mid s,a_1\right)\right)-\left(w\left(P\left(s_2 \mid s,a_2\right)+P\left(s_1 \mid s,a_2\right)\right)-w\left(P\left(s_1 \mid s,a_2\right)\right)\right)}\\
&=\frac{P\left(s_1 \mid s,a_1\right)-P\left(s_1 \mid s,a_2\right)}{P\left(s_2 \mid s,a_1\right)-P\left(s_2 \mid s,a_2\right)}
\end{aligned}
$$

where $w(p)=p$ is the only solution, contradicting the distortion required by Assumption 3. $\square$

Proof of Theorem 4. The proof of Theorem 4 is divided into three-fold.

1. Proof of asymptotic convergence

We first prove asymptotic convergence (Equation (14)), then we prove Equation (15) in part 3 of this proof. Note that the empirical distribution function $\widehat{F}_n(r)$ generate Stielgies measure which takes mass $\frac{1}{t}$ each of the sample points on $U^+(R_i)$.

or equivalently, show that

$$
\lim_{n\to+\infty}\sum_{i=1}^{n-1} u^+(R_{[i]})(w^+(\frac{n-i+1}{n})-w^+(\frac{n-i}{n}))\xrightarrow{n\to\infty}\int_0^{+\infty} w^+(P(U>t))dt,\text{w.p. } 1 \tag{18}
$$

where n denotes the number of positive reward among $|\mathcal{S}||\mathcal{A}|$. Let $\xi^+_{\frac{i}{n}}$ and $\xi^-_{\frac{i}{n}}$ denote the $\frac{i}{n}$th quantile of $u^+(X)$ and $u^-(X)$, respectively.

For the convergence proof, we first concentrate on finding the following probability,

$$
P(\left|\sum_{i=1}^{n-1} u^+(R_{[i]})\cdot(w^+(\frac{n-i}{n})-w^+(\frac{n-i-1}{n}))-\sum_{i=1}^{n-1}\xi^+_{\frac{i}{n}}\cdot(w^+(\frac{n-i}{n})-w^+(\frac{n-i-1}{n}))\right|>\epsilon), \tag{19}
$$

for any given $\epsilon > 0$. It is easy to check that

$$P(\left|\sum_{i=1}^{n-1} u^+(R_{[i]}) \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n}))\right| > \epsilon)$$
$$\leq P(\bigcup_{i=1}^{n-1} \left\{\left|u^+(R_{[i]}) \cdot (w^+_(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \xi^+_{\frac{i}{n}} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n}))\right| > \frac{\epsilon}{n}\right\})$$
$$\leq \sum_{i=1}^{n-1} P(\left|u^+(R_{[i]}) \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \xi^+_{\frac{i}{n}} \cdot (w^+_(\frac{n-i}{n}) - w^+_(\frac{n-i-1}{n}))\right| > \frac{\epsilon}{n}) \tag{20}$$
$$= \sum_{i=1}^{n-1} P(\left|(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}}) \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n}))\right| > \frac{\epsilon}{n})$$
$$\leq \sum_{i=1}^{n-1} P(\left|(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}}) \cdot \cdot (\frac{1}{n})^{\alpha}\right| > \frac{\epsilon}{n})$$
$$= \sum_{i=1}^{n-1} P(\left|(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}})\right| > \frac{\epsilon}{\cdot n^{1-\alpha}}). \tag{21}$$

The Right hand side of Inequality (21) could be expressed as follows.

$$P(\left|u^+(R_{[i]}) - \xi^+_{\frac{i}{n}}\right| > \frac{\epsilon}{n^{(1-\alpha)}})$$
$$= P(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}} > \frac{\epsilon}{n^{(1-\alpha)}}) + P(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}} < -\frac{\epsilon}{n^{(1-\alpha)}}).$$

We focus on the term $P(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}} > \frac{\epsilon}{n^{1-\alpha}})$. Now, let us define an event $A_t = I_{(u^+(X_t) > \xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}})}$ where $t = 1, \ldots, n$. Since the Cumulative distribution is non-decrasing function, we have the following,

$$P(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}} > \frac{\epsilon}{1-\alpha}) = P(\sum_{t=1}^{n} A_t > n \cdot (1 - \frac{i}{n^{(1-\alpha)}}))$$
$$= P(\sum_{t=1}^{n} A_t - n \cdot [1 - F^+(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}})] > n \cdot [F^+(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}) - \frac{i}{n}]).$$

Using the fact that $\mathbb{E}A_t = 1 - F^+(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}})$ in conjunction with Hoeffding's inequality, we obtain

$$P(\sum_{i=1}^{n} A_t - n \cdot [1 - F^+(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}})] > n \cdot [F^+(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}) - \frac{i}{n}]) < e^{-2n \cdot \delta'_t}, \tag{22}$$

where $\delta'_i = F^+(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}) - \frac{i}{n}$. Since $F^+(x)$ is Lipschitz, we have that $\delta'_i \leq L_{F^+} \cdot (\frac{\epsilon}{1-\alpha})$. Hence, we obtain

$$P(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}} > \frac{\epsilon}{1-\alpha}) < e^{-2n \cdot L_{F^+} \frac{\epsilon}{1-\alpha}} = e^{-2n^{\alpha} \cdot L^+ \epsilon} \tag{23}$$

In a similar fashion, one can show that

$$P(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}} < -\frac{\epsilon}{1-\alpha}) \leq e^{-2n^\alpha \cdot L_{F^+\epsilon}} \tag{24}$$

Combining (23) and (24), we obtain

$$P(\left|u^+(R_{[i]}) - \xi^+_{\frac{i}{n}}\right| > \frac{\epsilon}{1-\alpha}) \leq 2 \cdot e^{-2n^\alpha \cdot L_{F^+\epsilon}}, \ \forall i \in \mathbb{N} \cap (0,1)$$

Plugging the above in (21), we obtain

$$\begin{aligned}&P(\left|\sum_{i=1}^{n-1} u^+(R_{[i]}) \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n}))\right| > \epsilon) \\ &\leq 2n \cdot e^{-2n^\alpha \cdot L_{F^+}}.\end{aligned} \tag{25}$$

Notice that $\sum_{n=1}^{+\infty} 2n \cdot e^{-2n^\alpha \cdot L_{F^+\epsilon}} < \infty$ since the sequence $2n \cdot e^{-2n^\alpha \cdot L_{F^+}}$ will decrease more rapidly than the sequence $\frac{1}{n^k}$, $\forall k > 1$.

By applying the Borel Cantelli lemma, we have that $\forall \epsilon > 0$

$$P(\left|\sum_{i=1}^{n-1} u^+(R_{[i]}) \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n}))\right| > \epsilon) = 0,$$

which implies

$$\sum_{i=1}^{n-1} u^+(R_{[i]}) \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) \xrightarrow{n \to +\infty} 0 \text{ w.p } 1,$$

which proves (18).

Also, the remaining part, conducting the proof of convergence of w^- and u^-,i.e.

$$\lim_{n \to +\infty} \sum_{i=1}^{n-1} u^-(R_{[i]})(w^-(\frac{n-i+1}{n}) - w^-(\frac{n-i}{n})) \xrightarrow{n \to \infty} \int_0^{+\infty} w^-(P(U > t))dt, \text{w.p. } 1 \tag{26}$$

also follows simliar manner. we omit the proof for this.

2. Proof of value function lower bound

By the definition, we have the following

$$\begin{aligned}
|V(s_0) - V^\dagger(s_0)| &= \left| \int_{-\infty}^0 w_\star^-(\mathbb{P}_r(u_\star^-(\mathcal{R} > r)))dr - \int_\infty^0 w^-(\mathbb{P}_r(u^-(\mathcal{R} > r)))dr \right| \\
&= \Bigg| \int_{-\infty}^0 w_\star^-(\mathbb{P}_r(u_\star^-(\mathcal{R} > r)))dr - \int_\infty^0 w_\star^-(\mathbb{P}_r(u^-(\mathcal{R} > r)))dr \\
&\quad - \left(\int_\infty^0 w^-(\mathbb{P}_r(u^-(\mathcal{R} > r)))dr - \int_\infty^0 w_\star^-(\mathbb{P}_r(u^-(\mathcal{R} > r)))dr \right) \Bigg| \\
&\geq \underbrace{\left| \int_{-\infty}^0 w_\star^-(\mathbb{P}_r(u_\star^-(\mathcal{R} > r)))dr - \int_{-\infty}^0 w_\star^-(\mathbb{P}_r(u^-(\mathcal{R} > r)))dr \right|}_{\text{Term (I)}} \\
&\quad - \underbrace{\left| \int_{-\infty}^0 w_\star^-(\mathbb{P}_r(u_\star^-(\mathcal{R} > r)))dr - \int_\infty^0 w_\star^-(\mathbb{P}_r(u^-(\mathcal{R} > r)))dr \right|}_{\text{Term (II)}}
\end{aligned}$$

We first under bound the term (I). For notation simplicity, we let $g(r) = \mathbb{P}_r(u^-(\mathcal{R} > r)))$ and $g_\star(r) = \mathbb{P}_r(u_\star^-(\mathcal{R} > r)))$. Then we have the following

$$\text{Term (I)} = \left| \int_{-R_{\max}}^0 w_\star^-(g_\star(r)) - w_\star^-(g(r)) \right|$$

Now, since $w_\star^-(x)$ is monotonically increasing in $x \in [0, a]$ and monotonically decreasing in $x \in [a, 1]$, we could say for any $x, y \in [0, 1], x \neq y$ that

$$\frac{w_\star^-(x) - w_\star^-(y)}{x - y} = (w_\star^-)'(z) \geq \min_{z \in [0,1]} (w_\star^-)'(z) = \min \left\{ (w_\star^-)'(0), (w_\star^-)'(1) \right\},$$

where $z \in (x, y)$. The first equality holds due to the mean value theorem. Therfore it holds that

$$\begin{aligned}
\text{Term (I)} &= \left| \int_{-R_{\max}}^0 w_\star^-(g_\star(r)) - w_\star^-(g(r)) \right| \\
&\geq \left| \int_{-R_{\max}}^0 \min \left\{ (w_\star^-)'(0), (w_\star^-)'(1) \right\} (g_\star(r) - g(r)) \right| \\
&= \min \left\{ (w_\star^-)'(0), (w_\star^-)'(1) \right\} \left| \int_{-R_{\max}}^0 (g_\star(r) - g(r)) \right|
\end{aligned}$$

Now, recall the definition of $g_\star(r)$ and $g(r)$, then we have the following

$$\left| \int_{-R_{\max}}^0 (g_\star(r) - g(r))\, dr \right| = \left| \mathbb{E}_{\mathcal{R} \sim \mathbb{P}_\pi} \left[u_\star^-(\mathcal{R}) - u^-(\mathcal{R}) \right] \right|$$

Now, let us denote the intersection of $u^-(R)$ and $y = R + C_{bs}$ as $R = -R_{bs}$. We can say if the blackswan happens, then its reward is bounded between $[-R_{\max}, -R_{bs}]$. Then we have the following,

$$\begin{aligned}
\left|\int_{-R_{\max}}^{0} (g_\star(r) - g(r))\right| &= \left|\mathbb{E}_{\mathcal{R}\sim\mathbb{P}_\pi}\left[u_\star^-(\mathcal{R}) - u^-(\mathcal{R})\right]\right| \\
&= \Big|\mathbb{E}_{\mathcal{R}\sim\mathbb{P}_\pi}\left[\mathbf{1}\left[\mathcal{R} < -R_{bs}\right]\left(u_\star^-(\mathcal{R}) - u^-(\mathcal{R})\right)\right] \\
&\quad - \mathbb{E}_{\mathcal{R}\sim\mathbb{P}_\pi}\left[\mathbf{1}\left[\mathcal{R} \geq -R_{bs}\right]\left(-u_\star^-(\mathcal{R}) + u^-(\mathcal{R})\right)\right]\Big| \\
&\geq \underbrace{\left|\mathbb{E}_{\mathcal{R}\sim\mathbb{P}_\pi}\left[\mathbf{1}\left[\mathcal{R} < -R_{bs}\right]\left(u_\star^-(\mathcal{R}) - u^-(\mathcal{R})\right)\right]\right|}_{\text{Term I-1}} \\
&\quad - \underbrace{\left|\mathbb{E}_{\mathcal{R}\sim\mathbb{P}_\pi}\left[\mathbf{1}\left[\mathcal{R} \geq -R_{bs}\right]\left(-u_\star^-(\mathcal{R}) + u^-(\mathcal{R})\right)\right]\right|}_{\text{Term I-2}} \\
&\geq \left|\mathbb{E}_{\mathcal{R}\sim\mathbb{P}_\pi}\left[\mathbf{1}\left[\mathcal{R} < -R_{bs}\right]\left(u_\star^-(\mathcal{R}) - u^-(\mathcal{R})\right)\right]\right|
\end{aligned}$$

To lower bound the Term I-1, let's denote the minimum reachability of blackswan events as $\epsilon_{bs}^{\min} \neq 0$. Then we have

$$\begin{aligned}
\text{Term I-1} &\geq \frac{R_{\max} - R_{bs}}{R_{\max}} \epsilon_{bs}^{\min} \min_{R \in [-R_{\max}, -R_{bs}]} \left|u^-(R) - u_\star^-(R)\right| \\
&\geq \frac{R_{\max} - R_{bs}}{R_{\max}} \epsilon_{bs}^{\min} \left|u^-(-R_{bs}) - u_\star^-(-R_{bs})\right| \qquad (27)
\end{aligned}$$

$$\begin{aligned}
\text{Term I-2} &\leq \frac{R_{bs}}{R_{\max}} \epsilon_{bs} \max_{R \in [-R_{bs}, 0]} \left|u^-(R) - u_\star^-(R)\right| \\
&\leq \frac{R_{bs}}{R_{\max}} \epsilon_{bs} \left|u^-(-R_{bs}) - u_\star^-(-R_{bs})\right| \qquad (28)
\end{aligned}$$

Therefore, we have the following equation,

$$\text{Term I} \geq \frac{(R_{\max} - R_{bs})\,\epsilon_{bs}^{\min} - R_{bs}\epsilon_{bs}}{R_{\max}} \left|u^-(-R_{bs}) - u_\star^-(-R_{bs})\right|$$

Also, since the function $u_\star^-(r)$ is convex, and $u_\star^-(-R_{\max}) < -R_{\max} + C_{bs}$ holds. Therefore, we could say $u_\star^-(r) < \frac{R_{\max} - C_{bs}}{R_{\max}} r$. This leads us to come up with $u_\star^-(-R_b s) < \frac{R_{\max} - C_{bs}}{R_{\max}}(-R_{bs})$.

Therefore, we have a gap lowerbound as

$$\begin{aligned}|u^-(-R_{bs}) - u^-_\star(-R_{bs})| &\geq (R_{\max} - C_{bs})\frac{R_{bs}}{R_{\max}} - (R_{bs} - C_{bs}) \\ &= \frac{(R_{\max} - R_{bs})C_{bs}}{R_{\max}}\end{aligned}$$

The above inequality could be minimized as

$$\begin{aligned}\text{Term I} &\geq \frac{(R_{\max} - R_{bs})\,\epsilon_{bs}^{\min} - R_{bs}\epsilon_{bs}}{R_{\max}}\left(\frac{(R_{\max} - R_{bs})C_{bs}}{R_{\max}}\right) \\ &= \frac{\left((R_{\max} - R_{bs})\,\epsilon_{bs}^{\min} - R_{bs}\epsilon_{bs}\right)(R_{\max} - R_{bs})C_{bs}}{R_{\max}^2}\end{aligned}$$

Now, let's upper bound Term 2. Before, recall that the definition of $g(r) = \mathbb{P}_r(u^-(\mathcal{R}) > r))$ and note that by the definition of black swans, we have $u^-(\mathcal{R}) > \mathcal{R} + C_{bs}$ holds for $R \in [-R_{\max}, -R_{bs})$. Therefore, we can say for all $r \in [-R_{\max}, -R_{bs}), g(r) = 1$ holds. Therefore, for all $r \in [-R_{\max}, -R_{bs}]$, we have $w^-_\star(g(r)) - w^-(g(r)) = w^-_\star(1) - w^-(1) = 1 - 1 = 0$

$$\begin{aligned}\left|\int_{-R_{\max}}^{0} w^-_\star(g(r)) - w^-(g(r))dr\right| &= \left|\int_{-R_{\max}+C_{bs}}^{0} w^-_\star(g(r)) - w^-(g(r))dr\right| \\ &= \left|\int_{-R_{\max}+C_{bs}}^{0} w^-(g(r)) - w^-_\star(g(r))dr\right| \\ &\leq \left|\int_{-R_{\max}+C_{bs}}^{0} L^-g(r) - g(r)dr\right| \\ &= (L^- - 1)\left|\int_{-R_{\max}+C_{bs}}^{0} g(r)dr\right| \\ &\leq (L^- - 1)\cdot\frac{R_{\max} - C_{bs}}{2R_{\max}}\epsilon_{bs} \\ &= (L^- - 1)\left|\int_{-R_{\max}+C_{bs}}^{0} 1 - \mathbb{P}_r(U^-(\mathcal{R}) < r)dr\right| \\ &= (L^- - 1)\left|\int_{-R_{\max}+C_{bs}}^{0} 1 - \mathbb{P}_r(U^-(\mathcal{R}) < r)dr\right| \\ &= (L^- - 1)\left|\left((R_{\max} - C_{bs}) - \int_{-R_{\max}+C_{bs}}^{0} \mathbb{P}_r(u^-(\mathcal{R}) < r)dr\right)\right| \\ &= (L^- - 1)\left|\left((R_{\max} - C_{bs}) - \mathbb{E}_{\mathcal{R}\sim\mathbb{P}_r}\left[u^-(\mathcal{R})\mathbf{1}[-R_{\max} + C_{bs} < \mathcal{R} < 0]\right]\right)\right|\end{aligned} \quad (29)$$

Note that if $-R_{\max} + C_{bs} < -R_{bs}$, then

$$\mathbf{1}[-R_{\max}+C_{bs} < \mathcal{R} < 0]\cdot\mathbb{E}_{\mathcal{R}\sim\mathbb{P}_r}\left[u^-(\mathcal{R})\right] \geq \left(\frac{R_{\max} - C_{bs} - R_{bs}}{2R_{\max}}\epsilon_{bs}^{\min} + \frac{R_{bs}}{2R_{\max}}\epsilon_{bs}\right)u^-(-R_{\max}+C_{bs}) \quad (30)$$

and if $-R_{\max} + C_{bs} < -R_{bs}$, then

$$\mathbf{1}[-R_{\max} + C_{bs} < \mathcal{R} < 0] \cdot \mathbb{E}_{\mathcal{R}\sim\mathbb{P}_r}\left[u^-(\mathcal{R})\right] \geq \left(\frac{R_{\max} - C_{bs}}{2R_{\max}}\epsilon_{bs}\right) u^-(-R_{\max} + C_{bs}) \tag{31}$$

Therefore, combining the Equations (29), (30), (31), we conclude that

$$\text{Term II} \leq C \cdot \frac{\left((R_{\max} - R_{bs})\,\epsilon_{bs}^{\min} - R_{bs}\epsilon_{bs}\right)(R_{\max} - R_{bs})C_{bs}}{R_{\max}^2}$$

where $C \in [0, 1]$ is a constant. This completes the proof.

3. Value function upper bound

For the proof of Equation (15) of Theorem 4, we utilized the following Lemma 4 which provides a concentration inequality on the distance between empirical distribution and true distribution.

Since $u^+(\mathcal{R})$ is bounded above by $u^+(R_{\max})$ and $w^+(p)$ is Lipschitz with constant $L^+(= (w^+)'(a))$, we have the following inequality,

$$\begin{aligned}
&\left|\int_0^\infty w^+(P(u^+(X)) > x)dx - \int_0^\infty w^+(1 - \hat{F}_t^+(x))dx\right| \\
=&\left|\int_0^{u^+(R_{\max})} w^+(P(u^+(X)) > x)dx - \int_0^{u^+(R_{\max})} w^+(1 - \hat{F}_t^+(x))dx\right| \\
\leq&\left|\int_0^{u^+(R_{\max})} L^+ \cdot |P(u^+(X) < x) - \hat{F}_t^+(x)|dx\right| \\
\leq& L^+ u^+(R_{\max}) \sup_{x\in\mathbb{R}} \left|P(u^+(X) < x) - \hat{F}_t^+(x)\right|.
\end{aligned}$$

Now, plugging in the DKW inequality, we obtain

$$\begin{aligned}
&P\left(\left|\int_0^\infty w^+(P(u^+(X)) > x)dx - \int_0^\infty w^+(1 - \hat{F}_t^+(x))dx\right| > \epsilon/2\right) \\
&\leq P\left(L^+ u^+(R_{\max}) \sup_{x\in\mathbb{R}} \left|(P(u^+(X) < x) - \hat{F}_t^+(x)\right| > \epsilon/2\right) \leq 2e^{-t\frac{\epsilon^2}{2(L^+u^+(R_{\max}))^2}}.
\end{aligned} \tag{32}$$

Along similar manner, we have

$$P\left(\left|\int_0^\infty w^-(P(u^-(X)) > x)dx - \int_0^\infty w^-(1 - \hat{F}_t^-(x))dx\right| > \epsilon/2\right) \leq 2e^{-t\frac{\epsilon^2}{2(L^-u^-(-R_{\max}))^2)}}. \tag{33}$$

Combining (32) and (33), we obtain

$$\begin{aligned}
P(|\widehat{V}_t^\dagger - V_t^\dagger| > \epsilon) \leq\ & P\left(\left|\int_0^\infty w^+(P(u^+(X)) > x)dx - \int_0^\infty w^+(1 - \hat{F}_t^+(x))dx\right| > \epsilon/2\right) \\
&+ P\left(\left|\int_0^\infty w^-(P(u^-(X)) > x)dx - \int_0^\infty w^-(1 - \hat{F}_t^-(x))dx\right| > \epsilon/2\right) \\
&\leq 4e^{-t\frac{\epsilon^2}{2c^2}}.
\end{aligned}$$

where $c = \max\{|L^+u^+(R_{\max})|, |L^-u^-(-R_{\max})|\}$

□ □

Proof of Theorem 5. For a given optimal policy $\pi_\star$, define the normalized occupancy measure as $d_{\pi_\star} = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}_\pi((s_t,a_t)=(s,a))$. Note that $d_{\pi_\star}$ represents the stationary distribution. Additionally, given the assumption that the reward function $R : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ is a bijection, it follows that the distribution $d_{\pi_\star}(R^{-1}(s,a))$ and $\mathbb{P}_r$ are identical. This indicates that the occurrence of black swan events can be entirely characterized by the reward values, rather than the specific state-action pairs.

Now, we define the event $E_{bs} := \{\mathcal{R}\in[-R_{\max}, -R_{bs}]\}$ where $\mathcal{R}\sim\mathbb{P}_r$. The probability of event E_{bs} happens is bounded as follows

$$
\begin{aligned}
\mathbb{P}(E_{bs}) &= F(-R_{bs}) - F(-R_{\max}) \\
&= F(-R_{bs}) \\
&\in \left(\left(\frac{R_{\max}-R_{bs}}{2R_{\max}}\right)\epsilon_{bs}^{\min}, \left(\frac{R_{\max}-R_{bs}}{2R_{\max}}\right)\epsilon_{bs}^{\max}\right) \\
&:= [p_{bs}^{\min}, p_{bs}^{\max}]
\end{aligned}
$$

Note that we have assumed the $0 < \mathbb{P}_r(r = R(s,a)) < \epsilon_{bs}$ and its minimum reachable probability as $\epsilon_{bs}^{\min}$ for all reward. now, for given trajectory, the reward instance is given as $(r_1, r_2, ...r_h, ...)$ where $r_h\sim\mathbb{P}_r$, the probability that the agent first visit the black swan event at step h would be defined as

$$
\begin{aligned}
\mathbb{P}\left(r_1,\cdots,r_{h-1}\notin E_{bs}, r_h\in E_{bs}\right) &= (1-\mathbb{P}(E_{bs}))^{h-1}\mathbb{P}(E_{bs}) \\
&\le (1-p_{\min})^{h-1}p_{\max}
\end{aligned}
$$

Therefore, its probability is bounded as follows,

$$
(1-p_{\max})^{h-1}p_{\min} \le \mathbb{P}\left(r_1,\cdots,r_{h-1}\notin E_{bs}, r_h\in E_{bs}\right) \le (1-p_{\min})^{h-1}p_{\max}
$$

Now, to ensure that the blackswan probability to be lower bounded than δ, we need the following conditions,

$$
\begin{aligned}
\delta &\le (1-p_{\max})^{h-1}p_{\min} \\
\log\delta &\le (h-1)\log(1-p_{\max}) + \log p_{\min}
\end{aligned}
$$

Therefore, we have

$$
h \ge \log(\delta/p_{\min})/\log(1-p_{max}) + 1.
$$

Therefore, we can conclude that if $h = \Omega(\log(\delta/p_{\min})/\log(1-p_{max}))$, then the agent's probability to meet the black swan is at least δ.

□ □

Appendix B. Helpful Lemmas

Lemma 4. ***(Dvoretzky-Kiefer-Wolfowitz (DKW) inequality)***
Let $\hat{F}_n(u) = \frac{1}{n}\sum_{i=1}^{n} 1_{((u(X_i))\le u)}$ denote the empirical distribution of a r.v. U, with $u(X_1), \dots, u(X_n)$ being sampled from the r.v $u(X)$. The, for any n and $\epsilon > 0$, we have

$$P(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| > \epsilon) \le 2e^{-2n\epsilon^2}.$$

References

D. Abel, A. Barreto, B. Van Roy, D. Precup, H. P. van Hasselt, and S. Singh. A definition of continual reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

A. Agarwal, S. Kakade, and L. F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

S. Agrawal, S. K. Juneja, and W. M. Koolen. Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, pages 26–62. PMLR, 2021.

J. A. Bagnell, A. Y. Ng, and J. G. Schneider. Solving uncertain markov decision processes. 2001.

X. Barandiaran, E. D. Paolo, and M. Rohde. Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5):367–386, 2009.

BetterUp. The availability heuristic. https://www.betterup.com/blog/the-availability-heuristic, 2022. Accessed: 2024-05-12.

M. Bowling, J. D. Martin, D. Abel, and W. Dabney. Settling the reward hypothesis. In *International Conference on Machine Learning*, pages 3003–3020. PMLR, 2023.

B. O. Bradley and M. S. Taqqu. Financial risk and heavy tails. In *Handbook of heavy tailed distributions in finance*, pages 35–103. Elsevier, 2003.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

J. C. Carr, P. Panangaden, and D. Precup. Conditions on preference relations that guarantee the existence of optimal policies. In *International Conference on Artificial Intelligence and Statistics*, pages 3916–3924. PMLR, 2024.

K.-J. Chung and M. J. Sobel. Discounted mdp's: distribution functions and exponential utility maximization. *Siam Journal on Control and Optimization*, 25:49–62, 1987. URL https://api.semanticscholar.org/CorpusID:119760011.

S. P. Coraluppi. *Optimal control of Markov decision processes for performance and robustness*. University of Maryland, College Park, 1997.

S. P. Coraluppi and S. I. Marcus. Risk-sensitive and minimax control of discrete-time, finite-state markov decision processes. *Automatica*, 35(2):301–309, 1999.

S. P. Coraluppi and S. I. Marcus. Mixed risk-neutral/minimax control of discrete-time, finite-state markov decision processes. *IEEE Transactions on Automatic Control*, 45(3): 528–532, 2000.

D. Danis, P. Parmacek, D. Dunajsky, and B. Ramasubramanian. Multi-agent reinforcement learning with prospect theory. In *2023 Proceedings of the Conference on Control and its Applications (CT)*, pages 9–16. SIAM, 2023.

E. Delage and S. Mannor. Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213, 2010.

D. Di Castro, A. Tamar, and S. Mannor. Policy gradients with variance related risk criteria. *arXiv preprint arXiv:1206.6404*, 2012.

S. Dong, B. Van Roy, and Z. Zhou. Simple agent, complex environment: Efficient reinforcement learning with agent states. *Journal of Machine Learning Research*, 23(255): 1–54, 2022.

S. A. Drakopoulos and I. Theodossiou. Workers' risk underestimation and occupational health and safety regulation. *European Journal of Law and Economics*, 41:641–656, 2016.

M. Fard and J. Pineau. Pac-bayesian model selection for reinforcement learning. *Advances in Neural Information Processing Systems*, 23, 2010.

H. Fennema and P. Wakker. Original and cumulative prospect theory: A discussion of empirical differences. *Journal of Behavioral Decision Making*, 10(1):53–64, 1997.

M. J. Fleming and A. Sarkar. The failure resolution of lehman brothers. *Economic Policy Review, Forthcoming*, 2014.

J. Garcıa and F. Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

P. Geibel. Reinforcement learning for mdps with constraints. In *Machine Learning: ECML 2006: 17th European Conference on Machine Learning Berlin, Germany, September 18-22, 2006 Proceedings 17*, pages 646–653. Springer, 2006.

P. Geibel and F. Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24:81–108, 2005.

M. Gheshlaghi Azar, R. Munos, and H. J. Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91: 325–349, 2013.

T. Gilovich, D. Griffin, and D. Kahneman. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press, 2002.

A. Gosavi. Reinforcement learning for model building and variance-penalized control. In *Proceedings of the 2009 winter simulation conference (wsc)*, pages 373–379. IEEE, 2009.

M. Heger. Consideration of risk in reinforcement learning. In *Machine Learning Proceedings 1994*, pages 105–111. Elsevier, 1994.

D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

M. Housel. Penguin, 2023.

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

R. Huisman, K. G. Koedijk, and R. A. Pownall. Var-x: Fat tails in financial risk management. *Journal of risk*, 1(1):47–61, 1998.

G. N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30 (2):257–280, 2005.

N. Jiang. Pac reinforcement learning with an imperfect model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

N. Jiang. On value functions and the agent-environment boundary. *arXiv preprint arXiv:1905.13341*, 2019.

C. Jie, L. Prashanth, M. Fu, S. Marcus, and C. Szepesvári. Stochastic optimization in a cumulative prospect theory framework. *IEEE Transactions on Automatic Control*, 63 (9):2867–2882, 2018.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

E. Lecarpentier and E. Rachelson. Non-stationary markov decision processes, a worst-case approach using model-based reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

H. Lee, Y. Ding, J. Lee, M. Jin, J. Lavaei, and S. Sojoudi. Tempo adaptation in non-stationary reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

L. Leqi, A. Prasad, and P. K. Ravikumar. On human-aligned risk minimization. *Advances in Neural Information Processing Systems*, 32, 2019.

B. D. Martino, D. Kumaran, B. Seymour, and R. J. Dolan. Frames, biases, and rational decision-making in the human brain. *Science*, 313:684 – 687, 2006.

J. K. M. Mawutor. The failure of lehman brothers: causes, preventive measures and recommendations. *Research Journal of Finance and Accounting*, 5(4), 2014.

L. McDonald and P. Robinson. *A colossal failure of common sense: The incredible inside story of the collapse of Lehman Brothers*. Random House, 2009.

T. Moldovan and P. Abbeel. Risk aversion in markov decision processes via near optimal chernoff bounds. *Advances in neural information processing systems*, 25, 2012.

J. Neummann and O. Morgenstern. Theory of games and economic behaviour, 1944.

A. Nilim and L. El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

J. J. Opaluch and K. Segerson. Rational roots of "irrational" behavior: new theories of economic decision-making. *Northeastern Journal of Agricultural and Resource Economics*, 18(2):81–95, 1989.

L. Orseau and M. Ring. Space-time embedded intelligence. In *Artificial General Intelligence: 5th International Conference, AGI 2012, Oxford, UK, December 8-11, 2012. Proceedings 5*, pages 209–218. Springer, 2012.

T. Osogami. Robustness and risk-sensitivity in markov decision processes. *Advances in Neural Information Processing Systems*, 25, 2012.

B. Pandit, A. Albert, Y. Patil, and A. J. Al-Bayati. Impact of safety climate on hazard recognition and safety risk perception. *Safety science*, 113:44–53, 2019.

S. D. Patek. On terminating markov decision processes with a risk-averse objective function. *Automatica*, 37(9):1379–1386, 2001.

S. Pitis. Consistent aggregation of objectives with diverse time preferences requires non-markovian rewards. *Advances in Neural Information Processing Systems*, 36, 2024.

S. S. Ponda, L. B. Johnson, and J. P. How. Risk allocation strategies for distributed chance-constrained task allocation. In *2013 American Control Conference*, pages 3230–3236. IEEE, 2013.

L. Prashanth, C. Jie, M. Fu, S. Marcus, and C. Szepesvári. Cumulative prospect theory meets reinforcement learning: Prediction and control. In *International Conference on Machine Learning*, pages 1406–1415. PMLR, 2016.

M. Rabin. Risk aversion and expected-utility theory: A calibration theorem. In *Handbook of the fundamentals of financial decision making: Part I*, pages 241–252. World Scientific, 2013.

L. J. Ratliff and E. Mazumdar. Inverse risk-sensitive reinforcement learning. *IEEE Transactions on Automatic Control*, 65(3):1256–1263, 2019.

P. Rogers. The cognitive psychology of lottery gambling: A theoretical review. *Journal of gambling studies*, 14(2):111–134, 1998.

L. J. Savage. *The foundations of statistics*. Courier Corporation, 1972.

M. Shakerinava and S. Ravanbakhsh. Utility theory for sequential decision making. In *International Conference on Machine Learning*, pages 19616–19625. PMLR, 2022.

Y. Shen, M. J. Tobia, T. Sommer, and K. Obermayer. Risk-sensitive reinforcement learning. *Neural computation*, 26(7):1298–1328, 2014.

H. A. Simon. Decision making: Rational, nonrational, and irrational. *Educational administration quarterly*, 29(3):392–411, 1993.

R. Sugden. Rational choice: a survey of contributions from economics and philosophy. *The economic journal*, 101(407):751–785, 1991.

P. Sunehag and M. Hutter. Axioms for rational reinforcement learning. In *Algorithmic Learning Theory: 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings 22*, pages 338–352. Springer, 2011.

P. Sunehag and M. Hutter. Rationality, optimism and guarantees in general reinforcement learning. *The Journal of Machine Learning Research*, 16(1):1345–1390, 2015.

R. S. Sutton. The reward hypothesis, 2004. *URL http://incompleteideas. net/rlai. cs. ualberta. ca/RLAI/rewardhypothesis. html.*

R. S. Sutton. The quest for a common model of the intelligent decision maker. *arXiv preprint arXiv:2202.13252*, 2022.

H. A. Taha. Operations research an introduction. 2007.

N. N. Taleb. *The Black Swan:: The Impact of the Highly Improbable: With a new section:" On Robustness and Fragility"*, volume 2. Random house trade paperbacks, 2010.

A. M. Turing. *Computing machinery and intelligence.* Springer, 2009.

T. G. van der Meer, A. C. Kroon, and R. Vliegenthart. Do news media kill? how a biased news reality can overshadow real societal risks, the case of aviation and road traffic accidents. *Social forces*, 101(1):506–530, 2022.

P. Vasterman, C. J. Yzermans, and A. J. Dirkzwager. The role of the media and media hypes in the aftermath of disasters. *Epidemiologic reviews*, 27(1):107–114, 2005.

A. A. Wahlberg and L. Sjoberg. Risk perception and the media. *Journal of risk research*, 3(1):31–50, 2000.

G. Wheeler and G. Wheeler. A review of the lottery paradox. *Probability and inference: Essays in honour of Henry E. Kyburg, Jr*, pages 1–31, 2007.

W. Wiesemann, D. Kuhn, and B. Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

H. Xu and S. Mannor. Distributionally robust markov decision processes. *Advances in Neural Information Processing Systems*, 23, 2010.

Y. Zhang, X. Li, and S. Guo. Portfolio selection problems with markowitz's mean–variance framework: a review of literature. *Fuzzy Optimization and Decision Making*, 17:125–158, 2018.