A NOTATIONS

This section provides a table summarizing all the notations and their meanings introduced in the main paper.

Notation	Meaning	Defintion
$\mathcal{M}$	Ground truth MDP	Section 2
$\mathcal{S}$	State space	Section 2
$\mathcal{A}$	Action space	Section 2
T	Time horizon	Section 2
P	Transition probability function of $\mathcal{M}$	Section 2
R	Reward function of $\mathcal{M}$	Section 2
γ	Discount factor	Section 2
π	Policy	Section 2
P^{π}	Normalized visitation probability of $\mathcal{M}$	Section 2
$V^{\pi}_{\mathcal{M}}$	Value function of $\mathcal{M}$	Section 2
u, u^{-}, u^{+}	Value distortion function	Section 2 (Figure 1a)
w, w^{-}, w^{+}	Probability distorction function	Section 2 (Figure 1b)
$\mathcal{M}_d$	Distorted MDP	Section 4
$w(P)$	Transition probability of $\mathcal{M}_d$	Section 4
$u(R)$	Reward function of $\mathcal{M}_d$	Section 4
$\mathcal{M}^{\dagger}$	Human MDP (HMDP)	Subsection 5.1
$P^{\dagger}$	Transition probability of $\mathcal{M}^{\dagger}$	Subsection 5.1
$R^{\dagger}$	Reward function of $\mathcal{M}^{\dagger}$	Subsection 5.1
$P^{\dagger,\pi}$	Normalized visitation probability of $\mathcal{M}^{\dagger}$	Subsection 5.1
$\int P^{\pi}$	Cumulative visitation distribution of $\mathcal{M}$	Subsection 5.1
$\int P^{\dagger,\pi}$	Cumulative visitation distribution of $\mathcal{M}^{\dagger}$	Subsection 5.1
$V^{\pi}_{\mathcal{M}^{\dagger}}$	Value function of $\mathcal{M}^{\dagger}$	Subsection 5.1
ϵ_r, ϵ_d	Perception gap	Subsection 5.1
$\widehat{\mathcal{M}}^{\dagger}$	Human Estimation MDP (HEMDP)	Subsection 5.2
$\widehat{P}^{\dagger}$	Transition probability of $\widehat{\mathcal{M}}^{\dagger}$	Subsection 5.2
$\widehat{R}^{\dagger}$	Reward function of $\widehat{\mathcal{M}}^{\dagger}$	Subsection 5.2
$V^{\pi}_{\widehat{\mathcal{M}}^{\dagger}}$	Value function of $\widehat{\mathcal{M}}^{\dagger}$	Subsection 5.2
κ_r, κ_d	Estimation gap	Subsection 5.2
$R_{[\cdot]}, P^{\pi}_{[\cdot]}$	Order statistics of reward and visitation probability	Section 6
$\mathbb{P}_r$	Probability of reward	Section 6
$F(r)$	Cumulative distribution function of $\mathbb{P}_r$	Section 6
$\mathcal{B}$	Collection of all S-BLACK SWAN	Section 6

Notation	Meaning	Defintion
C_{bs}, ϵ_{bs}	The extent of distortion of functions u and w	Section 6 (Figures 1c and 1d)
$\epsilon_{bs}^{\min}$	Minimum probability of S-BLACK SWAN	Section 6
$u_\star^-$	u^- that satisfies $\mathcal{B} = \varnothing$, i.e. safe reward perception	Section 6
$w_\star^-$	w^- that satisfies $\mathcal{B} = \varnothing$, i.e. safe probability perception	Section 6

B 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. Turing first proposed the concept of a boundary between the agent and environment 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.

C RELATED WORKS: 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 C.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 C.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*.

C.1 DECISION MAKING UNDER RISK

Based on the comprehensive survey on safe reinforcement learning in Garcıa & 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 $\max_{\pi\in\Pi}\min_{w\in\mathcal{W}} V_{\mathcal{M}}^{\pi}(s;w)$, 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 & 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 ($\min_{w\in\mathcal{W}} V_{\mathcal{M}}^{\pi}(s;w)$) in order to maximize its income ($\max_{\pi\in\Pi}(\triangle)$), or maintaining constant health precautions such as wearing mask or maintaining distance with groups (π) to prepare for a possible pandemic ($\triangle = \min_{w\in\mathcal{W}} V_{\mathcal{M}}^{\pi}(s;w)$) 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 & El Ghaoui, 2005; Wiesemann et al., 2013; Xu & 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 ($\min_{win\mathcal{W}} V_{\mathcal{M}}^{\pi}(s;w)$), thus influencing asset allocation strategies ($\max_{\pi\in\Pi}\min_{w\in\mathcal{W}} V_{\mathcal{M}}^{\pi}(s;w)$) 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 & Matheson, 1972; Chung & Sobel, 1987; Patek, 2001). These algorithms optimize an alternative value function $V_{\mathcal{M}}^{\pi}(s) = \beta^{-1}\log\mathbb{E}_{\pi}[\exp^{\beta G}|P, s_0 = s]$, where β controls the desired level of risk and $G := \sum_{t=0}^{T}\gamma^t R(s_t, a_t)$ is a cumulative return. However, it is recognized that associating risk with the variance of the return is practical, as in $V_{\mathcal{M}}^{\pi}(s) = \beta^{-1}\log\mathbb{E}_{\pi}[\exp^{\beta G}] = \max_{\pi\in\Pi}\mathbb{E}_{\pi}[G] + \frac{\beta}{2}\text{var}(G) + \mathcal{O}(\beta^2)$, and the existence of black swan events does not significantly affect the returns of variance ($\text{var}(G)$) 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 & 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 $\text{Var}(G)$. 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 & 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 & 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 & 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 $h_i(G) \le \alpha_i$, for $i \in [N]$, where $h_i : \mathbb{R} \to \mathbb{R}$ is a function of return $G := \sum_{t=0}^{T}\gamma^t R(s_t, a_t)$ (Geibel, 2006). Typical constraints include ensuring the expectation of return exceeds a specific minimum threshold (α), such as $\mathbb{E}_{\pi}[G] \ge \alpha$, or softening these hard constraints by allowing a permissible probability of violation (ϵ), such as $\mathbb{P}(\mathbb{E}_\pi[G] \geq \alpha) \geq 1 - \epsilon$, known as chance-constraint (Delage & Mannor (2010); Ponda et al. (2013)). Constraints might also limit the return variance, such as $\text{Var}(G) \leq \alpha$ (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 3).

C.2 How irrationality relates with spatial black swans.

Before starting Subsection C.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 Ω where $|\Omega| = N$, through a *preference* relation in a *rational* manner. In von Neumann (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 $o_i \in \Omega$, 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 $\mathcal{L}$, 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 & Hutter (2015) extends rational choice theory to encompass the full reinforcement learning problem, further axiomatizing the concept in Sunehag & 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 & Ravanbakhsh (2022); Bowling et al. (2023), concentrate on the axioms of assigning utilities to all finite trajectories of a Markov Decision Process. Specifically, Shakerinava & 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 & 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 & Segerson (1989) explores subjective probabilities as an early modification to the expected utility model from von Neumann (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 & 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 & 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 & 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 & Wheeler (2007); BetterUp (2022)).

D CUMULATIVE PROSPECT THEOREM AND RISK

We note that existing works 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 & 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 & 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 ground truth MDP $\mathcal{M}$ is suggested as a way to achieve antifragility.

E PRELIMINARY FOR PROOFS

This subsection covers the preliminary concepts necessary for proving the theorems and lemmas presented in the paper.

First, in a discrete state and action space, the value function $\mathcal{M}$ could be expressed as an inner product of reward function R and normalized occupancy measure P^π as follows,

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

Based on Equations (5), (1), and (2), the *CPT* distorts the reward and its visitation probability as follows,

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

where † denotes the value function that 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 $w(\int P^\pi(s,a))$ into $u^-(R(s,a)), w^-(\int P^\pi(s,a))$ where $R(s,a) < 0$, and $u^+(R(s,a)), w^+(\int P^\pi(s,a))$ 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, the Equation (6) can be represented as follows:

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

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 rewrite the Equation (7) based on continuous state and actions space as follows.

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

We use the fact that for real-value function g, it holds that $\mathbb{E}[g(\mathcal{R})] = \int_0^\infty \Pr(g(\mathcal{R}) > r) dr$. Within the above problem setting, the agent's goal is to estimate the value function under safe perception $u^-_\star, w^-_\star$ as follows:

$$V_{\mathcal{M}}(s_0) = \int_0^\infty w^+\left(\mathbb{P}_r(u^+(X) > r)\right) dr - \int_0^\infty \boldsymbol{w}^-_\star\left(\mathbb{P}_r(\boldsymbol{u}^-_\star(X) > r)\right) dr \tag{9}$$

Note that the safe perception is only defined over w^- and u^- as $w^-_\star$ and $u^-_\star$. However, the agent possesses its own perceptions $\mathcal{M}^\dagger$, for which we assume the risk perception is represented as:

$$V_{\mathcal{M}^\dagger}(s_0) = \int_0^\infty w^+\left(\mathbb{P}_r(u^+(X) > r)\right) dr - \int_0^\infty \boldsymbol{w}^-\left(\mathbb{P}_r(\boldsymbol{u}^-(X) > r)\right) dr \tag{10}$$

As time goes by, the agent's goal is approximating 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. $\{h(s_i), a_i, u(r_i), h(s_{i+1})\}_{i=0}^t$ where the reward value itself and its sampling distribution are distorted due to the functions u and w, respectively (see Lemma 1 for definition of function h). Since function h maps state space to state space, we just use the notation $\{s'_i, a_i, u(r_i), s'_{i+1})\}_{i=0}^t$ to denote Let $r_i, i = 1, .., t$ denote n samples of the reward random variable X. We define the empirical distribution function (EDF) 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,

$$V_{\widehat{\mathcal{M}}^\dagger}(s_0) = \int_0^\infty w^+\left(1 - \hat{F}_t^+(r)\right)dr - \int_0^\infty w^-\left(1 - \hat{F}_t^-(r)\right)dr \tag{11}$$

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.

F Proofs

We first like to note that the following lemma helps to quantify how much the distortion on transition probability is related to the distortion on the visitation probability.

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

We begin with Lemma 3 to prove Lemma 2. Recall that $P^{\dagger,\pi}(s,a)$ is the ϵ_d-perceived visitation probability if $\max_{(s,a)}|P^\pi(s,a) - P^{\pi,\dagger}(s,a)| < \epsilon_d$. This perception gap arises from factors such as transition probabilities, policy, and state space. In the following lemma, we show how the perception gap in transition probability accumulates into the visitation probability. Before, we define ϵ_p-perceived transition probability if $\max_{(s,a)}||P(\cdot|s,a) - P^\dagger(\cdot|s,a)||_1 < \epsilon_p$ holds. We denote $\mathbb{P}_t^\pi(s,a)$ as the probability of visiting (s,a) at time t with policy π.

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

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

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

***Proof of Lemma** 3.* Proof by induction. We use short notation for $P(s_t = s \mid s_{t-1} = s', a_{t-1} = a')$ as $P_t(s \mid s', a')$ and $P^\dagger(s_t = s \mid s_{t-1} = s', a_{t-1} = a')$ as $P_t^\dagger(s \mid s', a')$. By the definition of rational transition probability the statement holds at $t = 1$ for any policy π. Now, suppose the statement holds for $t - 1$ for any policy π. Then, we have

$$\begin{aligned}
&\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|\mathbb{P}_t^{\pi}(s,a) - \mathbb{P}_t^{\pi,\dagger}(s,a)\right| \\
&\quad= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big|\pi(a_t = a \mid s_t = s) \sum_{s',a'} \left(P_t(s \mid s',a')\mathbb{P}_{t-1}^{\pi}(s',a')\right) \\
&\qquad - \pi(a_t = a \mid s_t = s) \sum_{s',a'} \left(P_t^{\dagger}(s \mid s',a')\mathbb{P}_{t-1}^{\pi,\dagger}(s',a')\right)\Big| \\
&\quad\le \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi(a_t = a \mid s_h = s)\Big| \sum_{s',a'} \left(P_t(s \mid s',a')\mathbb{P}_{t-1}^{\pi}(s',a')\right) - \sum_{s',a'} \left(P_t^{\dagger}(s \mid s',a')\mathbb{P}_{t-1}^{\pi,\dagger}(s',a')\right)\Big| \\
&\quad= \sum_{s\in\mathcal{S}} \Big| \sum_{s',a'} \left(P_t(s \mid s',a')\mathbb{P}_{t-1}^{\pi}(s',a')\right) - \sum_{s',a'} \left(P_t^{\dagger}(s \mid s',a')\mathbb{P}_{t-1}^{\pi,\dagger}(s',a')\right)\Big| \\
&\quad= \sum_{s\in\mathcal{S}} \Big| \sum_{s',a'} \left(P_t - P_t^{\dagger}\right)\mathbb{P}_{t-1}^{\pi}(s',a') + \sum_{s',a'} P_t^{\dagger}(s \mid s',a')\left(\mathbb{P}_{t-1}^{\pi}(s',a') - \mathbb{P}_{t-1}^{\pi,\dagger}(s',a')\right)\Big| \\
&\quad\le \sum_{s',a'} \Big| \sum_{s\in\mathcal{S}} \left(P_t - P_t^{\dagger}\right)\mathbb{P}_{t-1}^{\pi}(s',a')\Big| + \sum_{s',a'} \Big| \sum_{s\in\mathcal{S}} P_t^{\dagger}(s \mid s',a')\left(\mathbb{P}_{t-1}^{\pi}(s',a') - \mathbb{P}_{t-1}^{\pi,\dagger}(s',a')\right)\Big| \\
&\quad\le \epsilon_p \sum_{s',a'} \mathbb{P}_{t-1}^{\pi}(s',a') + 1\cdot(t-1)\epsilon_p \\
&\quad= \epsilon_p \cdot 1 + (t-1)\epsilon_p \\
&\quad\le t\epsilon_p
\end{aligned}$$

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

Now, we prove the Lemma 2.

Proof of Lemma 2. Lemma 2 is almost a corollary that stems from Lemma 3. By the definition of visitation probability, we have

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

Let $S = \sum_{t=0}^{\infty} \gamma^t t$, then $\gamma S = \sum_{t=0}^{\infty} \gamma^{t+1} t = \sum_{t=1}^{\infty} \gamma^t (t-1)$. Then by subtracting those two equations, we have $(1-\gamma)S = \sum_{t=1}^{\infty} \gamma^t = \frac{\gamma}{1-\gamma}$. 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 1. 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 h satisfies the set Z to in be permutation-invariant. Namely, if $R(s_1,a) = R(s_2,a)$, then $R(h(s_1)) = R(h(s_2),a)$ holds then there exists a one-to-one mapping function $h : [-R_{\max}, R_{\max}] \to [-R_{\max}, R_{\max}]$ such that

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

holds. The proof can be divided into two folds. The existence of such a function and its one-to-one mapping function exists. We first prove the existence of such function h. This is because for any state and action s,a, suppose its reward value is 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 $s_1' = h(s_1')$ and $s_2' = h(s_2')$. Now, suppose $R(s_1',a) \neq R(s_2',a)$ holds. Then, due to the property of h, then it should also satisfy $R(s_1,a) \neq R(s_2,a)$. Therefore, this concludes that h is the one-to-one mapping, and the following holds

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

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

$$w^-\left(\int_{-R_{max}}^{r} d\mathbb{P}_\mathcal{R}\right) = \int_{-R_{max}}^{r} d\mathbb{P}_{h^-(\mathcal{R})}, \quad \text{and} \quad w^+\left(\int_{-R_{max}}^{r} d\mathbb{P}_\mathcal{R}\right) = \int_{-R_{max}}^{r} d\mathbb{P}_{h^+(\mathcal{R})} \tag{12}$$

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 (12) 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 (12). This completes the proof. $\square$ $\square$

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 $a^\star$ is the argument that maximizes the $R(s,a)$. Since u is the non-decreasing convex function, $a^\star$ 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 $\pi^\star_{t'+1} = \pi^{\star,\dagger}_{t'+1}$ 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 $Q_t^{\pi^\star}(s,a) \ge Q_t^\pi(s,a')$, then $Q_t^{\dagger,\pi^\star}(s,a) \ge Q_t^{\dagger,\pi^\star}(s,a')$ 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

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

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 **??**. 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 $\pi^\star_{t'} = \pi^{\star,\dagger}_{t'}$ 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 $R_2(s_1) \le R_2(s_2) \le R_2(s_3)$. The condition $Q_1^{\dagger,\pi}(s,a_1) \ge Q_1^{\dagger,\pi}(s,a_2)$ 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 Definition 2. □

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

1. Proof of asymptotic convergence

We first prove Equation (3) of Theorem 4 in this part 1, then we prove Equation (4) of Theorem 4 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

$$\lim_{n\to+\infty}\sum_{i=1}^{n-1} u^+(R_{[i]})(w^+(\frac{n-i+1}{n})-w^+(\frac{n-i}{n})) \xrightarrow{n\to\infty} \int_0^{+\infty} w^+(P(U>t))dt, \text{w.p. } 1 \tag{13}$$

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,

$$P\left(\left|\sum_{i=1}^{n-1} u^+(R_{[i]})\cdot\left(w^+\left(\left(\frac{n-i}{n}\right)-w^+\left(\frac{n-i-1}{n}\right)\right)-\sum_{i=1}^{n-1}\xi^+_{\frac{i}{n}}\cdot\left(w^+\left(\frac{n-i}{n}\right)-w^+\left(\frac{n-i-1}{n}\right)\right)\right)\right|>\epsilon\right), \tag{14}$$

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

$$
\begin{aligned}
&P(\left|\sum_{i=1}^{n-1} u^+(R_{[i]}) \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n}))\right| > \epsilon) \\
&\leq P(\bigcup_{i=1}^{n-1} \left\{\left|u^+(R_{[i]}) \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \xi^+_{\frac{i}{n}} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n}))\right| > \frac{\epsilon}{n}\right\}) \\
&\leq \sum_{i=1}^{n-1} P(\left|u^+(R_{[i]}) \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \xi^+_{\frac{i}{n}} \cdot (w^+_(\frac{n-i}{n}) - w^+_(\frac{n-i-1}{n}))\right| > \frac{\epsilon}{n}) \qquad (15) \\
&= \sum_{i=1}^{n-1} P(\left|(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}}) \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n}))\right| > \frac{\epsilon}{n}) \\
&\leq \sum_{i=1}^{n-1} P(\left|(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}}) \cdot (\frac{1}{n})^{\alpha}\right| > \frac{\epsilon}{n}) \\
&= \sum_{i=1}^{n-1} P(\left|(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}})\right| > \frac{\epsilon}{\cdot n^{1-\alpha}}). \qquad (16)
\end{aligned}
$$

The right-hand side of Inequality (16) could be expressed as follows.

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

We focus on the term $P\left(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}} > \frac{\epsilon}{n^{1-\alpha}}\right)$. 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,

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

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

$$
P(\sum_{i=1}^{n} A_t - n \cdot [1 - F^+(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}})] > n \cdot [F^+(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}) - \frac{i}{n}]) < e^{-2n \cdot \delta'_t}, \qquad (17)
$$

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

$$
P(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}} > \frac{\epsilon}{1-\alpha}) < e^{-2n \cdot L_{F^+} \frac{\epsilon}{1-\alpha}} = e^{-2n^{\alpha} \cdot L^+ \epsilon} \qquad (18)
$$

In a similar fashion, one can show that

$$
P(u^+(R_{[i]}) - \xi^+_{\frac{i}{n}} < -\frac{\epsilon}{1-\alpha}) \leq e^{-2n^{\alpha} \cdot L_{F^+} \epsilon} \qquad (19)
$$

Combining (18) and (19), 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 (16), we obtain

$$P(\left|\sum_{i=1}^{n-1} u^+(R_{[i]}) \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n}))\right| > \epsilon)$$
$$\leq 2n \cdot e^{-2n^{\alpha} \cdot L_{F^+}}. \tag{20}$$

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 (13).

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

$$\lim_{n \to +\infty} \sum_{i=1}^{n-1} u^-(R_{[i]})(w^-(\frac{n-i+1}{n}) - w^-(\frac{n-i}{n})) \xrightarrow{n \to \infty} \int_0^{+\infty} w^-(P(U > t))dt, \text{w.p. } 1 \tag{21}$$

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

2. Proof of value function lower bound

By the definition, we have the following

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

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,

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

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

$$\begin{aligned}
\text{Term I-1} &\geq \frac{R_{\max} - R_{bs}}{R_{\max}} \epsilon_{bs}^{\min} \min_{R \in [-R_{\max}, -R_{bs}]} \left| u^-(R) - u_\star^-(R) \right| \\
&\geq \frac{R_{\max} - R_{bs}}{R_{\max}} \epsilon_{bs}^{\min} \left| u^-(-R_{bs}) - u_\star^-(-R_{bs}) \right|
\end{aligned} \tag{22}$$

$$\begin{aligned}\text{Term I-2} &\leq \frac{R_{bs}}{R_{\max}}\epsilon_{bs} \max_{R\in[-R_{bs},0]} |u^-(R) - u_\star^-(R)| \\ &\leq \frac{R_{bs}}{R_{\max}}\epsilon_{bs} |u^-(-R_{bs}) - u_\star^-(-R_{bs})| \end{aligned} \tag{23}$$

Therefore, we have the following equation,

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

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

$$w_\star^-(g(r)) - w^-(g(r)) = w_\star^-(1) - w^-(1) = 1 - 1 = 0$$

$$\begin{aligned}
\left|\int_{-R_{\max}}^{0} w_\star^-(g(r)) - w^-(g(r))dr\right| &= \left|\int_{-R_{\max}+C_{bs}}^{0} w_\star^-(g(r)) - w^-(g(r))dr\right| \\
&= \left|\int_{-R_{\max}+C_{bs}}^{0} w^-(g(r)) - w_\star^-(g(r))dr\right| \\
&\leq \left|\int_{-R_{\max}+C_{bs}}^{0} L^- g(r) - g(r)dr\right| \\
&= (L^- - 1)\left|\int_{-R_{\max}+C_{bs}}^{0} g(r)dr\right| \\
&\leq (L^- - 1)\cdot\frac{R_{\max} - C_{bs}}{2R_{\max}}\epsilon_{bs} \\
&= (L^- - 1)\left|\int_{-R_{\max}+C_{bs}}^{0} 1 - \mathbb{P}_r(U^-(\mathcal{R}) < r)dr\right| \\
&= (L^- - 1)\left|\int_{-R_{\max}+C_{bs}}^{0} 1 - \mathbb{P}_r(U^-(\mathcal{R}) < r)dr\right| \\
&= (L^- - 1)\left|\left((R_{\max} - C_{bs}) - \int_{-R_{\max}+C_{bs}}^{0} \mathbb{P}_r(u^-(\mathcal{R}) < r)dr\right)\right| \\
&= (L^- - 1)\left|((R_{\max} - C_{bs}) - \mathbb{E}_{\mathcal{R}\sim\mathbb{P}_r}\left[u^-(\mathcal{R})\mathbf{1}[-R_{\max} + C_{bs} < \mathcal{R} < 0]\right])\right|
\end{aligned} \tag{24}$$

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

$$\mathbf{1}[-R_{\max}+C_{bs} < \mathcal{R} < 0]\cdot\mathbb{E}_{\mathcal{R}\sim\mathbb{P}_r}\left[u^-(\mathcal{R})\right] \geq \left(\frac{R_{\max} - C_{bs} - R_{bs}}{2R_{\max}}\epsilon_{bs}^{\min} + \frac{R_{bs}}{2R_{\max}}\epsilon_{bs}\right)u^-(-R_{\max}+C_{bs}) \tag{25}$$

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

$$\mathbf{1}[-R_{\max} + C_{bs} < \mathcal{R} < 0]\cdot\mathbb{E}_{\mathcal{R}\sim\mathbb{P}_r}\left[u^-(\mathcal{R})\right] \geq \left(\frac{R_{\max} - C_{bs}}{2R_{\max}}\epsilon_{bs}\right)u^-(-R_{\max} + C_{bs}) \tag{26}$$

Therefore, combining the Equations (24), (25), (26), 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 (4) 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

$$\begin{aligned}
&P\left(\left|\int_0^\infty w^+(P(u^+(X))>x)dx - \int_0^\infty w^+(1-\hat{F}_t^+(x))dx\right|>\epsilon/2\right) \\
&\leq P\left(L^+u^+(R_{\max})\sup_{x\in\mathbb{R}}\left|(P(u^+(X)<x)-\hat{F}_t^+(x)\right|>\epsilon/2\right)\leq 2e^{-t\frac{\epsilon^2}{2(L^+u^+(R_{\max}))^2}}.
\end{aligned} \tag{27}$$

Along similar manner, we have

$$P\left(\left|\int_0^\infty w^-(P(u^-(X))>x)dx - \int_0^\infty w^-(1-\hat{F}_t^-(x))dx\right|>\epsilon/2\right)\leq 2e^{-t\frac{\epsilon^2}{2(L^-u^-(-R_{\max}))^2)}}. \tag{28}$$

Combining (27) and (28), we obtain

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

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

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

Therefore, its probability is bounded as follows,

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

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

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

Therefore, we have

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

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

$\square$ $\square$

G Helpful Lemmas

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

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