Provably Efficient Partially Observable Risk-sensitive Reinforcement Learning with Hindsight Observation

Tonghe Zhang [*1] **Yu Chen** [*2] **Longbo Huang** [2]

Abstract

This work pioneers regret analysis of risk-sensitive reinforcement learning in partially observable environments with hindsight observation, addressing a gap in theoretical exploration. We introduce a novel formulation that integrates hindsight observations into a Partially Observable Markov Decision Process (POMDP) framework, where the goal is to optimize accumulated reward under the entropic risk measure. We develop the first provably efficient RL algorithm tailored for this setting. We also prove by rigorous analysis that our algorithm achieves polynomial regret $\tilde{O}\left(\frac{e^{|\gamma|H}-1}{|\gamma|H}H^2\sqrt{KHS^2OA}\right)$, which outperforms or matches existing upper bounds when the model degenerates to risk-neutral or fully observable settings. We adopt the method of change-of-measure and develop a novel analytical tool of beta vectors to streamline mathematical derivations. These techniques are of particular interest to the theoretical study of reinforcement learning.

1. Introduction

Reinforcement learning (RL) is a sequential decision-making problem in which an agent learns to maximize accumulated rewards through interactions with an unknown environment (Sutton & Barto, 2018). In many practical scenarios such as derivative hedging (Cao et al., 2020) and actuarial science (Richman, 2021), decision-makers have to consider the associated risks, leading to the study of risk-sensitive RL (Fei et al., 2020).

It is also a common practice to make costly decisions based on unreliable or incomplete information, such as in autonomous driving (Huang et al., 2019), stock market prediction (Kabbani & Duman, 2022), and cybersecurity (Vassilev et al., 2022). The Partially Observable Markov Decision Process (POMDP) (Monahan, 1982) is widely employed as the mathematical framework for these problems.

Empirical studies have been conducted on risk-sensitive POMDPs to address the planning or learning problem in various application scenarios (Skoglund, 2021; Choi et al., 2021; Qiu et al., 2021; Shen et al., 2023). However, these studies often lack a performance guarantee. On the other hand, theoretical studies primarily focus on demonstrating the existence of an optimal policy (Di Masi & Sthttner, 1999) or addressing the planning problem with full knowledge of the transition probabilities (James et al., 1994; Bäauerle & Rieder, 2017), but have yet to develop the sample complexity analysis. Consequently, an open theoretical question remains from prior research:

Can we devise a sample-efficient and theoretically grounded risk-sensitive RL algorithm in partially observable environments?

Obtaining a conclusive answer to this question is challenging due to several technical obstacles. Firstly, the complex structure of the POMDP becomes even more intricate when incorporating a non-linear risk measure, raising doubts about whether mathematical analysis alone can effectively simplify the problem. Secondly, partial observations pose challenges in learning the model with limited sample complexity and designing the exploration bonus with incomplete information.

In this work, we devise a novel algorithm that addresses these problems in a POMDP model equipped with hindsight observations. Due to the fact that reinforcement learning in general POMDPs is intractable (Papadimitriou & Tsitsiklis, 1987; Jin et al., 2020), we involve hindsight observations(Lee et al., 2023) in the learning protocol, which will be introduced in Section 3. We also excavate the dynamic programming structure implicit in the risk-sensitive POMDP model and subsequently derive a fresh set of Bellman equations that matched our problem setting. Moreover, we uncover a simple representation of the value functions with explicit analytical forms, leading to the creation of a

[*]Equal contribution [1]Department of Electronic Engineering, Tsinghua University, Beijing, China [2]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. Correspondence to: Longbo Huang <longbohuang@tsinghua.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

new exploration bonus that exploits the partial information gleaned from the environment while considering the agent's risk sensitivity. Our algorithm efficiently estimates the accumulated risks in all the possible hidden states and selects the optimal actions.

In addition, we provide theoretical guarantee for the algorithm design: the upper bound of the regret presented in Eq.(23) is not only polynomial in all the parameters but also explains how the risk measure, partial observations and empirical estimator affects the learning efficiency of each component of the POMDP model. Moreover, when the model degenerates to risk-neutral or fully observable settings, our result improves or matches existing upper bounds and nearly reaches the lower bounds in these scenarios.

The contributions of this work are summarized as follows:

- *Formulation.* We propose a novel theoretical formulation for risk-sensitive reinforcement learning in a partially observable environment with hindsight observations. We adopt the entropic risk measure in the framework, which accommodates general underlying POMDP models with non-stationary decision process. We also generalize certain results to arbitrary utility risk measures.
- *Algorithm.* We develop a new algorithm that incorporates a belief propagation process before the value iteration begins. Our design allows the agent to preemptively estimate the accumulated risks within the hidden states, before she optimizes the value functions through a greatly simplified Bellman equation. We also introduce a new bonus function that exploits partial information to encourage risk-sensitive exploration.
- *Regret.* We provide the first regret analysis of the problem. By disregarding lower-order terms, our algorithm successfully attains the regret $O\left(\frac{e^{|\gamma|H}-1}{|\gamma|H} \cdot H^2\sqrt{KS^2OA} \cdot \sqrt{H\ln KHSOA/\delta}\right)$, which demonstrates the risk awareness of the agent and the history-dependency. When the model degenerates to risk-neutral or fully observable settings, our regret improves or matches existing upper bounds and nearly reaches the lower bound of risk-sensitive RL.
- *Techniques.* We introduce a novel analytical tool called the beta vector, which plays a pivotal role in designing our bonus function, resulting in simplified value iteration and regret analysis. We also adopt the change-of-measure technique, which decouples the state and observations to streamline analytical derivations.

2. Related Work

Due to space limits, we only discuss the most relevant works below. A thorough overview is provided in Appendix F.

Risk-sensitive RL. Our analysis draws inspiration from studies about risk-sensitive RL. For instance, (Fei et al., 2021a) introduces a new bonus to improve the regret bound of learning an MDP using entropic risk. Additionally, (Liang & Luo, 2023) uses the concept of Lipschitz continuity to linearize various risk measures. These works contribute to our technical toolkit.

POMDP. It is well-known that planning or learning a general POMDP is intractable (Papadimitriou & Tsitsiklis, 1987; Jin et al., 2020). Consequently, a body of research in partially observable RL restricts attention to subclasses of POMDP with structural assumptions (Golowich et al., 2022b; Zhan et al., 2022), such as the emission matrices having a minimum positive singular value, denoted by α (Liu et al., 2022a;b). In these subclasses of POMDPs, the emission process reveals enough information for the agent to decode the hidden states (Liu et al., 2022a; Cai et al., 2022; Golowich et al., 2022b). Our work significantly diverges by not imposing such structural constraints on the probability kernels. This foundational difference in formulation allows our regret bound to remain independent of α, which enlarges the scope of applicability to instances where emission kernels lack full column rank.

Risk-sensitive POMDP. Our theoretical framework builds on prior research such as (James et al., 1994; Bäauerle & Rieder, 2017). However, our study significantly diverges from the predecessors because we do not presuppose the transition and emission matrices are time-invariant and fully known. Our agent learns a non-stationary model through online interactions, focusing on the exploration-exploitation trade-off typical of reinforcement learning.

3. Problem Formulation

Notations In this study we use $\gtrsim$ and $\tilde{O}$ to hide problem-independent constants or logarithmic terms in the sample complexity. γ^- represents $\min\{\gamma, 0\}$ and γ^+ is for $\max\{\gamma, 0\}$, while ι is the shorthand for $\ln(KHSOA/\delta)$ and $N \vee 1$ is for $\max\{N, 1\}$. Random variables are in bold, while their realizations are in roman. For random vectors $\boldsymbol{v_A}$ and $\boldsymbol{v_B}$, we use $\boldsymbol{v_A} \setminus \boldsymbol{v_B}$ to denote the subvector of $\boldsymbol{v_A}$ with the components from $\boldsymbol{v_B}$ removed.

The POMDP Model We consider a tabular, episodic and finite horizon POMDP with non-stationary transition matrices (Monahan, 1982; Liu et al., 2022a). The underlying model of the POMDP can be specified by a tuple $\mathcal{P} = (\mathscr{S}, \mathscr{O}, \mathscr{A}; \mu_1, \mathbb{T}, \mathbb{O}; K, H, r)$, where $\mathscr{S}$, $\mathscr{O}$ and $\mathscr{A}$ are the spaces of the hidden states, observations, and actions with cardinality S, O, and A respectively. The agent plays with the model in K episodes and each episode contains H steps. $\mu_1 \in \Delta(\mathscr{S})$ is the prior distribution of the hid-

den states which can be represented as an S-dimensional vector $\vec{\mu}_1$. $\mathbb{T} = \{\mathbb{T}_{h,a} \in \mathbb{R}^{S\times S} \mid (h,a) \in [H]\times\mathscr{A}\}$ and $\mathbb{O} = \{\mathbb{O}_h \in \mathbb{R}^{O\times S} \mid h \in [H]\}$ are the transition and emission matrices respectively. If the environment is in state s at step h, then $\mathbb{T}_{h,a}(\cdot|s)$ represents the distribution of the next hidden state when the agent takes action a, while $\mathbb{O}_h(\cdot|s)$ is the distribution of the observations generated by the current hidden state. $r = \{\, r_h(\cdot,\cdot) : \mathscr{S}\times\mathscr{A} \to [0,1] \mid h \in [H]\,\}$ is the collection of reward functions that measure the performance of actions in each hidden state.

In each episode, the initial state $\boldsymbol{S}_1$ is sampled from $\mu_1(\cdot)$. For all steps $h \in [H]$, the agent decides an action $\boldsymbol{A_h}$ based on previous observations and receives a reward $r_h(s_h, a_h)$. The environment then transits to a new hidden state $\boldsymbol{S}_{h+1} \sim \mathbb{T}_{h,a_h}(\cdot|s_h)$ and emits an observation $\boldsymbol{O}_{h+1} \sim \mathbb{O}_{h+1}(\cdot|s_{h+1})$, after which the a new action will be taken. In a POMDP, the agent never detects the hidden states, so she makes decisions according to the observable history $\boldsymbol{F}_h := (\boldsymbol{A_1}, \boldsymbol{O_2}, \cdots, \boldsymbol{O_{h-1}}, \boldsymbol{A_{h-1}}, \boldsymbol{O_h}) \in \mathscr{F}_h$, from which her policy $\pi_h(\cdot)$ maps to the action space. [1]The POMDP evolves until the last state $\boldsymbol{S}_{H+1}$, before a new episode begins.

Reinforcement Learning with Hindsight Observation In RL, the agent plans under empirical models $\{\widehat{\mathcal{P}}^k\}_{k=1}^K$ learned from data samples of the history. The corresponding empirical distributions will be denoted as $\widehat{\mu}_1^k$, $\widehat{\mathbb{T}}^k$, and $\widehat{\mathbb{O}}^k$. In this work, we incorporate *hindsight observation* (Lee et al., 2023) in the interaction protocol of the POMDP: in test time, the agent is allowed to review the hidden states that occurred in the last H steps at the end of each episode. The concept of hindsight observability in RL is proposed by (Lee et al., 2023) and echoed by (Sinclair et al., 2023; Shi et al., 2023; Guo et al., 2023). This setting also makes efficient learning possible. According to (Lee et al., 2023), hindsight observations are common in real-world applications of the POMDP, such as data center scheduling (Sinclair et al., 2023) and online imitation learning (Ross et al., 2011).[2] Hindsight observability is also satisfied in asymmetric RL settings (Pinto et al., 2017; Baisero et al., 2022) and sim-to-real robotic learning tasks (Chen et al., 2020).

Reinforcement Learning using Entropic Risk Measure In risk-sensitive RL, the agent seeks an *optimal policy* $\pi^\star$ that maximizes the following *optimization objective*, which is the entropic risk measure of accumulated reward:

$$J(\pi;\mathcal{P},\gamma) := \frac{1}{\gamma}\ln\mathbb{E}_{\mathcal{P}}^{\pi}\left[e^{\gamma\sum_{t=1}^{H} r_t(\boldsymbol{S_t},\boldsymbol{A_t})}\right] \qquad (1)$$

where $\mathcal{P}$ is the POMDP model and $\gamma \neq 0$ is the given parameter of risk-sensitivity.

Remark 3.1. Following the convention of risk-sensitive RL (Bäuerle & Rieder, 2014; Liang & Luo, 2023), we adopt the entropic risk measure (Fei et al., 2020) to define the risk-sensitive objective function in (1). The connection between the exponential function and risk-awareness can be illustrated by the following expansion:

$$\frac{1}{\gamma}\ln\mathbb{E}e^{\gamma R} = \mathbb{E}R + \frac{\gamma}{2}\cdot\operatorname{Var}R + o(\operatorname{Var}R)$$

where $R = \sum_{h=1}^{H} r_h(\boldsymbol{S_h}, \boldsymbol{A_h})$ is the accumulated reward. This relation intuitively reveals a close connection between the value of the risk measure and the uncertainty (variance) in the agents' returns, implying that the agent is risk-seeking when $\gamma > 0$ while risk-averse when $\gamma < 0$ (Bäuerle & Rieder, 2014). There exists various other risk measures in the studies of risk-sensitive RL, such as Conditional Value-at-Risk (CVaR) (Du et al., 2022), Lipschitz risks (Liang & Luo, 2023) and utility risks (Bäuerle & Rieder, 2014). Many of the results in this study can be generalized to utility risks, and we defer the discussion to Appendix B.

Learning Objective In this study, we aim to devise an algorithm whose output policies $\{\widehat{\pi}^k\}_{k=1}^K$ minimizes the difference with $\pi^\star$, which will be measured by the *"regret"* defined below:

$$\operatorname{Regret}(K;\mathcal{P},\gamma) := \sum_{k=1}^{K} J(\pi^\star;\mathcal{P},\gamma) - J(\widehat{\pi}^k;\mathcal{P},\gamma)$$

We are also concerned with the *sample complexity* of the algorithm, which is the smallest episode number K that ensures $\frac{1}{K}\sum_{k=1}^{K} J(\widehat{\pi}^k,\mathcal{P},\gamma) \geq J(\pi^\star,\mathcal{P},\gamma) - \epsilon$ with probability at least $1-\delta$.

4. Value Function and the Bellman Equations

In the much more complex setting of POMDP, we need to design special value functions to simplify the intensive computation caused by the history dependency in the policies. However, as is shown in Remark B.27 in the appendix, a naive adaptation of the value functions in the POMDP literature (Monahan, 1982) fails to capture the nonlinear structure of a risk-POMDP. In this work, we introduce a new definition for value functions based on the studies of (Bäauerle & Rieder, 2017; James et al., 1994), which not only simplifies the analysis but also helps us derive a new set of Bellman equations tailored to our problem. We will use these concepts in our algorithm presented in Section 5.

Change of Measure To simplify the mathematical analysis, we adopt the technique called "change of measure" and investigate the risk-POMDP problem in a simpler model $\mathcal{P}'$

[1]We use the notation $\boldsymbol{F}$ because the histories constitute a filter process.

[2]Please refer to Section 3.2 and 5 in (Lee et al., 2023) for more examples.

(obtained via transformation), in which the observations $\boldsymbol{O}_t$ and hidden states $\boldsymbol{S_t}$ are independent.

The technique of change-of-measure originates from stochastic calculus (Oksendal, 2013) and is vastly adopted in derivative pricing (Baxter & Rennie, 1996) and filtering theory (Jazwinski, 2007). In the study of POMDP, this method is used to decouple the transition and emission processes (Baras & James, 1997; Fernandez-Gaucherand & Marcus, 1997), which will also facilitate the statistical complexity analysis using hindsight observations.

In this work, we refer to $\mathcal{P}'$ as the "reference model", whose rigorous definition is presented in Appendix B.1. The following analysis comes from the studies of (James et al., 1994; Cavazos-Cadena & Hernández-Hernández, 2005). The relationship between $\mathcal{P}'$ and the original POMDP can be described by their Radon-Nikodym derivative: [3]

$$\boldsymbol{D_h} := \frac{d\mathbb{P}^\pi_{\mathcal{P}}}{d\mathbb{P}^\pi_{\mathcal{P}'}} = \prod_{t=2}^{h} \frac{\mathbb{O}_t(\mathbf{O}_t|\mathbf{S}_t)}{\mathbb{O}'(\mathbf{O}_t)}$$

In the reference model, the observations are irrelevant to the hidden states. As a result, they are separate from the underlying process and independent of the history. To further simplify the model, we can specify $\mathbb{O}'(\cdot)$ as the uniform distribution, so that $\boldsymbol{O}_h \overset{i.i.d.}{\sim} \text{Unif}\mathscr{O}$ in the model $\mathcal{P}'$.[4] Consequently, we can significantly simplify the posterior distribution of $\mathbf{O}_t$ in model $\mathcal{P}'$, which not only reduces space consumption dramatically but also reduces a series of analytical computations. After planning in $\mathcal{P}'$, we will use $\boldsymbol{D_h}$ as a bridge to convert the results back to the environment $\mathcal{P}$, according to the following rule derived from the Lebesgue-Radon-Nikodyn theorem (Rankin, 1968):

$$\mathbb{E}^\pi_{\mathcal{P}} = \mathbb{E}^\pi_{\mathcal{P}'}\boldsymbol{D_h}$$

We define our value functions in the model $\mathcal{P}'$.

Definition 4.1. (Value functions)

$$\begin{aligned}\mathsf{V}^\pi_h(\boldsymbol{F}_h) :=& \frac{1}{\gamma}\ln \mathbb{E}^\pi_{\mathcal{P}'}\left[e^{\gamma\sum_{t=1}^H r_t(\boldsymbol{S}_t,\boldsymbol{A}_t)}\middle|\boldsymbol{F}_h\right]\\ \mathsf{Q}^\pi_h(\boldsymbol{F}_h,\boldsymbol{A}_h) :=& \frac{1}{\gamma}\ln \mathbb{E}^\pi_{\mathcal{P}'}\left[e^{\gamma\sum_{t=1}^H r_t(\boldsymbol{S}_t,\boldsymbol{A}_t)}\middle|\boldsymbol{F}_h,\boldsymbol{A}_h\right]\end{aligned} \tag{2}$$

In a POMDP, the policy passes the history-dependency down to the value functions, whose variables contain $\boldsymbol{F_h}$ but not $\boldsymbol{S_h}$, since the latter is not even observable. We can also derive the Bellman equation for our problem:

[3] Please refer to Appendix G.1 for a formal definition.

[4] Other distributions are also suitable for $\mathbb{O}'$. Please refer to Appendix G.1.

Definition 4.2. (Bellman equations)

$$\begin{cases}\mathsf{V}^\pi_{H+1}(f_{H+1}) =& \frac{1}{\gamma}\ln \mathbb{E}^\pi_{\mathcal{P}'}\left[e^{\gamma\sum_{t=1}^H r_t(\boldsymbol{S}_t,\boldsymbol{A}_t)}\middle|f_{H+1}\right]\\ \mathsf{Q}^\pi_h(f_h,a_h) =& \frac{1}{\gamma}\ln \mathbb{E}_{\mathbf{O}_{h+1}\sim\text{Unif}\mathscr{O}}\left[e^{\gamma\mathsf{V}^\pi_{h+1}(f_h,a_h,\mathbf{O}_{h+1})}\right]\\ \mathsf{V}^\pi_h(f_h) =& \mathbb{E}_{\mathbf{A_h}\sim\pi_h(\cdot|f_h)}\mathsf{Q}^\pi_h(f_h,\mathbf{A}_h)\end{cases} \tag{3}$$

The Bellman equations in Eq.(3) are novel in the literature. Specifically, the value function V^π_{H+1} is not zero, which necessitates a belief propagation in the algorithm to initiate the value iteration. Moreover, the computation of Q function is greatly simplified since it is defined in the reference model. For the derivation of Eq.(3) and a generalization to arbitrary utility functions, please refer to AppendixB.4.

5. Algorithm Design

In the subsequent section, we introduce a sample-efficient algorithm tailored for solving the partially observable risk-sensitive RL problem. Algorithm 1 presents a concise overview, while a comprehensive description can be found in Appendix C. We name the algorithm Beta Vector Value Iteration (BVVI).

Algorithm 1 Beta Vector Value Iteration (BVVI)

Input K, H, risk level $\gamma \neq 0$, confidence $\delta \in (0,1)$
Initialize $\widehat{\mu}^1_1(\cdot), \widehat{\mathbb{T}}^1_{h,a}(\cdot|s) \leftarrow \text{Unif}(\mathscr{S})$
Initialize $\widehat{\mathbb{O}}^1_h(\cdot|s) \leftarrow \text{Unif}(\mathscr{O})$
for $k = 1 : K$ **do**
  $\widehat{\sigma}^k_1 \leftarrow \widehat{\mu}^k_1$ ▷ Belief propagation
  for $h = 1 : H$ **do**
    Update risk belief $\widehat{\sigma}^k_{h+1,f_{h+1}}$ by Eq.(9)
    Update bonus function $b^k_h(s_h,a_h)$ by Eq.(4)
  end for
  $\widehat{\beta}^k_{H+1,f_{H+1}} \leftarrow \vec{1}_S$ ▷ Value iteration
  for $h = H : 1$ **do**
    $\widehat{\mathsf{Q}}^k_h(f_h,a_h) \leftarrow \frac{1}{\gamma}\ln \mathbb{E}_{o_{h+1}\sim\text{Unif}\mathscr{O}}\langle\widehat{\sigma}^k_{h+1},\widehat{\beta}^{k,\widehat{\pi}^k}_{h+1}\rangle$
    $\widehat{\mathsf{V}}^k_h(f_h) \leftarrow \max_{a\in\mathscr{A}} \widehat{\mathsf{Q}}^k_h(f_h,a)$
    $\widehat{\pi}^k_h(f_h) \leftarrow \operatorname{argmax}_{a\in\mathscr{A}} \widehat{\mathsf{Q}}^k_h(f_h,a)$
    Update beta vector $\widehat{\beta}^{k,\widehat{\pi}^k}_{h,f_h}$ by Eq.(13)
    Restrict $\widehat{\beta}^{k,\widehat{\pi}^k}_{h,f_h}$ in $\left[e^{\gamma^-(H-h+1)}, e^{\gamma^+(H-h+1)}\right]$
  end for ▷ Statistical Learning
  Play with $\mathcal{P}$ under $\widehat{\pi}^k$ while gleaning $\{\widehat{o}^k_t,\widehat{a}^k_t\}$
  Review hidden states $\{\widehat{s}^k_t\}^{H+1}_{t=1}$ *in hindsight*.
  $\widehat{N}^{k+1}_h(s,a) \leftarrow \sum_{\kappa=1}^k \mathbb{1}\{\hat{s}^\kappa_h = s, \hat{a}^\kappa_h = a\}$
  $\widehat{N}^{k+1}_h(s) \leftarrow \sum_{\kappa=1}^k \mathbb{1}\{\hat{s}^\kappa_h = s\}$
  Update $\widehat{\mu}^k_1(\cdot), \widehat{\mathbb{T}}^{k+1}_{h,a}(\cdot|s), \widehat{\mathbb{O}}^{k+1}_h(\cdot|s)$ by Eq.(6)
end for

BVVI is a UCB algorithm (Lattimore & Szepesvári, 2020) that encourages explorations by a bonus function:

$$\mathsf{b}_h^k(s_h,a_h) = \left|e^{\gamma(H-h+1)}-1\right| \cdot \min\{1, \mathsf{t}_h^k(s_h,a_h) + \sum_{s'} \widehat{\mathbb{T}}_{h,a_h}^k(s'|s_h)\mathsf{o}_{h+1}^k(s')\} \quad (4)$$

where $\mathsf{t}_h^k(s_h,a_h)$ and $\mathsf{o}_{h+1}^k(s')$ are defined as

$$\begin{aligned}\mathsf{t}_h^k(s_h,a_h) &:= \min\left\{1, 3\sqrt{\frac{SH\cdot\iota}{\widehat{N}_h^k(s_h,a_h)\vee 1}}\right\} \\ \mathsf{o}_{h+1}^k(s_{h+1}) &:= \min\left\{1, 3\sqrt{\frac{OH\cdot\iota}{\widehat{N}_{h+1}^{k+1}(s_{h+1})\vee 1}}\right\}\end{aligned} \quad (5)$$

In BVVI, the agent updates the empirical model using hindsight observations:

$$\begin{aligned}\widehat{\mu}_1^{k+1}(s) &\leftarrow \sum_{\kappa=1}^{k} \frac{\mathbb{1}\{\hat{s}_1^\kappa = s\}}{k} \\ \widehat{\mathbb{T}}_{h,a}^{k+1}(s'|s) &\leftarrow \sum_{\kappa=1}^{k} \frac{\mathbb{1}\{\hat{s}_{h+1}^\kappa = s',\ \hat{s}_h^\kappa = s,\ \hat{a}_h^\kappa = a\}}{\widehat{N}_h^{k+1}(s,a)\vee 1} \\ \widehat{\mathbb{O}}_h^{k+1}(o|s) &\leftarrow \sum_{\kappa=1}^{k} \frac{\mathbb{1}\{\hat{o}_h^\kappa = o,\ \hat{s}_h^\kappa = s\}}{\widehat{N}_h^{k+1}(s)\vee 1}\end{aligned} \quad (6)$$

In the planning phase, the agent preemptively estimates the accumulated risks across the hidden states in a belief propagation process (Line 1), before she plans for the optimal policy in the empirical model, according to the Bellman equations provided in Eq.(3). To simplify computations, we express the value function with the help of a set of new variables $\widehat{\sigma}_{h,f_h}^k$ and $\widehat{\beta}_{h,f_h}^{k,\widehat{\pi}^k}$, which will be introduced later in Section 7. BVVI provides exact solution to the planning problem so assignments to the functions traverse their domains.

6. Main Results

In this section, we present the theoretical guarantee for algorithm 1. The proofs will be overviewed in Section 7.

Theorem 6.1. *(Regret) With probability at least $1-4\delta$, algorithm 1 achieves the following regret upper bound:*

$$\mathcal{O}\Bigg(\underbrace{\frac{e^{|\gamma|H}-1}{|\gamma|\,H}}_{\text{risk awareness}}\ \underbrace{H^2\sqrt{KS^2AO}}_{\text{Statistical error}}\ \underbrace{\sqrt{H\ln\frac{KHSOA}{\delta}}}_{\text{History-dependency}}\Bigg) \quad (7)$$

The detailed expression of the regret bound is presented in Eq.(23). Using the online-to-PAC conversion argument(Jin et al., 2018) we also obtain the sample complexity of BVVI:

Corollary 6.2. *(Sample complexity) For algorithm 1, the uniform mixture of its output policies ensures*

$$\mathbb{P}\left(V^\star - \sum_{k=1}^{K} V^{\widehat{\pi}^k} < \epsilon\right) \geq 1-\delta$$

when episode number K satisfies

$$K \gtrsim \frac{1}{\epsilon^2\delta^2}\left(\frac{e^{|\gamma|H}-1}{|\gamma|\,H}\right)^2 \cdot H^5S^2OA\cdot\ln\left(\frac{KHSOA}{\delta}\right)$$

Remark 6.3. The first factor in Eq.(7) reveals the risk awareness of the agent, which disappears in the risk-neutral scenario ($\gamma\to 0$). The second term demonstrates the statistical error brought by our UCB-style algorithm. The last factor represents how the inherent history-dependency of the POMDP discourages confidence in exploration, where the additional $\sqrt{H}$ comes from the uncertainty in the face of the large history space: $\sqrt{\ln\frac{|\mathscr{F}_h|}{\delta}} \leq \sqrt{\ln\frac{|OA|^H}{\delta}} \leq \sqrt{H\iota}$.

The mathematical analysis behind the main results will be presented in the following two sections.

7. Risk Belief and Beta Vector

In what follows we lay the foundation for comprehending Theorem 6.1 by the introduction of several essential concepts and lemmas.

A crucial element of the analysis is the risk belief vector, which plays a pivotal role in deriving novel value functions and Bellman equations presented in Eq.(4.1) and (4.2). Moreover, we will put forward the concept of beta vector, which is the cornerstone for the bonus design in Eq.(4) and the regret analysis presented in Section 8.

7.1. Risk Belief and the Bellman Equations

To capture the structural properties characteristic of our problem, we construct a *risk belief* that estimates the accumulated risk based on historical observations, which also excavates the dynamic programming structure implicit in the risk-sentive POMDP model.

Definition 7.1. (Risk Belief) (James et al., 1994)
For all $h\in[H+1]$, $f_h\in\mathscr{F}_h$, $s_h\in\mathscr{S}$, the risk beliefs are random vectors in $\mathbb{R}^S$, in which $\vec{\sigma}_1$ is defined as $\vec{\mu}_1$ and

$$[\vec{\sigma}_{h,\mathbf{F}_h}]_{s_h} := \mathbb{E}_{\mathcal{P}'}^{\pi}\left[\mathbf{D}_h\mathbb{1}\{\mathbf{S}_h=s_h\}e^{\gamma\sum_{t=1}^{h-1}r_t(\mathbf{S}_t,\mathbf{A}_t)}\middle|\mathbf{F}_h\right]$$

We can view the vector $\vec{\sigma}_{h,\mathbf{F}_h}$ as a list of cumulative risks estimated from the observable history, when the agent is in each of the hidden states. Using the risk beliefs, the optimization objective in Eq.(1) takes a simple form:

$$J(\pi;\mathcal{P},\gamma) = \frac{1}{\gamma}\ln\mathbb{E}_{\mathcal{P}'}^{\pi}\left[\langle\vec{\sigma}_{H+1,\mathbf{F}_{H+1}},\vec{1}_S\rangle\right] \quad (8)$$

With some derivation (cf. Theorem B.8), we can obtain the evolution law of the stochastic process $\vec{\sigma}_{h,\boldsymbol{F}_h}$:[5]

$$[\vec{\sigma}_{h+1,\boldsymbol{F}_{h+1}}]_{s_{h+1}} = \sum_{s_h} \mathbb{T}_{h,\boldsymbol{A_h}}(s_{h+1}|s_h) \cdot e^{\gamma r_h(s_h,\boldsymbol{A_h})} \frac{\mathbb{O}_{h+1}(\boldsymbol{O_{h+1}}|s_{h+1})}{\mathbb{O}'_{h+1}(\boldsymbol{O_{h+1}})} [\vec{\sigma}_{h,\boldsymbol{F_h}}]_{s_h} \quad (9)$$

which has a matrix representation in the tabular case: $\vec{\sigma}_{h+1,f_{h+1}} = \mathsf{U}_{a_h,o_{h+1}} \vec{\sigma}_{h,f_h}$. The inner product in Eq.(8) illuminates the existence of linear structure in $J(\pi;\mathcal{P},\gamma)$. Next, we will introduce another stochastic process, called the conjugate beliefs, which works in a concerted effort with the risk beliefs to expose the linearity within the optimization objective.

Definition 7.2. (Conjugate beliefs) (James et al., 1994) Let $\vec{\nu}_{H+1} := \vec{1}_S$. For all $h \in [H]$ and $\bar{\boldsymbol{F}}_h = \boldsymbol{F_{H+1}} \setminus \boldsymbol{F_h}$, the conjugate beliefs is a series of random vectors in $\mathbb{R}^S$ which is defined iteratively: $\vec{\nu}_{h,\bar{\boldsymbol{F}}_h} = \mathsf{U}^\top_{\boldsymbol{A_h},\boldsymbol{O_{h+1}}} \vec{\nu}_{h+1,\bar{\boldsymbol{F}}_{h+1}}$

The the update operator of the conjugate beliefs is the transpose of that of the risk belief, which immediately implies their inner product is invariant with time

$$\langle \vec{\sigma}_{H+1,f_{H+1}}, \vec{1} \rangle \equiv \langle \vec{\sigma}_{h,f_h}, \vec{\nu}_{h,\bar{f}_h} \rangle, \quad \forall h \in [H+1] \quad (10)$$

Consequently, we have

$$J(\pi;\mathcal{P},\gamma) \equiv \frac{1}{\gamma} \ln \mathbb{E}^\pi_{\mathcal{P}'} \left[\langle \vec{\sigma}_{h,\mathbf{F}_h}, \vec{\nu}_{h,\bar{\mathbf{F}}_h} \rangle \right], \forall h \in [H+1]$$

Motivated by the equation above we introduce the value functions and Q-functions in our problem setting, which take equivalent forms presented in Definition 4.1.

$$\begin{aligned} \mathsf{V}^\pi_h(\boldsymbol{F_h}) &:= \frac{1}{\gamma} \ln \mathbb{E}^\pi_{\mathcal{P}'} \left[\langle \vec{\sigma}_{h,\boldsymbol{F_h}}, \vec{\nu}_{h,\bar{\boldsymbol{F}}_h} \rangle \middle| \boldsymbol{F_h} \right] \\ \mathsf{Q}^\pi_h(\boldsymbol{F_h}, \boldsymbol{A_h}) &:= \frac{1}{\gamma} \ln \mathbb{E}_{\mathcal{P}'} \left[\langle \vec{\sigma}_{h,\boldsymbol{F_h}}, \vec{\nu}_{h,\bar{\boldsymbol{F}}_h} \rangle \middle| \boldsymbol{F_h}, \boldsymbol{A_h} \right] \end{aligned} \quad (11)$$

We can immediately obtain the Bellman equations introduced in Section 3 by the iterative expectation formula.

7.2. Beta Vector and the Bonus Design

To recover the Markov property of the value functions, in this work, we put forward the concept of beta vectors $\vec{\beta}^\pi_{h,f_h}$, and take it as the surrogate for $\mathsf{V}^\pi_h(f_h)$. Beta vectors utilize the hindsight observations to obtain a polynomial regret for our algorithm. We will study the statistical error of approximating a beta vector and then design a bonus that ensures optimism in the value functions and encourages greedy exploration.

[5]For the evolution law of the empirical beliefs $\widehat{\sigma}^k_{h+1}$, we replace the transition and emission matrices with their empirical approximations.

Definition 7.3. (Beta vector) The beta vectors of a risk-sensitive POMDP $\mathcal{P}$ with policy π is a series of random vectors in $\mathbb{R}^S$, which are specified by

$$\begin{aligned} \vec{\beta}^\pi_{H+1,\boldsymbol{F_{H+1}}} &:= \vec{1}_S \\ \vec{\beta}^\pi_{h,\boldsymbol{F_h}} &:= \mathbb{E}^\pi_{\mathcal{P}'} \left[\vec{\nu}_{h,\bar{\boldsymbol{F}}_h} | \boldsymbol{F_h} \right], \quad \forall 2 \le h \le H. \\ \vec{\beta}^\pi_1 &:= \mathbb{E}^\pi_{\mathcal{P}'} \left[\vec{\nu}_{1,\bar{\boldsymbol{F}}_1} \right] \end{aligned} \quad (12)$$

where $\vec{\nu}_{f,\bar{\boldsymbol{F}}_h}$ are the conjugate beliefs specified in Definition 7.2.

The beta vectors can be viewed as the risk-sensitive counterpart of the alpha-vector well-known in the POMDP literature (Pineau et al., 2006).[6] Similarly, they help to represent the value function in a simple form:

Theorem 7.4. *(Beta Vector Representation)* $\forall h \in [H+1], f_h \in \mathscr{F}_h$: $\mathsf{V}^\pi_h(f_h) = \frac{1}{\gamma} \ln \langle \vec{\sigma}_{h,f_h} \vec{\beta}^\pi_{h,f_h} \rangle$

As will be shown in Theorem B.26, the beta vector $\left[\vec{\beta}^\pi_{h,\boldsymbol{F_h}}\right]_{s_h}$ is equal to the following quantity:

$$\mathbb{E}_{\mathbf{A}_h \sim \pi_h(\cdot|\boldsymbol{F_h})} \Big[e^{\gamma r_h(s_h,\mathbf{A}_h)} \sum_{s_{h+1} \in \mathscr{S}} \mathbb{T}_{h,\mathbf{A}_h}(s_{h+1}|s_h) \sum_{o_{h+1} \in \mathscr{O}} \mathbb{O}_{h+1}(o_{h+1}|s_{h+1}) \left[\vec{\beta}^\pi_{h+1,f_{h+1}=(\boldsymbol{F_h},\mathbf{A}_h,o_{h+1})}\right]_{s_{h+1}} \Big]$$

where the expectation is computed with respect to probability measures that solely rely on the previous states s_h, rather than the entire history. We will see in Section 8 that the Markov property of the stochastic process $\{\beta^\pi_{h,\boldsymbol{F_h}}\}_{h \ge 1}$ will cooperate with the hindsight observations to secure a polynomial sample complexity for our algorithm. Consequently, we will focus on studying the beta vectors in the proceeding bonus design and regret analysis.

Bonus Function In reinforcement learning, we are concerned with the statistical error of the value functions brought by the inaccurate estimate of the environment $\mathcal{P}$. Using Theorem 7.4, it suffices to calculate the error that occurred in the evolution of the beta vectors. Based on concentration inequalities, detailed analysis in Appendix D.2.1 shows that the empirical error of beta vectors is controlled by the following bound with probability at least $1-2\delta$:

$$\begin{aligned} &\left| \left(\mathbb{E}\pi_{\widehat{\mathcal{P}}^k(\cdot,\cdot|s_h)} - \mathbb{E}^\pi_{\mathcal{P}(\cdot,\cdot|s_h)} \right) e^{\gamma r_h(s_h,\pi_h(f_h))} \beta^\pi_{h+1}(\cdot,\cdot;f_h) \right| \\ \le &\left| e^{\gamma(H-h+1)} - 1 \right| \cdot \left[\min\left\{1, 3\sqrt{\frac{SH\iota}{\widehat{N}^k_h(s_h,\pi_h(f_h)) \vee 1}}\right\} \right. \\ &\left. + \sum_{s'} \widehat{\mathbb{T}}^k_{h,\pi_h(f_h)}(s'|s_h) \min\left\{1, 3\sqrt{\frac{OH\iota}{\widehat{N}^{k+1}_{h+1}(s') \vee 1}}\right\} \right] \end{aligned}$$

[6]For a detailed comparison, please refer to Appendix B.27.

where we have temporarily viewed the beta vector as a binary function over $\mathscr{O} \times \mathscr{S}$. We refer to $\widehat{N}_{h+1}^{k+1}(s)$ and $\widehat{N}_{h+1}^{k+1}(s,a)$ as the occurrence frequencies of states and actions in the data samples obtained in the learning process.

We will abbreviate the two minimums in Eq.(5) as the *the transition error residue* $\mathsf{t}_h^k(s_h,a_h)$ and *the emission error residue* $\mathsf{o}_{h+1}^k(s_{h+1})$ respectively. With some minor adjustments, the upper bound in Eq.(5) will be used to define our bonus function for all the state-action pairs under the risk level γ, which is presented in Eq.(4).

Optimism In the algorithm, we have the freedom to design the empirical value function, as well as the empirical beliefs $\widehat{\sigma}_{h,f_h}^k$ and the beta vectors $\widehat{\beta}_{h,f_h}^{k,\pi}$. In this study, the empirical beliefs are determined by Eq.(9) and we define the empirical beta vectors by the iterative formulas below:

$$\begin{aligned}\left[\widehat{\beta}_{H+1}^{k,\pi}\right]_{s_1} :=& 1\\ \left[\widehat{\beta}_{h,f_h}^{k,\pi}\right]_{s_h} :=& \mathbb{E}_{a_h\sim\pi_h(\cdot|f_h)}\Big[e^{\gamma r_h(s_h,a_h)}\sum_{s'}\widehat{\mathbb{T}}_{h,a_h}^k(s'|s_h)\\ &\sum_{o'}\widehat{\mathbb{O}}_{h+1}^k(o'|s')\left[\widehat{\beta}_{h+1,f_{h+1}=(f_h,a_h,o')}^{k,\widehat{\pi}^k}\right]_{s'}+\\ &\operatorname{sgn}\gamma\cdot\mathsf{b}_h^k(s_h,a_h;\gamma)\Big]\end{aligned} \tag{13}$$

Finally, we mimic the representation Theorem 7.4 and construct the empirical value function by

$$\widehat{V}_h^\pi(f_h) := \frac{1}{\gamma}\ln\langle\widehat{\sigma}_{h,f_h}^k,\widehat{\beta}_{h,f_h}^{k,\pi}\rangle$$

As will be shown in Appendix D.2.2, the introduction of an additional bonus term in Eq.(13) will ensure that the value functions will be over-estimated in the empirical model:

Corollary 7.5. *(Optimism in value functions)*
For any risk-sensitivity level $\gamma \neq 0$ and episode number $k \in [K]$, we have $V_1^{\pi^\star} \leq \widehat{V}_1^{\widehat{\pi}^k}$, where $\widehat{\pi}^k$ is the optimal policy in $\widehat{\mathcal{P}}^k$.

8. Regret Analysis

We now give an overview of the proof of Theorem 6.1. Technical details are provided in Appendix D.

8.1. From Regret to Beta Vectors

With the help of Corollary 7.5 and Theorem 7.4, we can control the regret by the risk beliefs and the beta vectors:

$$\begin{aligned}&\operatorname{Regret}(K;\mathcal{P},\gamma)\\ \leq&\mathsf{K}_\gamma\sum_{k=1}^K\left|\langle\widehat{\sigma}_1^k-\vec{\sigma}_1,\widehat{\beta}_1^{k,\widehat{\pi}^k}\rangle\right|+\left|\mathbb{E}_{\mathcal{P}}\left(\widehat{\beta}_1^{k,\widehat{\pi}^k}-\vec{\beta}_1^{\widehat{\pi}^k}\right)\right|\end{aligned} \tag{14}$$

where K_γ is the Lipschitz constant of the entropic risk measure, which is $\frac{e^{(-\gamma)^+}}{\gamma}$ for all $\gamma \neq 0$.[7]

Using concentration inequalities G.5 and G.6, the first term in Eq.(14) is controlled by the following upper bound

$$\mathsf{K}_\gamma\sum_{k=1}^K\left|\langle\widehat{\sigma}_1^k-\vec{\sigma}_1,\widehat{\beta}_1^{k,\widehat{\pi}^k}\rangle\right| \leq \frac{e^{|\gamma|H}-1}{|\gamma|}\sqrt{2KS\ln\frac{K}{\delta}} \tag{15}$$

with probability at least $1-\delta$. We name the right-hand side of Eq.(15) as "the prior error," which arises from the inaccurate estimate of the prior distribution μ_1.

Next, our attention turns to bounding the second term in Eq.(14), which we refer to as the initial beta vector error and denote as Δ_1^k.

8.2. Control the Error Between Beta Vectors

We extend the definition of beta vector errors to $h \in [H+1]$:

$$\Delta_h^k := \left|\mathbb{E}_{\mathcal{P}}\left[\widehat{\beta}_{h,\boldsymbol{F_h}}^{k,\widehat{\pi}^k}(\mathbf{S}_h)-\vec{\beta}_{h,\boldsymbol{F_h}}^{\widehat{\pi}^k}(\mathbf{S}_h)\right]\right|$$

which demonstrate the average error of the beta vectors at step h incurred by the inaccurate empirical estimate of the POMDP model. We observe that [8]

$$\begin{aligned}\Delta_h^k =&\Big|\underbrace{\operatorname{sgn}(\gamma)\mathbb{E}_{\mathcal{P}}[\mathsf{b}_h^k]}_{\mathrm{I}}+\underbrace{\mathbb{E}_{\mathcal{P}}[(\widehat{\mathbb{T}}_h^k\widehat{\mathbb{O}}_{h+1}^k-\mathbb{T}_h\mathbb{O}_{h+1})e^{\gamma r_h}\vec{\beta}_{h+1}^{\widehat{\pi}^k}]}_{\mathrm{II}}\\ &+\underbrace{\mathbb{E}_{\mathcal{P}}[(\widehat{\mathbb{T}}_h^k\widehat{\mathbb{O}}_{h+1}^k-\mathbb{T}_h\mathbb{O}_{h+1})(e^{\gamma r_h}\widehat{\beta}_{h+1}^{\widehat{\pi}^k}-e^{\gamma r_h}\vec{\beta}_{h+1}^{\widehat{\pi}^k})]}_{\mathrm{III}}\\ &+\underbrace{\mathbb{E}_{\mathcal{P}}[\mathbb{T}_h\mathbb{O}_{h+1}(e^{\gamma r_h}\widehat{\beta}_{h+1}^{\widehat{\pi}^k}-e^{\gamma r_h}\vec{\beta}_{h+1}^{\widehat{\pi}^k})]}_{\mathrm{IV}}\Big|\end{aligned} \tag{16}$$

The terms in Eq.(16) are controlled by the concentration inequalities. In Appendix D.3.3, we show that

$$\Delta_h^k \leq e^{\gamma^+}\Delta_{h+1}^k + 4\mathbb{E}_{\mathcal{P}}\left[\mathsf{b}_h^k\right] \quad \forall h\in[H] \tag{17}$$

Then we invoke Lemma G.15 to obtain an upper bound on Δ_1^k based on Eq.(17). In the end, we prove that the sum of the initial beta vector error is dominated by the value of the bonus functions on the sampled trajectories and the bias in

[7] Please refer to Lemma G.14 for a detailed derivation.

[8] We have abbreviated the transition and emission matrices as operators in Eq.(16). We also omitted the variables of $\mathsf{b}_h^k(\cdot,\cdot)$.

the bonus function incurred by the empirical estimator:

$$\begin{aligned}\mathsf{K}_\gamma \sum_{k=1}^{K} \Delta_1^k < 4\mathsf{K}_\gamma \sum_{h=1}^{H} e^{\gamma^+(h-1)}.\\ \Bigg[\sum_{k=1}^{K} \underbrace{\mathsf{b}_h^k(\hat{s}_h^k, \hat{a}_h^k; \gamma)}_{\text{Bonus samples}} \\ + \sum_{k=1}^{K} \underbrace{\mathbb{E}_{\mathcal{P}} \mathsf{b}_h^k(\boldsymbol{S_h}, \boldsymbol{A_h}; \gamma) - \mathsf{b}_h^k(\hat{s}_h^k, \hat{a}_h^k; \gamma)}_{\text{Empirical bias}} \Bigg]\end{aligned} \tag{18}$$

We will first study the second summation. Concentration inequality G.9 implies that with probability at least $1-\delta$,

$$\begin{aligned}\sum_{k=1}^{K} \underbrace{\mathbb{E}_{\mathcal{P}} \mathsf{b}_h^k(\boldsymbol{S_h}, \boldsymbol{A_h}; \gamma) - \mathsf{b}_h^k(\hat{s}_h^k, \hat{a}_h^k; \gamma)}_{\text{Empirical bias}} \\ \leq \left|e^{\gamma(H-h+1)} - 1\right| \sqrt{K/2 \cdot \iota}\end{aligned} \tag{19}$$

A bound on the first term can also be derived from Eq.(4):

$$\begin{aligned}\sum_{k=1}^{K} \underbrace{\mathsf{b}_h^k(\hat{s}_h^k, \hat{a}_h^k; \gamma)}_{\text{Bonus samples}} \leq \left|e^{\gamma(H-h+1)} - 1\right| \\ \cdot \sum_{k=1}^{K} \mathsf{t}_h^k(\hat{s}_h^k, \hat{a}_h^k) + \sum_{s'} \widehat{\mathbb{T}}_{h,\hat{a}_h^k}^k(s'|\hat{s}_h^k) \mathsf{o}_{h+1}^k(s')\end{aligned} \tag{20}$$

To establish an upper bound for the summation in Eq.(20), we will telescope the equation twice and utilize the results from Lemmas G.6 and G.9 to bound the statistical error in the transition and emission processes. Subsequently, we demonstrate that with a probability of at least $1-2\delta$,

$$\begin{aligned}&\sum_{k=1}^{K} \mathsf{t}_h^k(\hat{s}_h^k, \hat{a}_h^k) + \sum_{s'} \widehat{\mathbb{T}}_{h,\hat{a}_h^k}^k(s'|\hat{s}_h^k) \mathsf{o}_{h+1}^k(s') \\ =&\sum_{k=1}^{K} \Big[\underbrace{\mathbb{T}_{h,\hat{a}_h^k}^k(\cdot|\hat{s}_h^k) \mathsf{o}_{h+1}^k(\cdot) - \mathsf{o}_{h+1}^k(\hat{s}_{h+1}^k)}_{\text{Concentration of MDS}} \\ &+ \underbrace{\mathsf{t}_h^k(\hat{s}_h^k, \hat{a}_h^k) + \mathsf{o}_{h+1}^k(\hat{s}_{h+1}^k)}_{\text{Residues}} \\ &+ \underbrace{\widehat{\mathbb{T}}_{h,\hat{a}_h^k}^k(\cdot|\hat{s}_h^k) \mathsf{o}_{h+1}^k(\cdot) - \mathbb{T}_{h,\hat{a}_h^k}^k(\cdot|\hat{s}_h^k) \mathsf{o}_{h+1}^k(\cdot)}_{\text{Azuma Hoeffding}}\Big] \\ \leq&\sqrt{2K \frac{\ln HSA}{\delta}} + 2\sum_{k=1}^{K} \mathsf{t}_h^k(\hat{s}_h^k, \hat{a}_h^k) + \mathsf{o}_{h+1}^k(\hat{s}_{h+1}^k)\end{aligned} \tag{21}$$

The relations above imply that the right-hand side of Eq.(18) is dominated by the transition and emission residues, which are defined in Eq.(5).

8.3. Sum Up the Residue

Finally, the pigeon-hole Lemma G.12 will help us compute the summation of the residue terms. Indeed,

$$\begin{aligned}\sum_{k=1}^{k} \mathsf{t}_h^k(\hat{s}_h^k, \hat{a}_h^k) &= \sum_{k=1}^{k} \min\left\{1,\ 3\sqrt{\frac{SH\ln\frac{KHSOA}{\delta}}{\widehat{N}_h^k(\hat{s}_h^k, \hat{a}_h^k) \vee 1)}}\right\} \\ &\leq \left(3\sqrt{SH\iota}\right) 2\sqrt{KSA}\end{aligned}$$

Similarly, we have

$$\sum_{k=1}^{K} \mathsf{o}_{h+1}^k(\hat{s}_{h+1}^k) \leq 3\sqrt{OH} \cdot \iota \cdot 2\sqrt{KS}$$

Bringing these relations back to Eqs.(21) and (20), we obtain

$$\begin{aligned}&\sum_{k=1}^{K} \mathsf{b}_h^k(\hat{s}_h^k, \hat{a}_h^k; \gamma) + \mathbb{E}_{\mathcal{P}}[\mathsf{b}_h^k(\boldsymbol{S_h}, \boldsymbol{A_h}; \gamma)] - \mathsf{b}_h^k(\hat{s}_h^k, \hat{a}_h^k; \gamma) \\ \leq&12 \cdot \underbrace{\left|e^{\gamma(H-h+1)} - 1\right|}_{\text{Bonus magnitude}} \cdot \sqrt{H} \cdot \underbrace{\sqrt{\ln\left(\frac{KHSOA}{\delta}\right)}}_{\text{History-dependency of POMDP}} \\ &\cdot \left(\underbrace{\sqrt{KS^2A}}_{\text{Hidden state error}} + \underbrace{\sqrt{KSO}}_{\text{Observation error}} + \underbrace{\sqrt{K}}_{\text{Empirical bias}}\right)\end{aligned} \tag{22}$$

The last step remaining is to take Eq.(22) back to Eq.(18) and then bring Eq.(18) with Eq.(15) to Eq.(14). Rearranging terms, we conclude that with probability at least $1-4\delta$,

$$\begin{aligned}\text{Regret}(K; \mathcal{P}, \gamma) \leq 48 \underbrace{\frac{e^{|\gamma|H} - 1}{|\gamma|}}_{\text{Risk measure}} \underbrace{\sqrt{H \cdot \ln\frac{KHSOA}{\delta}}}_{\text{History-dependency}} \\ \cdot \Big(\underbrace{\sqrt{KS}}_{\text{Prior error}} + \underbrace{H\sqrt{KS^2A}}_{\text{Transition error}} + \underbrace{H\sqrt{KSO}}_{\text{Emission error}} + \underbrace{H\sqrt{K}}_{\text{Empirical bias}}\Big)\end{aligned} \tag{23}$$

Neglecting lower order terms, we obtain the upper bound presented in Theorem 6.1:

$$\text{Regret}(K; \mathcal{P}, \gamma) \leq \tilde{\mathcal{O}}\left(\frac{e^{|\gamma|H} - 1}{|\gamma|\, H} H^{\frac{5}{2}} \sqrt{KS^2AO}\right)$$

Discussion In the risk-neutral setting, our regret improves the result given by (Lee et al., 2023) [9]

$$\begin{aligned}\text{Regret}(K; \mathcal{P}) \leq \tilde{O}\Big(&\sqrt{SAH^4K} + \boldsymbol{H^3S}\sqrt{O} \\ &+ \boldsymbol{H^4S^2A}(1 + \ln K) + H\boldsymbol{SA}\sqrt{H^3}\Big)\end{aligned}$$

in the order of S, A, and H. The improvement is attributed to the refined analysis in this work. Our sample complexity

[9] For details, please refer to Theorem C.1 of (Lee et al., 2023).

also nearly reaches the lower bound of learning a hindsight POMDP, which is $\Omega\left(\frac{SO}{\epsilon^2}\right)$ according to (Lee et al., 2023).

In the completely observable setting, with some adjustments, our algorithm can degenerate to the algorithm 1 in (Fei et al., 2021a) and thus matches their upper bound [10]

$$\text{Regret}(K;\mathcal{M},\gamma) \leq \tilde{O}\left(\frac{e^{|\gamma|H}-1}{|\gamma|H}\sqrt{KH^4S^2A}\right)$$

Moreover, our regret achieves the lower bound of risk-sensitive RL (Fei et al., 2020) concerning K and γH, with the order of H only slightly higher.

$$\text{Regret}(K;\mathcal{M},\gamma) \geq \tilde{O}\left(\frac{e^{|\gamma|\frac{H}{2}}-1}{|\gamma|H}\cdot H^{\frac{3}{2}}\sqrt{K}\right)$$

9. Experiments

In this section, we present experimental results to evaluate the empirical performance of our algorithm, BVVI, which serves as a reference to validate our theoretical findings in Section 8. For details of our experimental setup and more results, please refer to Section E.

Although tailored for solving POMDPs, the BVVI algorithm also adapts to MDP setups with lower regret due to the reduction of the uncertainty in the environment. In both scenarios, BVVI yields sublinear regret, growing approximately at a rate of $O\left(\sqrt{K}\right)$, consolidating our theoretical guarantee proposed in Theorem 6.1.

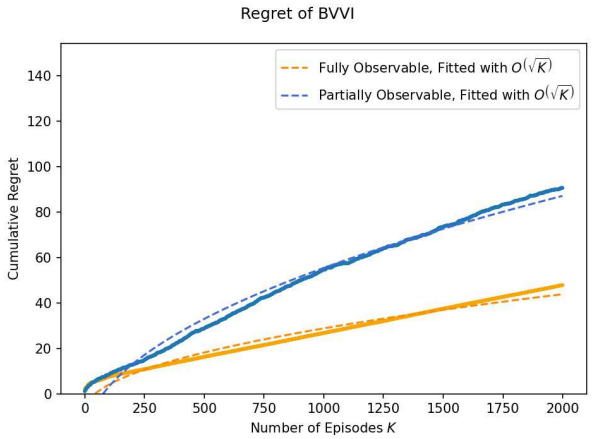

Figure 1: Regret of BVVI (Algorithm 1) in MDP and POMDP with $\gamma = 1$. Solid lines indicate cumulative regrets. Dashed curves are the regrets fitted with Theorem 6.1.

We also test BVVI in a POMDP under different risk sensitivity parameters. Our experiment shows that BVVI yields sublinear regret across multiple risk levels,demonstrating its versatility in handling various risk-sensitive RL scenarios.

[10] Please refer to Appendix D.4.4 for details.

Furthermore, our experiments demonstrate that the regret increases with the absolute value of γ, aligning with the theoretical reasoning presented in Section 6.

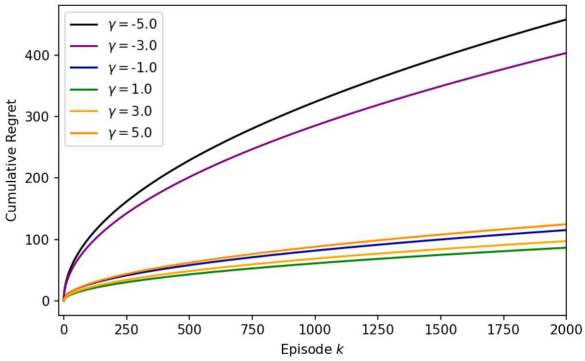

Figure 2: Cumulative regret of BVVI (Algorithm 1) in a POMDP with various risk-sensitivity. Risk-level γ ranges in $\{-5.0, -3.0, -1.0, 1.0, 3.0, 5.0\}$.

10. Conclusion and Future Work

In this study, we introduce a novel formulation of risk-sensitive RL in a partially observable environment with hindsight observations. We provide the first provably sample-efficient algorithm tailored for the new setting, whose regret improves existing upper bounds and nearly reaches the lower bounds in the degenerated cases. Our analysis also explains how the sample complexity is affected by the risk-awareness and history-dependency inherent in our problem. We validate the theoretical findings through numerical experiments, which demonstrates the algorithm's capability in solving POMDP problems across various levels of risk sensitivity.

There are several potential future directions. One is to derive theoretical results for POMDPs with function-approximation, for the case when the state space is extremely large or even infinite. Another avenue is to extend our findings to various risk measures such as conditional value-at-risk, coherent risk and optimized certainty equivalents.

Acknowledgement

This work was supported by the Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grant 2020AAA0108400 and 2020AAA0108403.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

Bäauerle, N. and Rieder, U. Partially observable risk-sensitive markov decision processes. *Mathematics of Operations Research*, 42(4):1180–1196, 2017.

Bai, Y. and Jin, C. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pp. 551–560. PMLR, 2020.

Baisero, A., Daley, B., and Amato, C. Asymmetric dqn for partially observable reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 107–117. PMLR, 2022.

Baras, J. S. and James, M. R. Robust and risk-sensitive output feedback control for finite state machines and hidden markov models. *J. Math. Systems, Estimation & Control*, 7:371–374, 1997. URL http://spigot.anu.edu.au/people/mat/home.html.

Bäuerle, N. and Rieder, U. *Markov decision processes with applications to finance*. Springer Science & Business Media, 2011.

Bäuerle, N. and Rieder, U. More risk-sensitive markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

Baxter, M. and Rennie, A. *Financial calculus: an introduction to derivative pricing*. Cambridge university press, 1996.

Bercu, B., Delyon, B., Rio, E., et al. *Concentration inequalities for sums and martingales*. Springer, 2015.

Boda, K. and Filar, J. A. Time consistent dynamic risk measures. *Mathematical Methods of Operations Research*, 63:169–186, 2006.

Boucheron, S., Lugosi, G., and Bousquet, O. *Concentration inequalities*. Springer, 2003.

Cai, Q., Yang, Z., and Wang, Z. Reinforcement learning from partial observation: Linear function approximation with provable sample efficiency, 2022.

Cao, J., Chen, J., Hull, J., and Poulos, Z. Deep hedging of derivatives using reinforcement learning. *The Journal of Financial Data Science*, 2020.

Cassandra, A. R. *Exact and approximate algorithms for partially observable Markov decision processes*. Brown University, 1998.

Cavazos-Cadena, R. and Hernández-Hernández, D. Successive approximations in partially observable controlled markov chains with risk-sensitive average criterion. *Stochastics: An International Journal of Probability and Stochastics Processes*, 77(6):537–568, 2005.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. On the generalization ability of on-line learning algorithms. *Information Theory, IEEE Transactions on*, 50:2050 – 2057, 10 2004. doi: 10.1109/TIT.2004.833339.

Chen, D., Zhou, B., Koltun, V., and Krähenbühl, P. Learning by cheating. In *Conference on Robot Learning*, pp. 66–75. PMLR, 2020.

Choi, J., Dance, C., Kim, J.-E., Hwang, S., and Park, K.-s. Risk-conditioned distributional soft actor-critic for risk-sensitive navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8337–8344. IEEE, 2021.

Di Masi, G. and Sthttner, L. Risk sensitive control of discrete time partially observed markov processes with infinite horizon. *Stochastics: An International Journal of Probability and Stochastic Processes*, 67(3-4):309–322, 1999.

Du, Y., Wang, S., and Huang, L. Provably efficient risk-sensitive reinforcement learning: Iterated cvar and worst path. In *The Eleventh International Conference on Learning Representations*, 2022.

Elliott, R. J., Moore, J. B., and Dey, S. Risk-sensitive maximum likelihood sequence estimation. *IFAC Proceedings Volumes*, 29(1):4616–4621, 1996. ISSN 1474-6670. doi: https://doi.org/10.1016/S1474-6670(17)58410-4. URL https://www.sciencedirect.com/science/article/pii/S1474667017584104. 13th World Congress of IFAC, 1996, San Francisco USA, 30 June - 5 July.

Fei, Y., Yang, Z., Chen, Y., Wang, Z., and Xie, Q. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395, 2020.

Fei, Y., Yang, Z., Chen, Y., and Wang, Z. Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 34:20436–20446, 2021a.

Fei, Y., Yang, Z., and Wang, Z. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on*

Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pp. 3198–3207. PMLR, 18–24 Jul 2021b. URL https://proceedings.mlr.press/v139/Fei21improve.html.

Fernandez-Gaucherand, E. and Marcus, S. Risk-sensitive optimal control of hidden markov models: structural results. *IEEE Transactions on Automatic Control*, 42(10): 1418–1422, 1997. doi: 10.1109/9.633830.

Golowich, N., Moitra, A., and Rohatgi, D. Planning in observable pomdps in quasipolynomial time. *arXiv preprint arXiv:2201.04735*, 1 2022a. URL http://arxiv.org/abs/2201.04735.

Golowich, N., Moitra, A., and Rohatgi, D. Learning in observable pomdps, without computationally intractable oracles. *Advances in Neural Information Processing Systems*, 6 2022b. URL http://arxiv.org/abs/2206.03446.

Guo, J., Chen, M., Wang, H., Xiong, C., Wang, M., and Bai, Y. Sample-efficient learning of pomdps with multiple observations in hindsight. *arXiv preprint arXiv:2307.02884*, 2023.

Hau, J. L., Delage, E., Ghavamzadeh, M., and Petrik, M. On dynamic programming decompositions of static risk measures in markov decision processes, 2023.

Huang, X., Hong, S., Hofmann, A., and Williams, B. C. Online risk-bounded motion planning for autonomous vehicles in dynamic environments. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29, pp. 214–222, 2019.

James, M. R., Baras, J. S., and Elliott, R. J. Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems. *IEEE Transactions on Automatic Control*, 39:780–792, 1994. ISSN 15582523. doi: 10.1109/9.286253.

Jazwinski, A. H. *Stochastic processes and filtering theory*. Courier Corporation, 2007.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient?, 2018.

Jin, C., Kakade, S., Krishnamurthy, A., and Liu, Q. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33:18530–18539, 2020.

Kabbani, T. and Duman, E. Deep reinforcement learning approach for trading automation in the stock market. *IEEE Access*, 10:93564–93574, 2022.

Kreyszig, E. *Introductory functional analysis with applications*, volume 17. John Wiley & Sons, 1991.

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

Lee, J. N., Agarwal, A., Dann, C., and Zhang, T. Learning in pomdps is sample-efficient with hindsight observability, 2023.

Liang, H. and Luo, Z.-q. Regret bounds for risk-sensitive reinforcement learning with lipschitz dynamic risk measures. *arXiv preprint arXiv:2306.02399*, 2023.

Liu, Q., Chung, A., Szepesvári, C., Ca, S., Jin, C., Loh, P.-L., and Raginsky, M. When is partially observable reinforcement learning not scary? *Proceedings of Machine Learning Research*, 178:1–46, 2022a.

Liu, Q., Netrapalli, P., Szepesvári, C., and Jin, C. Optimistic mle – a generic model-based algorithm for partially observable sequential decision making, 2022b.

Monahan, G. E. State of the art—a survey of partially observable markov decision processes: theory, models, and algorithms. *Management science*, 28(1):1–16, 1982.

Oksendal, B. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Papadimitriou, C. H. and Tsitsiklis, J. N. The complexity of markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/3689975.

Pineau, J., Gordon, G., and Thrun, S. Anytime point-based approximations for large pomdps. *Journal of Artificial Intelligence Research*, 27:335–380, November 2006. ISSN 1076-9757. doi: 10.1613/jair.2078. URL http://dx.doi.org/10.1613/jair.2078.

Pinto, L., Andrychowicz, M., Welinder, P., Zaremba, W., and Abbeel, P. Asymmetric actor critic for image-based robot learning. *arXiv preprint arXiv:1710.06542*, 2017.

Qiu, W., Wang, X., Yu, R., Wang, R., He, X., An, B., Obraztsova, S., and Rabinovich, Z. Rmix: Learning risk-sensitive policies for cooperative reinforcement learning agents. *Advances in Neural Information Processing Systems*, 34:23049–23062, 2021.

Rankin, R. Real and complex analysis. by w. rudin. pp. 412. 84s. 1966.(mcgraw-hill, new york.). *The Mathematical Gazette*, 52(382):412–412, 1968.

Richman, R. Ai in actuarial science–a review of recent advances–part 1. *Annals of Actuarial Science*, 15(2): 207–229, 2021.

Righi, M. B. A theory for combinations of risk measures. *arXiv preprint arXiv:1807.01977*, 2018.

Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Sereda, E., Bronshtein, E., Rachev, T., Fabozzi, F., Sun, E., and Stoyanov, S. Distortion risk measures in portfolio optimization. *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, pp. 649–673, 01 2010. doi: 10.1007/978-0-387-77439-8_25.

Shen, S., Ma, C., Li, C., Liu, W., Fu, Y., Mei, S., Liu, X., and Wang, C. Riskq: risk-sensitive multi-agent reinforcement learning value factorization. *arXiv preprint arXiv:2311.01753*, 2023.

Shi, M., Liang, Y., and Shroff, N. Theoretical hardness and tractability of pomdps in rl with partial hindsight state information. *arXiv preprint arXiv:2306.08762*, 2023.

Sinclair, S. R., Frujeri, F. V., Cheng, C.-A., Marshall, L., Barbalho, H. D. O., Li, J., Neville, J., Menache, I., and Swaminathan, A. Hindsight learning for mdps with exogenous inputs. In *International Conference on Machine Learning*, pp. 31877–31914. PMLR, 2023.

Skoglund, C. Risk-aware autonomous driving using pomdps and responsibility-sensitive safety, 2021.

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

Tirinzoni, A., Al-Marjani, A., and Kaufmann, E. Optimistic pac reinforcement learning: the instance-dependent view, 2022.

Vassilev, V., Donchev, D., and Tonchev, D. Risk assessment in transactions under threat as partially observable markov decision process. In *Optimization in Artificial Intelligence and Data Sciences: ODS, First Hybrid Conference, Rome, Italy, September 14-17, 2021*, pp. 199–212. Springer, 2022.

Von Neumann, J. and Morgenstern, O. Theory of games and economic behavior princeton. *Princeton University Press*, 1947:1953, 1944.

WATER, H. V. D. and Willems, J. C. The certainty equivalence property in stochastic control theory. *IEEE TRANSACTIONS ON AUTOMATIC CONTROL*, 1981.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Whittle, P. A risk-sensitive maximum principle: the case of imperfect state observation. *IEEE Transactions on Automatic Control*, 36(7):793–801, 1991. doi: 10.1109/9.85059.

Zhan, W., Uehara, M., Sun, W., and Lee, J. D. Pac reinforcement learning for predictive state representations, 2022.

A. Notations and Concepts

In this section we provide several additional concepts and notations not mentioned in Section 1.

Additional Notations Given a vector $\mathbf{x} \in \mathbb{R}^d$, we denote its i^{th} entry as $\mathbf{x}(i)$ or $[\mathbf{x}]_i$. For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we use $\mathbf{A}_{ij}$ or $[\mathbf{A}]_{i,j}$ to indicate the $(i,j)^{\text{th}}$ entry. The comparison and expectation of random vectors are defined component-wise. We employ $\mathbb{1}\{\cdot\}$ to represent the indicator operator, and the signature function is denoted as $\text{sgn}(\cdot)$. Additionally, we use $\text{Unif}(\mathcal{X})$ to express the uniform distribution over the finite space $\mathcal{X}$.

Remark on the POMDP

Remark A.1. Following the convention of POMDP literature (Monahan, 1982; James et al., 1994; Cavazos-Cadena & Hernández-Hernández, 2005; Golowich et al., 2022a), no observation is made at the first step. The definition of $\mathbf{O}_h$ begins from $h = 2$. The first action $\mathbf{A}_1$ is chosen based on the agent's prior knowledge. For the sake of consistent notations, we still adopt the notation of $\boldsymbol{F_1}$ and we use $\mathbb{P}(\cdot|\boldsymbol{F}_1)$ and $\mathbb{P}(\cdot)$ interchangeably. We permit the environment to generate the observation $\mathbf{O}H+1$ when in state $\mathbf{S}H+1$, but the agent abstains from taking any actions at $H+1$.

We will use the following fact extensively which expresses the recursive relation of the "history" defined in Section 3.

Fact A.2. $\forall h \in [H-1]:\quad \boldsymbol{F_{h+1}} = (\boldsymbol{F_h}, \mathbf{A}_h, \mathbf{O}_{h+1}), \qquad \text{where } \mathbf{A}_h = \pi_h(\boldsymbol{F}_h)$

In this study, we use f_{h+1} and (f_h, a_h, o_{h+1}) interchangeably.

Another concept related to "history" is the trajectory $\boldsymbol{\tau}_h$ of the Markov process.

Definition A.3. (Trajectory)

$$\begin{aligned} &\text{Full trajectory} \quad \bar{\boldsymbol{\tau}}_{\boldsymbol{h}} := (\mathbf{S}_1, \mathbf{A}_1, \dots \mathbf{S}_h, \mathbf{O}_h, \mathbf{A}_h), \forall h \in [H] \quad , \bar{\boldsymbol{\tau}}_{\boldsymbol{H+1}} := (\mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_H, \mathbf{O}_H, \mathbf{A}_H, \mathbf{S}_{H+1}) \\ &\text{Observable trajectory} \quad \boldsymbol{\tau}_{\boldsymbol{h}} := (\mathbf{A}_1, \dots, \mathbf{O}_h, \mathbf{A}_h), \forall h \in [H] \end{aligned} \tag{24}$$

Optimization Objective using General Utility Risk Measure In this work we refer the utility risk as any strictly increasing function that is continuously differentiable. We can extend many results in this work to general utility risk measures. We will present our proofs using the utility function U and instantiate it to the entropic risk ($U(\cdot) = \gamma e^{\gamma(\cdot)}$) when necessary.

The optimization objective using arbitrary utility risk measure U is defined as

$$\underset{\pi}{\text{maximize}} \;\; U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}} U\left[\sum_{t=1}^{H} r_t(\boldsymbol{S_t}, \boldsymbol{A_t})\right] \tag{25}$$

B. The Structure of Risk-sensitive POMDP

In what follows, we present the theoretical framework of partially observable reinforcement learning using arbitrary utility risk measures. Our framework builds upon the studies of (James et al., 1994; Cavazos-Cadena & Hernández-Hernández, 2005; Bäauerle & Rieder, 2017). Furthermore, we introduce novel concepts and provide several new proofs in a more comprehensive setting, enhancing the existing literature.

Given that the studies of risk-sensitive POMDP are relatively historical, we will provide detailed discussions about the intuition and implications behind various concepts and results. We aim to elucidate these findings, as they will serve as a foundation for the algorithm design and regret analysis in the subsequent sections.

B.1. Change of Measure

In reinforcement learning, the lack of knowledge about the emission process $\mathbf{O}_h$ presents a significant challenge for statistical inference, which motivates us to devise a surrogate POMDP $\mathcal{P}'$ named "reference model", which possesses a simplified emission process.

Definition B.1. (Reference model of a POMDP)
Given a POMDP model $\mathcal{P} = (\mathscr{S}, \mathscr{O}, \mathscr{A}; \mu_1, \mathbb{T}, \mathbb{O}; K, H, r)$ and a reference measure $\mathbb{O}'(\cdot) \in \Delta(\mathscr{O})$, the reference model of P specified by $\mathbb{O}'$ is another partially observable Markov decision process $\mathcal{P}' = (\mathscr{S}, \mathscr{O}, \mathscr{A}; \mu_1, \mathbb{T}, \mathbb{O}'; K, H, r)$, in which for all $h \in [H]$ and $s_h \in \mathscr{S}$, we have $\mathbb{O}'_h(\cdot|s_h) = \mathbb{O}'(\cdot)$.

In the reference model, the initial distribution and transition matrices mirror those of the real-world POMDP $\mathcal{P}$. However, the observations and hidden states are statistically independent and the emission process is stationary with a predefined observation probability. Consequently, the observations are separate from the underlying transition process and are independent of the history.

The probability of generating a full trajectory in the two models can be expressed as

$$\begin{aligned}\mathbb{P}^{\pi}_{\mathcal{P}}(\bar{\tau}_{\mathbf{h}}) =&\mu_1(\mathbf{S}_1)\mathbb{O}_1(\mathbf{O}_1|\mathbf{S}_1)\pi_1(\mathbf{A}_1|\mathbf{O}_1)\cdot\mathbb{T}_{h,\mathbf{A}_1}(\mathbf{S}_2|\mathbf{S}_1)\mathbb{O}_2(\mathbf{O}_2|S_2)\pi_2(\mathbf{A}_2|\boldsymbol{F}_2)\\ &\dots\mathbb{T}_{h-1,\mathbf{A}_{h-1}}(\mathbf{S}_{h+1}|\mathbf{S}_h)\mathbb{O}_h(\mathbf{O}_h|\mathbf{S}_h)\pi_h(\mathbf{A}_h|\boldsymbol{F}_h)\\ =&\left[\mu_1(\mathbf{S}_1)\prod_{t=1}^{h-1}\mathbb{T}_{t,\mathbf{A}_t}(\mathbf{S}_{t+1}|\mathbf{S}_t)\right]\cdot\left[\prod_{t=2}^{h}\mathbb{O}_t(\mathbf{O}_t|\mathbf{S}_t)\right]\cdot\left[\prod_{t=1}^{h}\pi_t(\mathbf{A}_t|\boldsymbol{F}_t)\right]\\ \mathbb{P}^{\pi}_{\mathcal{P}'}(\bar{\tau}_{\mathbf{h}}) =&\left[\mu_1(\mathbf{S}_1)\prod_{t=1}^{h-1}\mathbb{T}_{t,\mathbf{A}_t}(\mathbf{S}_{t+1}|\mathbf{S}_t)\right]\cdot\left[\prod_{t=2}^{h}\mathbb{O}'_t(\mathbf{O}_t|\mathbf{S}_t)\right]\cdot\left[\prod_{t=1}^{h}\pi_t(\mathbf{A}_t|\boldsymbol{F}_t)\right]\end{aligned}\tag{26}$$

Eq.(26) suggests that conditioned on the generated sigma-algebra $\mathscr{G}_h = \sigma\left(\{\mathbf{S}_t,\mathbf{O}_t,\mathbf{A}_t\}_{t=1}^{h}\right)$, the Radon-Nykodym derivative between the two trajectory probabilities takes the form of

$$\left.\frac{d\mathbb{P}^{\pi}_{\mathcal{P}}}{d\mathbb{P}^{\pi}_{\mathcal{P}'}}\right|_{\mathscr{G}_h} = \prod_{t=2}^{h}\frac{\mathbb{O}_t(\mathbf{O}_t|\mathbf{S}_t)}{\mathbb{O}'_t(\mathbf{O}_t|\mathbf{S}_t)} := D_h(\mathbf{O}_{2:h},\mathbf{S}_{2:h}) := \mathbf{D}_h$$

By Theorem G.1, for any measurable function f of the full trajectory $\bar{\tau}_h$, [11]

$$\mathbb{E}^{\pi}_{\mathcal{P}}[f(\bar{\tau}_{\mathbf{h}})] = \mathbb{E}^{\pi}_{\mathcal{P}'}[\mathbf{D}_h\cdot f(\bar{\tau}_{\mathbf{h}})] = \int_{\mathcal{T}_h}(D_h f)_{(\bar{\tau}_h)}\cdot d\mathbb{P}^{\pi}_{\mathcal{P}'}(\bar{\tau}_h)\tag{27}$$

The two expectations are taken with respect to the randomness in the transitions, emissions and the same policy. Then we can rewrite our optimization objective in Eq.(25) by the change of measure

$$J(\pi;\mathcal{P}) := \frac{1}{\gamma}\ln\mathbb{E}^{\pi}_{\mathcal{P}}\left[e^{\gamma\sum_{h=1}^{H}r_h(\mathbf{S_h},\mathbf{A_h})}\right] = \frac{1}{\gamma}\ln\mathbb{E}^{\pi}_{\mathcal{P}'}\left[\mathbf{D}_H\cdot e^{\gamma\sum_{h=1}^{H}r_h(\mathbf{S_h},\mathbf{A_h})}\right]\tag{28}$$

Remark B.2. In general, the conditional expectation $\mathbb{E}^{\pi}_{\mathcal{P}'}[\mathbf{D}_h\cdot\ |f_h]$ in Definition B.3 cannot be replaced by $\mathbb{E}^{\pi}_{\mathcal{P}}[\cdot\ |f_h]$, as our RN derivative $\mathbf{D}_h$ is calculated from the joint but not conditional probability.

B.2. Risk-sensitive Belief

One of the key concepts in the study of risk-neutral POMDP (Monahan, 1982)is the "belief state', which is the posterior distribution of the hidden states given the observable history.

$$\vec{b}_h(\cdot;f_h) = \mathbb{P}^{\pi}_{\mathcal{P}}\left\{\mathbf{S}_h=\cdot|\mathbf{F}_h=f_h\right\}$$

Since we can always view a probability from the perspective of expectation, we observe that [12]

$$\begin{aligned}\vec{b}_h(\cdot;f_h) =&\mathbb{P}^{\pi}_{\mathcal{P}}\left\{\mathbf{S}_h=\cdot|\mathbf{F}_h=f_h\right\}\\ \equiv&\mathbb{E}^{\pi}_{\mathcal{P}}\left[\mathbf{1}\{\mathbf{S}_h=\ \cdot\ \}\mid\mathbf{F}_h=f_h\right]\\ =&\mathbb{E}^{\pi}_{\mathcal{P}}\left[\mathbf{1}\{\mathbf{S}_h=\ \cdot\ \}e^{\gamma\sum_{t=1}^{h-1}r_t(\mathbf{S}_t,\mathbf{A}_t)}\mid\mathbf{F}_h=f_h\right]\Big|_{\gamma=0}\end{aligned}\tag{29}$$

The risk-sensitive counterpart of the belief is inspired by Eq.(29).

[11] We should also guarantee that the reference measure $\mathbb{O}'$ is strictly positive a.s. and $\mathbb{P}^{\pi}_{\mathcal{P}'} << \mathbb{P}^{\pi}_{\mathcal{P}}$. See Section G.1 for details.

[12] In the continuous case we replace the indicator $\mathbb{1}\{\cdot\}$ with the Dirac-delta function and the following definitions should be modified accordingly.

Definition B.3. (Risk-sensitive belief, Definition 2.9 in (Cavazos-Cadena & Hernández-Hernández, 2005))

For all $h \in [H+1]$, $f_h \in \mathscr{F}_h$, $s_h \in \mathscr{S}$:

$$\begin{aligned}[\vec{\sigma}_1]_{s_1} :=& \mu_1(s_1) \\ [\vec{\sigma}_{h,f_h}]_{s_h} :=& \mathbb{E}^{\pi}_{\mathcal{P}'}\left[\mathbf{D}_h \cdot \mathbb{1}\{\mathbf{S}_h = s_h\} \exp\gamma \sum_{t=1}^{h-1} r_t(\mathbf{S}_t, \mathbf{A}_t) \middle| \mathbf{F}_h = f_h\right]\end{aligned} \tag{30}$$

Remark B.4. The risk belief in this study is not normalized, since we make it a carrier of the one-step risk-sensitive reward. However, some literature (Cavazos-Cadena & Hernández-Hernández, 2005; Bäauerle & Rieder, 2017) still defines a normalized belief.

Remark B.5. In reinforcement learning, the risk beliefs corresponding to the empirical models $\widehat{\mathcal{P}}^k = (\widehat{\mu}_1^k, \widehat{\mathbb{T}}^k, \widehat{\mathbb{O}}^k)$ will be defined in a similar manner and referred to as the empirical belief $\widehat{\sigma}_h^k$.

Relationship with the Optimization Objective We can use the risk belief to express the optimization objective defined in Eq.(29).

$$\begin{aligned}J(\pi;\mathcal{P}) :=& U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}}\left[U\sum_{h=1}^{H} r_h(\mathbf{S_h}, \mathbf{A_h})\right] \\ \equiv& U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}'}\left[\mathbf{D}_{H+1} \cdot U\sum_{h=1}^{H} r_h(\mathbf{S_h}, \mathbf{A_h})\right] \quad //\text{Change of measure} \\ =& U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}'}\left[\mathbb{E}^{\pi}_{\mathcal{P}'}\left[\mathbf{D}_{H+1} \cdot U\sum_{h=1}^{H} r_h(\mathbf{S_h}, \mathbf{A_h}) \middle| \mathbf{F}_{H+1}\right]\right] \\ \equiv& U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}'}\left[\sum_{s_{H+1}\in\mathscr{S}} \mathbb{E}^{\pi}_{\mathcal{P}'}\left[\mathbf{1}\{\mathbf{S}_{H+1} = s_{H+1}\}\mathbf{D}_{H+1} \cdot U\sum_{h=1}^{H} r_h(\mathbf{S_h}, \mathbf{A_h}) \middle| \mathbf{F}_{H+1}\right] \cdot 1\right] \quad //\text{Lemma G.3} \\ \equiv& U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}'}\left[\langle \vec{\sigma}_{H+1,\mathbf{F}_{H+1}}, \vec{1}_S\rangle\right]\end{aligned} \tag{31}$$

If we can discover the evolution law of the new belief then we will be able to break the structure of $J(\pi;\mathcal{P})$ down by dynamic programming equations, as is presented in Section B.3.

Closed-form expression To gain more concrete understanding of the the specific structure of σ we may first utilize the Markov property of the hidden states to expand the condition measure of $\mathbf{S}_h$ given f_h:

Observation B.6. (Expansion of conditional probability) $\forall \pi \in \Pi, \mathcal{P}' = (\mu_1, \{\mathbb{T}_h\}, \{\mathbb{O}'_h\}), h \in [H+1], s_{1:h} \in \mathscr{S}^h$, $f_h = (o_1, a_1, \ldots, o_{h-1}, a_{h-1}, o_h) \in \mathscr{F}_{h+1}$,

$$\mathbb{P}^{\pi}_{\mathcal{P}'}(s_{1:h}, a_{1:h}|f_h) = \mu_1(s_1) \cdot \prod_{t=1}^{h-1} \mathbb{T}_{t,a_t}(s_{t+1}|s_t) \prod_{t=1}^{h} \pi_t(a_t|f_t) \tag{32}$$

The belief can then be computed by

$$\begin{aligned}&[\vec{\sigma}_{h+1,f_{h+1}}]_{s_{h+1}} \\ =& \sum_{\tilde{s}_{1:h+1}} \mathbb{P}^{\pi}_{\mathcal{P}'}(\tilde{s}_{1:h+1}, a_{1:h+1}|f_{h+1}) \cdot \left[\mathbb{1}\{\tilde{s}_{h+1} = s_{h+1}\} \cdot D_{h+1}(\tilde{s}_{2:h+1}, o_{2:h+1}) \cdot \exp\gamma \sum_{t=1}^{h} r_t(\tilde{s}_t, a_t)\right]\end{aligned} \tag{33}$$

A Special Case In tabular case, when we select the emission matrix as the uniform distribution, $D_h(o_{2:h}; s_{2:h})$ will become $\frac{1}{\mathbb{O}'_1(o_1)}^h \prod_{t=2}^{h} \mathbb{O}_t(o_t|s_t)$. Moreover, if the policies are deterministic, $\mathbb{P}^{\pi}_{\mathcal{P}'}(s_{1:h}, a_{1:h}|f_h) = \mu_1(s_1)\prod_{t=1}^{h-1}\mathbb{T}_{t,a_t}(s_{t+1}|s_t)$. Plugging the two terms together in Eq.(33) we conclude that

Corollary B.7. *(Belief vector using uniform emission matrix, Eq. (2.9) in (Cavazos-Cadena & Hernández-Hernández, 2005)) Suppose that $\mathbb{O}'_t(\cdot|\mathbf{S}_t) = Unif\mathscr{O}$ and the policies are deterministic, then*

$$\begin{aligned}[\vec{\sigma_1}]_{s_1} =&\mu_1(s_1)\\ [\vec{\sigma}_{h,f_h}]_{s_h} =&|\mathscr{O}|^h\,\mathbb{E}^{\pi}_{\mathcal{P}}\left[\prod_{t=2}^{h}\mathbb{O}_t(o_t|\mathbf{S}_t)\mathbb{1}\{\mathbf{S}_h=s_h\}e^{\gamma\ \sum_{t=1}^{h-1}r_t(\mathbf{S}_t,\mathbf{A}_t)}\right],\quad \forall 2\le h\le H+1\end{aligned} \tag{34}$$

Evolution law The risk belief evolves in a Markovian manner, incorporating recent action a_h and observation o_{h+1} to the new belief. We will use $\Psi(\cdot,a_h,o_{h+1}):\mathbb{R}^S\to\mathbb{R}^S$ to denote the update operator of $\vec{\sigma}_{h,f_h}$, whose matrix representation will be denoted as $\mathsf{U}_{a_h,o_{h+1}}\in\mathbb{R}^{S\times S}$. The details of the update process is specified by the following theorem.

Theorem B.8. *(Evolution of risk-sensitive belief, adapted from theorem 2.2 of (James et al., 1994))*

$$\begin{aligned}&\forall h=H,H-1,\cdots,1, f_{h+1}=(f_h,a_h,o_{h+1})\in\mathscr{F}_{h+1}, s_{h+1}\in\mathscr{S}:\\ &\quad[\vec{\sigma}_{h+1,f_{h+1}}]_{s_{h+1}} =\Psi(\vec{\sigma}_{h,f_h},a_h,o_{h+1})=\left[\mathsf{U}_{a_h,o_{h+1}}\vec{\sigma}_{h,f_h}\right]\\ &\qquad\qquad=\sum_{s_h}\mathbb{T}_{h,a_h}(s_{h+1}|s_h)\mathbb{O}_{h+1}(o_{h+1}|s_{h+1})\cdot\left(\frac{e^{\gamma r_h(s_h,a_h)}}{\mathbb{O}'_{h+1}(o_{h+1}|s_{h+1})}\right)[\vec{\sigma}_{h,f_h}]_{s_h}\end{aligned} \tag{35}$$

Remark B.9. The proof for this theorem in the continuous case is provided by (James et al., 1994). However, their proof was written in the language of functional analysis and they have restricted the transition and observation probabilities to be i.i.d. Gaussian distributions. In the tabular setting, though (Cavazos-Cadena & Hernández-Hernández, 2005) have presented a similar result in Eq.(2.10), they have omitted the proof and restricted the reference measure $\mathbb{O}'$ as the uniform distribution. For the reader's convenience, in what follows we will prove Theorem B.8 in the tabular case using simple algebraic calculations, which also accommodates arbitrary structures of $\mathbb{O}'$, $\mathbb{T}$ and $\mathbb{O}$.

Proof.

$$\begin{aligned}RHS=&\sum_{s_h}\mathbb{T}_{h,a_h}(s_{h+1}|s_h)\frac{\mathbb{O}_{h+1}}{\mathbb{O}'_{h+1}}(o_{h+1}|s_{h+1})\exp\gamma r_h(s_h,a_h)\\ &\left[\sum_{\tilde{s}_{1:h}}\mathbb{1}\{\tilde{s}_h=s_h\}\prod_{t=2}^{h}\frac{\mathbb{O}_t}{\mathbb{O}'_t}(o_t|\tilde{s}_t)\exp\gamma\sum_{t=1}^{h-1}r_t(\tilde{s}_t,a_t)\mathbb{P}^{\pi}_{\mathcal{P}'}(\tilde{s}_{1:h}|f_h)\right]\quad //(\text{Definition B.3})\end{aligned} \tag{36}$$

$$\begin{aligned}=&\sum_{s_h}\mathbb{T}_{h,a_h}(s_{h+1}|s_h)\frac{\mathbb{O}_{h+1}}{\mathbb{O}'_{h+1}}(o_{h+1}|s_{h+1})\exp\gamma r_h(s_h,a_h)\\ &\sum_{\tilde{s}_{1:h}}\prod_{t=2}^{h}\frac{\mathbb{O}_t}{\mathbb{O}'_t}(o_t|\tilde{s}_t)1\{\tilde{s}_h=s_h\}\exp\gamma\sum_{t=1}^{h-1}r_t(\tilde{s}_t,a_t)\mu_1(\tilde{s}_1)\prod_{t=1}^{h-1}\mathbb{P}^{\pi}_{\mathcal{P}'}(\tilde{s}_{t+1}|\tilde{s}_t,a_t)\quad //(\text{Observation B.6})\end{aligned} \tag{37}$$

Since $\mathcal{P}'$ and $\mathcal{P}$ share the same transition matrix, rearranging terms by Fubini's theorem we have

$$\begin{aligned}RHS=&\sum_{\tilde{s}_{1:h-1}}\left(\sum_{s_h}\mathbb{P}^{\pi}_{\mathcal{P}'}(s_{h+1}|s_h,a_h)\exp\gamma r_h(s_h,a_h)\sum_{\tilde{s}_h}\mathbb{1}\{\tilde{s}_h=s_h\}\mathbb{P}^{\pi}_{\mathcal{P}'}(\tilde{s}_h|\tilde{s}_{h-1},a_h)\right)\\ \cdot&\left(\mu_1(\tilde{s}_1)\prod_{t=1}^{h-2}\mathbb{P}^{\pi}_{\mathcal{P}'}(\tilde{s}_{t+1}|\tilde{s}_t,a_t)\right)\cdot\left(\frac{\mathbb{O}_{h+1}}{\mathbb{O}'_{h+1}}(o_{h+1}|s_{h+1})\prod_{t=2}^{h}\frac{\mathbb{O}_t}{\mathbb{O}'_t}(o_t|\tilde{s}_t)\right)\cdot\left(\exp\gamma r_h(s_h,a_h)\exp\gamma\sum_{t=1}^{h-1}r_t(\tilde{s}_t,a_t)\right)\end{aligned} \tag{38}$$

$$\begin{aligned}=&\sum_{\tilde{s}_{1:h-1}}\left(\sum_{s_h}\mathbb{P}^{\pi}_{\mathcal{P}'}(s_{h+1}|s_h,a_h)\cdot\mathbb{P}^{\pi}_{\mathcal{P}'}(s_h|\tilde{s}_{h-1},a_h)\right)\cdot\left(\mu_1(\tilde{s}_1)\prod_{t=1}^{h-2}\mathbb{P}^{\pi}_{\mathcal{P}'}(\tilde{s}_{t+1}|\tilde{s}_t,a_t)\right)\\ &\cdot\left(\frac{\mathbb{O}_{h+1}}{\mathbb{O}'_{h+1}}(o_{h+1}|s_{h+1})\prod_{t=2}^{h}\frac{\mathbb{O}_t}{\mathbb{O}'_t}(o_t|\tilde{s}_t)\right)\cdot\left(\exp\gamma r_h(s_h,a_h)\exp\gamma\sum_{t=1}^{h-1}r_t(\tilde{s}_t,a_t)\right)\end{aligned} \tag{39}$$

Relabel s_h as $\tilde{s}_h$ and invoke the equality $f(s_{h+1}) \equiv \sum_{\tilde{s}_{h+1}} f(\tilde{s}_{h+1}) \mathbb{1}\{\tilde{s}_{h+1} = s_{h+1}\}$, we conclude

$$\begin{aligned}
RHS =& \sum_{\tilde{s}_{h+1}} \mathbb{1}\{\tilde{s}_{h+1} = s_{h+1}\} \left[\sum_{\tilde{s}_{1:h-1}} \sum_{\tilde{s}_h} (\mathbb{P}^{\pi}_{\mathcal{P}'}(\tilde{s}_{h+1}|\tilde{s}_h, a_h) \cdot \mathbb{P}^{\pi}_{\mathcal{P}'}(\tilde{s}_h|\tilde{s}_{h-1}, a_h)) \cdot \left(\mu_1(\tilde{s}_1) \prod_{t=1}^{h-2} \mathbb{P}^{\pi}_{\mathcal{P}'}(\tilde{s}_{t+1}|\tilde{s}_t, a_t) \right) \right] \\
& \cdot \left(\frac{\mathbb{O}_{h+1}}{\mathbb{O}'_{h+1}}(o_{h+1}|\tilde{s}_{h+1}) \prod_{t=2}^{h} \frac{\mathbb{O}_t}{\mathbb{O}'_t}(o_t|\tilde{s}_t) \right) \cdot \left(\exp \gamma r_h(\tilde{s}_h, a_h) \exp \gamma \sum_{t=1}^{h-1} r_t(\tilde{s}_t, a_t) \right) \\
=& \sum_{\tilde{s}_{1:h+1}} \mathbb{1}\{\tilde{s}_{h+1} = s_{h+1}\} \cdot \left[\mu_1(\tilde{s}_1) \prod_{t=1}^{h} \mathbb{P}^{\pi}_{\mathcal{P}'}(\tilde{s}_{t+1}|\tilde{s}_t, a_t) \right] \cdot \left(\prod_{t=2}^{h+1} \frac{\mathbb{O}_t}{\mathbb{O}'_t}(o_t|\tilde{s}_t) \right) \cdot \left(\exp \gamma \sum_{t=1}^{h} r_t(\tilde{s}_t, a_t) \right) \\
=& \sum_{\tilde{s}_{1:h+1}} \mathbb{P}^{\pi}_{\mathcal{P}'}(\tilde{s}_{1:h+1}|f_{h+1}) \cdot \left[\mathbb{1}\{\tilde{s}_{h+1} = s_{h+1}\} \cdot \mathbf{D}_{h+1}(o_{2:t}; \tilde{s}_{1:t}) \cdot \exp \gamma \sum_{t=1}^{h} r_t(\tilde{s}_t, a_t) \right] \\
=& \mathbb{E}^{\pi}_{\mathcal{P}'} \left[\mathbf{1}\{\mathbf{S}_{h+1} = s_{h+1}\} \cdot \mathbf{D}_{h+1}(\mathbf{O}_{2:t}; \mathbf{S}_{1:t}) \cdot \exp \gamma \sum_{t=1}^{h} r_t(\mathbf{S}_t, \mathbf{A}_t) \middle| \mathbf{F}_{h+1} = f_{h+1} \right] \\
=& LHS \quad //\text{Definition B.3}
\end{aligned}$$

□

Remark B.10. When we specify $\mathbb{O}'_h(\cdot|s_h)$ as the uniform distribution $\mathrm{Unif}(\mathscr{O})$, we can write the update formula as

$$\vec{\sigma}_{h+1,f_{h+1}} = \mathsf{U}_{a_h,o_{h+1}} \vec{\sigma}_{h,f_h} = |\mathscr{O}| \operatorname{diag}\left(\mathbb{O}_{h+1}(o_{h+1}|\cdot)\right) \mathbb{T}_{h,a_h} \operatorname{diag}\left(\exp \gamma r_h(\cdot, a_h)\right) \vec{\sigma}_{h,f_h} \tag{40}$$

Remark B.11. (Initial belief) There are multiple ways to define our initial belief according to Definition B.25, since $\sum_{h=1}^{0}$ is ill-defined in nature. The optimization problem also poses no restriction on σ_1, since we present the optimization objective by σ_{H+1} instead. However, since we wish to represent σ_{H+1} by its predecessors, an appropriate definition of σ_1 should be compatible with our update rule, so that we can derive σ_1 from the σ_2 by Eq.(35). A simple calculation will show that such constraint impels $\sigma_1(s_1) = \mu_1(s_1)$.

Remark B.12. For simplicity, we have presented the theorem in the tabular case. With slight modifications, similar result holds in the continuous case. However, when the spaces are infinite, the evolution operator $\mathsf{U}^\star$ may not have a matrix representation as presented in Eq.(40).

B.3. Conjugate Beliefs

In the analysis of Eq.(31), we can express the objective by the terminal belief $\{\vec{\sigma}_{H+1}\}$

$$J(\pi; \mathcal{P}) = U^{-1} \mathbb{E}^{\pi}_{\mathcal{P}'} \left[\langle \vec{\sigma}_{H+1, \mathbf{F}_{H+1}}, \vec{1}_S \rangle \right]$$

It is reasonable to express σ_{H+1} by its predecessors, which posses simpler structures. However, as Theorem B.8 suggests, $\vec{\sigma}_t$ evolves forward in time, which hinders us from writing σ_{H+1} in terms of $\{\sigma_t\}_{t \le H}$.

$$[\vec{\sigma}_1]_{s_1} := \vec{\mu}_1(s_1), \quad \vec{\sigma}_{h+1,f_{h+1}} = \mathsf{U}_{a_h,o_{h+1}} \vec{\sigma}_{h,f_h}, \forall h \in [H] \tag{41}$$

To bridge the gap in the the direction of evolution, we would like to introduce another process $\vec{\nu}_t$ that evolves backward in time, so that it is straightforward to express her initial state $\vec{\nu}_{H+1}$ by the predecessors, according to the new update rule.

We find it convenient to first introduce several concepts that describe how a stochastic process evolves backward in the episodic setting.

Definition B.13. (Backward History)

$$\begin{aligned}
&\bar{\boldsymbol{F}}_{H+1} = \emptyset \qquad \forall h = H, H-1, \ldots, 1 \; \bar{\boldsymbol{F}}_h = (\mathbf{A}_h, \mathbf{O}_{h+1}, \ldots, \mathbf{A}_{H-1}, \mathbf{O}_H, \mathbf{A}_H), \quad \bar{\boldsymbol{F}}_0 = \tau_H \\
&\forall h \in [H]: \quad \bar{\boldsymbol{F}}_h = (\mathbf{A}_h, \mathbf{O}_{h+1}, \bar{\boldsymbol{F}}_{h+1}) \qquad \sigma(\bar{\boldsymbol{F}}_0) \supset \sigma(\bar{\boldsymbol{F}}_1) \ldots \supset \sigma(\bar{\boldsymbol{F}}_{H+1})
\end{aligned} \tag{42}$$

Definition B.13 implies that $\{\bar{\boldsymbol{F}}_t\}_{t \ge 0}$ and $\{\boldsymbol{F}_t\}_{t \ge 0}$ are complementary at all times.

Observation B.14. (Complementary relation) $\forall h = H+1, H, \cdots, 0, \quad (\boldsymbol{F}_h, \bar{\boldsymbol{F}}_h) = \boldsymbol{\tau}_H$

Now we are ready to define the backward process $\{\vec{\nu}_t\}_{t\geq 0}$., whose update operator will be the Hilbert-adjoint operator [13] of that of $\vec{\sigma}_t$.

Definition B.15. (Conjugate Beliefs, Definition 2.8 in (James et al., 1994))

$$\begin{aligned} \vec{\nu}_{H+1}(\cdot) :&\equiv \vec{1}_S \\ \vec{\nu}_{h,\bar{f}_h} :&= \mathsf{U}^{\top}_{a_h,o_{h+1}} \vec{\nu}_{h+1,\bar{f}_{h+1}}, \text{for all } \ h = H, H-1, \cdots, 1, \ \ \bar{f}_h = (a_h, o_{h+1}, \bar{f}_{h+1}) \in \mathscr{A} \times \mathscr{O} \times \bar{\mathscr{F}}_h \end{aligned} \tag{43}$$

Remark B.16. In tabular case when we select the emission measure of the reference model as uniform distribution, we have:

$$[\vec{\nu}_{h,\bar{f}_h}]_{s_h} = \frac{e^{\gamma r_h(s_h,a_h)}}{\mathbb{O}'(o_{h+1})} \sum_{s_{h+1}\in\mathscr{S}} \mathbb{T}_{h,a_h}(s_{h+1}|s_h)\mathbb{O}_{h+1}(o_{h+1}|s_{h+1})[\vec{\nu}_{h+1,\bar{f}_{h+1}}]_{s_{h+1}} \tag{44}$$

We have carefully designed the update rule of the conjugate belief: the complementary relation (B.14) immediately implies the inner product between $\vec{\sigma}_t$ and $\vec{\nu}_t$ does not change with time. This result helps to excavate the dynamic programming structure hidden within the optimization objective.

Observation B.17. (Conjugate evolution, Eq.(2.9) in (James et al., 1994)) For all $h \in [H+1]$,

$$\langle \vec{\sigma}_{H+1,f_{H+1}}, \vec{1} \rangle = \langle \vec{\sigma}_{h,f_h}, \vec{\nu}_{h,\bar{f}_h} \rangle = \cdots \langle \vec{\mu}_1(s_1), \vec{\nu}_{1,\bar{f}_1} \rangle \tag{45}$$

Proof. By the definition of adjoint operator,

$$\langle \vec{\sigma}_{t,f_t}, \vec{\nu}_{t,\bar{f}_t} \rangle = \langle \mathsf{U}_{a_{t-1},o_t} \vec{\sigma}_{t-1,f_{t-1}}, \vec{\nu}_{t,\bar{f}_t} \rangle = \langle \vec{\sigma}_{t-1,f_{t-1}}, \mathsf{U}^{\top}_{a_{t-1},o_t} \vec{\nu}_{t,\bar{f}_t} \rangle = \langle \vec{\sigma}_{t-1,f_{t-1}}, \vec{\nu}_{t-1,\bar{f}_{t-1}} \rangle$$

□

Bringing Eq.(45) back to (B.3) we immediately conclude that for all $h \in [H]$,

$$J(\pi;\mathcal{P}) = U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}'}\left[U\langle \vec{\sigma}_{H+1,\mathbf{F_{H+1}}}, \vec{1}\rangle)\right] = U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}'}\left[U\langle \vec{\sigma}_{h,\mathbf{F_h}}, \vec{\nu}_{h,\bar{\mathbf{F}}_\mathbf{h}}\rangle\right] = U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}'}\left[U\langle \vec{\sigma}_{1,\mathbf{F}_1}, \vec{\nu}_{1,\bar{\mathbf{F}}_\mathbf{1}}\rangle\right] \tag{46}$$

B.4. Value functions, Q-functions and Bellman equations

In this section we derive the value functions and Bellman equations for general tabular POMDP models using arbitrary utility risk measure. Our derivation and Bellman equations are different from previous works including (James et al., 1994; Cavazos-Cadena & Hernández-Hernández, 2005; Bäauerle & Rieder, 2017).

From the subscripts in Eq.(46) we can witness the trace of time evolution hidden within $J(\pi;\mathcal{P})$. To expose the dynamic programming structure more explicitly, we will follow the rationale behind the design of belief states, utilizing the iterated expectation formula to define a series of intermediate variables that dissect the information at each step. These variables will be called the partially observable risk-sensitive value functions. For all $t \in [H+1]$,

$$\begin{aligned} J(\pi;\mathcal{P}) :&= U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}}\left[U\sum_{t=1}^{H} r_t(\mathbf{S}_t, \mathbf{A}_t)\right] \\ &= U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}'}[\langle \vec{\sigma}_{H+1,\mathbf{F}_{H+1}}, \vec{\nu}_{H+1,\bar{\mathbf{F}}_{H+1}}\rangle] \quad \text{//Belief representation by Eq.(31)} \\ &= U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}'}[\langle \vec{\sigma}_{t,\mathbf{F}_t}, \vec{\nu}_{t,\bar{\mathbf{F}}_t}\rangle] \quad \text{//Conjugate evolution property proved in Eq.(45)} \\ &= U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}'}\left[UU^{-1}\mathbb{E}^{\pi}_{\mathcal{P}'}\left[\langle \vec{\sigma}_{t,\mathbf{F}_t}, \vec{\nu}_{t,\bar{\mathbf{F}}_t}\rangle \middle| \mathbf{F}_t\right]\right] \\ &:= U^{-1}\mathbb{E}^{\pi}_{\mathcal{P}'}\left[U\mathsf{V}^{\pi}_t(\boldsymbol{F}_t)\right] \end{aligned}$$

Now we formally define the family of value functions in our problem.

[13] For a rigorous definition please refer to Section G.2.

Definition B.18. (Partially observable risk-sensitive value functions, Definition 2.13 in (James et al., 1994))

$$\begin{aligned}
\mathsf{V}_{H+1}^{\pi}(f_{H+1}) :=& U^{-1}\mathbb{E}_{\mathcal{P}'}^{\pi}\left[\langle \vec{\sigma}_{H+1,\mathbf{F}_{H+1}}, \vec{1}_S\rangle \middle| \mathbf{F}_{H+1} = f_{H+1}\right] = U^{-1}\|\vec{\sigma}_{h+1,f_{h+1}}\|_1 \\
\mathsf{V}_{h}^{\pi}(f_{h}) :=& U^{-1}\mathbb{E}_{\mathcal{P}'}^{\pi}\left[\langle \vec{\sigma}_{h,\mathbf{F}_h}, \vec{\nu}_{h,\bar{\mathbf{F}}_h}\rangle \middle| \mathbf{F}_h = f_h\right] \quad \forall 2 \le h \le H, f_h \in \mathscr{F}_h \\
\mathsf{V}_{1}^{\pi}(f_{1}) :=& U^{-1}\mathbb{E}_{\mathcal{P}'}^{\pi}\left[\langle \vec{\mu}_1, \vec{\nu}_{1,\bar{\mathbf{F}}_1}\rangle\right]
\end{aligned} \tag{47}$$

Remark B.19. The objective $J(\pi;\mathcal{P})$ can be expressed by the value function:

$$J(\pi;\mathcal{P}) = U^{-1}\mathbb{E}_{\mathcal{P}'}\left[U\mathsf{V}_1^{\pi}\right] = U^{-1}\mathbb{E}_{\mathcal{P}'}\langle \vec{\sigma}_1, \vec{\nu}_1\rangle$$

In reinforcement learning, we are also curious about how the action to take might affect future rewards. We can separate the stochasticity within π and $\mathcal{P}$ by the total expectation formula:

$$\begin{aligned}
\mathsf{V}_{h}^{\pi}(f_h) :=& U^{-1}\mathbb{E}_{\mathcal{P}'}^{\pi}\left[\langle \vec{\sigma}_h(\boldsymbol{F_h}), \vec{\nu}_h(\bar{\boldsymbol{F}}_h)\rangle \middle| \boldsymbol{F_h} = f_h\right] \quad //\text{Definition B.18} \\
=& U^{-1}\mathbb{E}^{\pi}\left[\mathbb{E}_{\mathcal{P}'}\left[\mathbb{E}_{\mathcal{P}'}^{\pi}\left[\langle \vec{\sigma}_{h+1,\boldsymbol{F_{h+1}}}, \vec{\nu}_{h+1,\bar{\mathbf{F}}_{h+1}}\rangle \middle| \boldsymbol{F}_h, \mathbf{A}_h, \mathbf{O}_{h+1}\right] \middle| \boldsymbol{F}_h, \mathbf{A}_h\right] \middle| \boldsymbol{F_h} = f_h\right] \\
=& U^{-1}\mathbb{E}^{\pi}\left[\mathbb{E}_{\mathcal{P}'}\left[U^{-1}\mathsf{V}_{h+1}^{\pi}(\boldsymbol{F_{h+1}}) \middle| \boldsymbol{F}_h, \mathbf{A}_h\right] \middle| \boldsymbol{F_h} = f_h\right] \\
:=& U^{-1}\mathbb{E}^{\pi}\left[UU^{-1}\mathbb{E}_{\mathcal{P}'}\left[U^{-1}\mathsf{V}_{h+1}^{\pi}(\boldsymbol{F_{h+1}}) \middle| \boldsymbol{F}_h, \mathbf{A}_h\right] \middle| \boldsymbol{F_h} = f_h\right] \\
=& U^{-1}\mathbb{E}^{\pi}\left[UU^{-1}\mathbb{E}_{\mathcal{P}'}\left[U^{-1}\mathsf{V}_{h+1}^{\pi}(\boldsymbol{F_{h+1}}) \middle| f_h, \mathbf{A}_h\right] \middle| \boldsymbol{F_h} = f_h\right] \\
:=& U^{-1}\mathbb{E}_{\mathbf{A_h}\sim\pi_h(\cdot|f_h)}U\mathsf{Q}_h^{\pi}(f_h, \mathbf{A}_h)
\end{aligned}$$

Now we formally introduce the Q functions in our problem:

Definition B.20. (Partially observable risk-sensitive Q-functions) For all $(h, s_h, a_h) \in [H] \times \mathscr{S} \times \mathscr{A}$,

$$\mathsf{Q}_h^{\pi}(f_h, a_h) := U^{-1}\mathbb{E}_{\mathcal{P}'}\left[U\mathsf{V}_{h+1}^{\pi}(\mathbf{F}_{h+1} = (f_h, a_h, \mathbf{O}_{h+1}))\middle| f_h, a_h\right]$$

Remark B.21. We can also represent the Q-function by risk beliefs. Indeed,

$$\begin{aligned}
\mathsf{Q}_h^{\pi}(f_h, a_h) =& U^{-1}\mathbb{E}_{\mathbf{O}_{h+1}\sim\mathbb{P}_{\mathcal{P}'}(\cdot|f_h,a_h)}\left[U\mathsf{V}_{h+1}^{\pi}(f_h, a_h, \boldsymbol{O}_{h+1})\right] \\
=& U^{-1}\mathbb{E}_{\mathcal{P}'}\left[\langle \vec{\sigma}_{H+1,\mathbf{F}_{H+1}}, \vec{\nu}_{H+1,\bar{\mathbf{F}}_{H+1}}\rangle \middle| f_h, a_h\right] = U^{-1}\mathbb{E}_{\mathcal{P}'}\left[\langle \vec{\sigma}_{h,\mathbf{F}_h}, \vec{\nu}_{h,\bar{\mathbf{F}}_h}\rangle \middle| f_h, a_h\right]
\end{aligned} \tag{48}$$

where the last step is due to Eq.(45). We can also use (48) as an alternative definition of the Q function.

The relationship between the value function and the Q function is summarized as the Bellman equations:

Corollary B.22. *(Bellman equations for risk-sensitive POMDP)*

$$\begin{aligned}
\mathsf{V}_{H+1}^{\pi} =& U^{-1}\|\vec{\sigma}_{h+1,f_{h+1}}\|_1 \\
\forall h = H:1,\ \mathsf{Q}_h^{\pi}(f_h, a_h) =& U^{-1}\mathbb{E}_{\mathbf{O}_{h+1}\sim\mathbb{O}'(\cdot)}\left[U\mathsf{V}_{h+1}^{\pi}(f_h, a_h, \mathbf{O}_{h+1})\right] \\
\mathsf{V}_h^{\pi}(f_h) =& \mathbb{E}_{\mathbf{A_h}\sim\pi_h(\cdot|f_h)}\mathsf{Q}_h^{\pi}(f_h, \mathbf{A}_h) \\
J(\pi;\mathcal{P}) =& U^{-1}\mathbb{E}_{\mathcal{P}'}\left[U\mathsf{V}_1^{\pi}\right]
\end{aligned} \tag{49}$$

The expectation in the second equation should have been taken with respect to $\mathbb{P}_{\mathcal{P}'}^{\pi}(\cdot|f_h, a_h)$. However, since the observations are disjoint from the POMDP in the reference model, we obtain a simpler expression in Eq.(49).

B.5. Optimal Policy

Optimal substructure Since our utility risk measure is increasing, to optimize our objective $U^{-1}\mathbb{E}U(\sum_{h=1}^{H} r_h)$ is equivalent to maximize $\mathbb{E}U(\sum_{h=1}^{H} r_h)$. According to Theorem 3.3 of (Bäauerle & Rieder, 2017), the latter problem has a globally optimal policy, and so does the former. Moreover, the following theorem shows the existence of an optimal substructure in the planning problem of risk-sensitive POMDP. This result justifies the use of dynamic programming equations in our algorithm.

Theorem B.23. *(Bellman optimality equations, extended from theorem 2.5 of (James et al., 1994)) When the utility risk function is strictly increasing and the referential emission matrix $\mathbb{O}'_t(\cdot|s)$ is irrelevant with the history $\mathbf{F}_h$, the locally optimized policy will bring globally optimized value. Formally, the locally optimal values defined by*

$$\mathsf{V}_1^\star := \max_\pi \mathsf{V}_1^\pi, \qquad \mathsf{V}_h^\star(f_h) := \max_\pi \mathsf{V}_h^\pi(f_h), \quad \forall 2 \le h \le H+1$$

can be computed recursively:

$$\begin{cases} V_1^\star = \max\limits_{a_1 \in \mathscr{A}} U^{-1}\mathbb{E}_{\mathcal{P}'}\left[UV_2^\star(a_1, \mathbf{O}_2)\right] \\ V_h^\star(f_h) = \max\limits_{a_h \in \mathscr{A}} U^{-1}\mathbb{E}_{\mathcal{P}'}\left[UV_{h+1}^\star(f_h, a_h, \mathbf{O}_{h+1})\right], \quad \forall h = H:2 \\ V_{H+1}^\star(f_{H+1}) = U^{-1}\|\vec{\sigma}_{h+1,f_{h+1}}\|_1 \end{cases} \tag{50}$$

Proof.

$$\begin{aligned} V_h^\star(f_h) :=& \max_{a_{h:H}} U^{-1}\mathbb{E}_{\mathcal{P}'}\left[\langle \vec{\sigma}_{h,\mathbf{F}_h}, \vec{\nu}_{h,\bar{\mathbf{F}}_h}\rangle \mid \mathbf{F}_h = f_h\right] \text{//Definition of value functions in B.18} \\ =& U^{-1}\max_{a_{h:H}} \mathbb{E}_{\mathcal{P}'}\left[\langle \vec{\sigma}_{h,\mathbf{F}_h}, \vec{\nu}_{h,\bar{\mathbf{F}}_h}\rangle \mid \mathbf{F}_h = f_h\right] \quad \text{//}U \text{ monotonically increases, so does } U^{-1}. \\ =& U^{-1}\max_{a_h}\left\{\max_{a_{h+1:H}} \mathbb{E}_{\mathcal{P}'}\left[\mathbb{E}_{\mathcal{P}'}\left[\langle \vec{\sigma}_{h,\mathbf{F}_h}, \vec{\nu}_{h,\bar{\mathbf{F}}_h}\rangle \middle| \mathbf{F}_{h+1}\right] \middle| \mathbf{F}_h = f_h\right]\right\} \\ =& U^{-1}\max_{a_h}\left\{\max_{a_{h+1:H}} \mathbb{E}_{\mathcal{P}'}\left[\mathbb{E}_{\mathcal{P}'}\left[\langle \vec{\sigma}_{H+1,\mathbf{F}_{H+1}}, \vec{\nu}_{h+1,\bar{\mathbf{F}}_{h+1}}\rangle \middle| \mathbf{F}_{h+1}\right] \middle| \mathbf{F}_h = f_h\right]\right\} \text{//Observation B.17} \\ =& U^{-1}\max_{a_h}\left\{\max_{a_{h+1:H}} \mathbb{E}_{\mathcal{P}'}\left[\mathbb{E}_{\mathcal{P}'}\left[\langle \vec{\sigma}_{h+1,\mathbf{F}_h,a_h,\mathbf{O}_{h+1}}, \vec{\nu}_{h+1,a_{h+1},\mathbf{O}_{h+2},\cdots,\mathbf{O}_H,a_H}\rangle \middle| \mathbf{F}_h, a_h, \mathbf{O}_{h+1}\right] \middle| \mathbf{F}_h = f_h\right]\right\} \\ =& U^{-1}\max_{a_h}\left\{\mathbb{E}_{\mathcal{P}'}\left[\max_{a_{h+1:H}} \mathbb{E}_{\mathcal{P}'}\left[\langle \vec{\sigma}_{h+1,\mathbf{F}_h,a_h,\mathbf{O}_{h+1}}, \vec{\nu}_{h+1,a_{h+1},\mathbf{O}_{h+2},\cdots,\mathbf{O}_H,a_H}\rangle \middle| \mathbf{F}_h, a_h, \mathbf{O}_{h+1}\right] \middle| \mathbf{F}_h = f_h\right]\right\} \\ =& U^{-1}\max_{a_h} \mathbb{E}_{\mathcal{P}'}[UV_{h+1}^\star(\mathbf{F}_{h+1} = (\mathbf{F}_h, a_h, \mathbf{O}_{h+1}) \big| \mathbf{F}_h = f_h] \\ =& U^{-1}\max_{a_h} \mathbb{E}_{\mathcal{P}'}[UV_{h+1}^\star(f_h, a_h, \mathbf{O}_{h+1})] \text{//Section B.1} \end{aligned}$$

□

The proof is inspired by (James et al., 1994; Bäauerle & Rieder, 2017) and we have generalized their result beyond Gaussian transition matrices and the entropic risk. We will present the proof in the tabular case. Regularity conditions will be needed in the sixth step when we generalize the theorem to the continuous setting.

Our proof also yields the following corollary, providing justification for selecting greedy policies in our algorithm:

Corollary B.24. *(Adapted from Theorem 3.3 of (Bäauerle & Rieder, 2017)) There always exists an optimal policy for a risk-sensitive tabular POMDP using utility risk measure, which is deterministic and history-dependent.*

B.6. Beta Vectors

Inspired by the alpha vector representation method in the study of POMDP((Monahan, 1982)), we will exploit the structure of risk-sensitive value functions and represent them in a simple form. Recall that in Definition B.18, the value functions are specified as inner products. Since $\sigma(\mathbf{F}_h)$ is already determined by the condition on f_h, we can write the value function as $\mathsf{V}_h^\pi(f_h) = U^{-1}\left\langle \vec{\sigma}_{h,f_h}, \mathbb{E}_{\mathcal{P}'}^\pi\left[\vec{\nu}_{h,\bar{\mathbf{F}}_h} \mid \mathbf{F}_h = f_h\right]\right\rangle$. , which motivates us to introduce the concept of "beta" vector.

Definition B.25. (Beta vector) The beta vector of a risk-sensitive POMDP model $\mathcal{P} = (\mu_1, \mathbb{T}, \mathbb{O})$ under policy π is a series of random vectors in $\mathbb{R}^S$, which are specified as

$$\begin{aligned} \vec{\beta}_{H+1,\boldsymbol{F_{H+1}}}^\pi :=& \vec{1}_S \\ \vec{\beta}_{h,\boldsymbol{F_h}}^\pi :=& \mathbb{E}_{\mathcal{P}'}^\pi\left[\vec{\nu}_{h,\bar{\boldsymbol{F}}_{\boldsymbol{h}}} | \boldsymbol{F_h}\right], \quad \forall 2 \le h \le H. \\ \vec{\beta}_1^\pi :=& \mathbb{E}_{\mathcal{P}'}^\pi\left[\vec{\nu}_{1,\boldsymbol{F_1}}\right] \end{aligned} \tag{51}$$

where ν_t is the conjugate belief defined in B.15.

Next, we will try to obtain the evolution law of the beta vector from the way $\vec{\nu}_h$ is updated:

$$[\vec{\nu}_{h,\bar{f}_h}]_{s_h} = \frac{1}{\mathbb{O}'_1(o_1)} \cdot e^{\gamma r_h(s_h,a_h)} \cdot \sum_{s_{h+1}\in\mathscr{S}} \mathbb{T}_{h,a_h}(s_{h+1}|s_h)\mathbb{O}_{h+1}(o_{h+1}|s_{h+1})[\vec{\nu}_{h+1,\bar{f}_{h+1}}]_{s_{h+1}} \tag{52}$$

Under the reference model $\mathcal{P}'$, we can compute the probability of witnessing an observable trajectory $\bar{f}_h$ given previous history f_h by the following equation:

$$\mathbb{P}^\pi_{\mathcal{P}'}(\bar{f}_h|f_h) = \pi_h(a_h|f_h)\mathbb{O}'_1(o_1)\mathbb{P}^\pi_{\mathcal{P}'}(\bar{f}_{h+1}|f_h,a_h,o_{h+1}) = (\mathbb{O}'_1(o_1))^{H-h+1} \cdot \prod_{t=h}^{H}\pi_h(a_t|f_t) \tag{53}$$

where $\bar{f}_h = (a_h, o_{h+1}, \bar{f}_{h+1}) = (a_h, o_{h+1}, \cdots, a_{H-1}, o_H, a_H)$. Combining Eqs.(53), (52), we obtain

$$\begin{aligned}
&\mathbb{E}^\pi_{\mathcal{P}'}[\nu_{h,\bar{\mathbf{F}}_\mathbf{h}}|f_h] \\
=&\sum_{a_h}\pi_h(a_h|f_h)\sum_{o_{h+1}}\mathbb{O}'_1(o_1)\sum_{\bar{f}_{h+1}}\mathbb{P}^\pi_{\mathcal{P}'}(\bar{f}_{h+1}|f_h,a_h,o_{h+1}) \\
&\frac{1}{\mathbb{O}'_1(o_1)}\cdot e^{\gamma r_h(s_h,a_h)}\cdot\sum_{s_{h+1}\in\mathscr{S}}\mathbb{T}_{h,a_h}(s_{h+1}|s_h)\mathbb{O}_{h+1}(o_{h+1}|s_{h+1})[\vec{\nu}_{h+1}(\bar{f}_{h+1})]_{s_{h+1}} \\
=&\sum_{a_h\in\mathscr{A}}\pi_h(a_h|f_h)\left\{e^{\gamma r_h(s_h,a_h)}\sum_{s_{h+1}\in\mathscr{S}}\mathbb{T}_{h,a_h}(s_{h+1}|s_h)\sum_{o_{h+1}\in\mathscr{O}}\mathbb{O}_{h+1}(o_{h+1}|s_{h+1})\mathbb{E}^\pi_{\mathcal{P}'}\left[\vec{\nu}_{h+1,\bar{\mathbf{F}}_{h+1}}|f_{h+1}\right]_{s_{h+1}}\right\}
\end{aligned} \tag{54}$$

The previous derivations leads to the fundamental theorem below, which forms the cornerstone of our subsequent analysis:

Theorem B.26. *(β-vector representation of value functions) For any $h\in\{1,\cdots,H+1\}$ and $f_h\in\mathscr{F}_h$, the value function defined in B.18 can be expressed as the inner product between the risk-sensitive beliefs defined in B.3 and the beta vector defined in B.25:*

$$\mathsf{V}^\pi_1 = U^{-1}\langle\vec{\sigma}_1\ ,\ \vec{\beta}^\pi_1\rangle = \frac{1}{\gamma}\ln\langle\vec{\sigma}_1\ ,\ \vec{\beta}^\pi_1\rangle$$

$$\mathsf{V}^\pi_h(f_h) = U^{-1}\langle\vec{\sigma}_{h,f_h}\ ,\ \vec{\beta}^\pi_{h,f_h}\rangle = \frac{1}{\gamma}\ln\langle\vec{\sigma}_{h,f_h}\ ,\ \vec{\beta}^\pi_{h,f_h}\rangle$$

Moreover, the beta vectors evolve by the following rule: For all $2\le h\le H$, $f_h\in\mathscr{F}_h$, $s_h\in\mathscr{S}$

$$\begin{aligned}
\vec{\beta}_{H+1,f_{H+1}} =&\vec{1}_S \\
\left[\vec{\beta}_{h,f_h}\right]_{s_h} =& \mathbb{E}_{a_h\sim\pi_h(\cdot|f_h)}e^{\gamma r_h(s_h,a_h)}\sum_{s_{h+1}}\mathbb{T}_{h,a_h}(s_{h+1}|s_h)\sum_{o_{h+1}}\mathbb{O}_{h+1}(o_{h+1}|s_{h+1})\left[\vec{\beta}_{h+1,f_{h+1}=(f_h,a_h,o_{h+1})}\right]_{s_{h+1}}
\end{aligned} \tag{55}$$

We remind the reader that this theorem holds for randomized policies.

Remark B.27. When it is clear from the context, we omit the policy sign $^\pi$ for the beta vectors. We can also define the beta vectors for the empirical POMDP model by replacing the matrices $\mathbb{T}$ and $\mathbb{O}$ with their empirical approximations $\widehat{\mathbb{T}}^k$ and $\widehat{\mathbb{O}}^k$. The corresponding beta vector will be called the "empirical beta vector", which is denoted as $\widehat{\beta}^{k,\pi}_{h,f_h}$ and abbreviated as $\widehat{\beta}_h$.

Remark B.28. (Motivation behind the beta vector) According to (Monahan, 1982; Pineau et al., 2006; Lee et al., 2023) the value function of a risk-neutral POMDP is the inner product of the risk-neutral belief $\vec{b}_h$ and another function named the "alpha vector". The concept of $\alpha-$vector has been widely adopted in the algorithm design of POMDPs (Pineau et al., 2006).

$$V^\pi_h(\vec{b}_h;f_h) = \mathbb{E}_{a_h\sim\pi(\cdot|f_h)}\left[r(\vec{b}_h,a_h;f_h) + \sum_{o_{h+1}\in\mathscr{O}}\eta_h(o_{h+1}|f_h,a_h)V^\pi_{h+1}(\vec{b}_{h+1};f_{h+1}=(f_h,a_h,o_{h+1}))\right]$$

$$V^\pi_h(\vec{b}_h;f_h) = \langle\vec{b}_h,\alpha^\pi_{h,f_h}\rangle$$

$$\begin{cases}\{\vec{\alpha}^\pi_{H+1,f_{H+1}}\equiv 0\\ \left[\vec{\alpha}^\pi_{h,f_h}\right]_{s_h} = \mathbb{E}_{a_h\sim\pi(\cdot|f_h)}\left[r(s_h,a_h)+\sum_{s_{h+1}}\mathbb{T}_{h,a_h}(s_{h+1}|s_h)\sum_{o_{h+1}}\mathbb{O}_{h+1}(o_{h+1}|s_{h+1})\left[\vec{\alpha}^\pi_{h+1,f_{h+1}=(f_h,a_h,o_{h+1})}\right]_{s_{h+1}}\right]\end{cases}$$

Though the update rule of the value function inevitably relies on the entire history, the alpha vectors evolve in a Markovian way. This finding helps us obtain a polynomial regret bound in risk-neutral POMDP using hindsight observations, which is discovered by (Lee et al., 2023). To gain some insights, recall that the pigeon-hole lemma G.12 suggests that the sum of concentration errors is the polynomial function of the cardinality of the space on which the transition probability is dependent.

$$\widehat{P}_h(s'|s,a) := \frac{\sum_{\kappa=1}^{k} \mathbb{1}\{\hat{s}_{h+1}^{\kappa} = s' \ , \ \hat{s}_h^{\kappa} = s \ , \ \hat{a}_h^{\kappa} = a\}}{\max\left\{1, \widehat{N}_h^{k+1}(s,a)\right\}} \quad \sum_{k=1}^{K} \frac{1}{\sqrt{\max\{1, N_h^{k+1}(\hat{s}_h^k, \hat{a}_h^k)\}}} < 2\sqrt{K \cdot SA}$$

In the POMDP setting, if we refuse to represent the value function by the alpha vector, the upper bound in the right should be replaced with $2\sqrt{K \cdot |\mathscr{F}_h \times \mathscr{A}|} = 2O^{\frac{h}{2}} A^{\frac{h}{2}}$, which then brings a factor of $O^H A^H$ to our regret. However, under hindsight observability, since we can calculate the occurrences of the hidden states after each episode, we use the alpha vectors to represent the value function. Consequently, we can replace $O^H A^H$ with $2\sqrt{S^2}$, which is polynomial in H again.

However, we cannot directly utilize the alpha vector in the risk-sensitive setting, as it may no longer preserve the Markov property. For the reader's convenience, we excerpt the critical steps in the proof of the evolution law (Monahan, 1982)

$$\begin{aligned}
&= \langle \vec{b}_h(f_h), r_h(\cdot, a_h) \rangle + \boldsymbol{\rho}_{o_{h+1} \sim \eta_{h+1}(\cdot|f_h,a_h)} \left[\left\langle \frac{\operatorname{diag}\left(\mathbb{O}_{h+1}(o_{h+1}|\cdot)\right) \ \mathbb{T}_{h,a_h} \ \vec{b}_h(f_h)}{\eta_{h+1}(o_{h+1}|f_h, a_h)}, \alpha^{\pi}_{h+1,f_{h+1}} \right\rangle \right] \\
&= \langle \vec{b}_h(f_h), r_h(\cdot, a_h) \rangle + \left\langle \sum_{o_{h+1} \in \mathscr{O}} \cancel{\eta_{h+1}(o_{h+1}|f_h, a_h)} \frac{\operatorname{diag}\left(\mathbb{O}_{h+1}(o_{h+1}|\cdot)\right) \ \mathbb{T}_{h,a_h} \ \vec{b}_h(f_h)}{\cancel{\eta_{h+1}(o_{h+1}|f_h, a_h)}}, \alpha^{\pi}_{h+1,f_{h+1}} \right\rangle
\end{aligned}$$

In the risk-neutral setting, the risk measure ρ is represented by the expectation operator $\mathbb{E}$. The simple linear structure of $\mathbb{E}$ helps to cancel the partition function $\eta_{h+1}(\cdot|f_h, a_h)$ when the alpha vector updates, thus eliminating the culprit of the exponential dependency on H. However, the previous analysis becomes invalid when using a non-linear risk measure. That's why the beta vector is introduced in risk-sensitive POMDPs to mitigate the impact of historical dependencies under a non-linear criterion (WATER & Willems, 1981; Whittle, 1991; Baras & James, 1997).

Remark B.29. (Comparison with the alpha vector in risk-neutral POMDP) The value function can be expressed as an inner product in risk-sensitive settings, because the utility risk measure keeps a layer of $\mathbb{E}$, which still bears a linear structure. Both α and beta vectors evolve in a Markovian way, making it possible to control the regret under a polynomial upper bound given hindsight observability. However, the terminal value of $\vec{\beta}_{h,f_h}$ is $\vec{1}$ while that of $\vec{\alpha}$ is $\vec{0}$; The beta vector undergoes multiplicative updates, whereas the alpha vector renews through additive increments. If we brutally set $\gamma = 0$ we cannot reduce the beta vector to the alpha vector, as $\gamma = 0$ is a singular point of the risk measure, which will cause all the beta vectors to collapse to $\vec{1}$. We will introduce the correct way to degenerate our result to the classical setting in Section D.4.4.

C. Detailed Algorithm Design

In what follows, we present the algorithm introduced in Section 5 in detail, along with several remarks additional to the discussion in Section 5.

Remark C.1. (Computation issues) The BVVI algorithm, as well as other exact algorithms for POMDP, are inefficient in computation complexity(Cassandra, 1998), which is due to the inherent complexity of the POMDP model(Papadimitriou & Tsitsiklis, 1987). Using similar techniques as the point-based algorithms(Pineau et al., 2006), we can develop approximate solutions to our problem based on BVVI.

Remark C.2. (Explanation of line 2) The operation in line 2 is equivalent to the assignment below:

$$\forall s \in \mathscr{S}: \quad \left[\widehat{\beta}^k_{h,f_h}\right]_s \leftarrow \begin{cases} e^{\gamma^+(H-h+1)} & , \left[\widehat{\beta}^k_{h,f_h}\right]_s \geq e^{\gamma^+(H-h+1)} \\ e^{\gamma^-(H-h+1)} & , \left[\widehat{\beta}^k_{h,f_h}\right]_s \leq e^{\gamma^-(H-h+1)} \\ \left[\widehat{\beta}^k_{h,f_h}\right]_s & , \text{else} \end{cases} \tag{56}$$

Algorithm 2 Beta Vector Value Iteration(BVVI)

Input risk sensitivity $\gamma \neq 0$, confidence level $\delta \in (0,1)$, episode number K, horizon length H.
Initialize $\widehat{\mu}_1^k(\cdot), \widehat{\mathbb{T}}_{h,a}^1(\cdot|s) \leftarrow \text{Unif}(\mathscr{S}), \quad \widehat{\mathbb{O}}_h^1(\cdot|s) \leftarrow \text{Unif}(\mathscr{O})$ for all $(h,s,a,o,s') \in [H] \times \mathscr{S} \times \mathscr{A} \times \mathscr{O} \times \mathscr{S}$.
for $k = 1 : K$ **do**
  //Planning
  //Forward belief propagation
  $\widehat{\sigma}_1^k \leftarrow \widehat{\mu}_1^k$
  for $h = 1 : H, \quad s_{h+1}, s_h \in \mathscr{S}, \quad f_{h+1} = (f_h, a_h, o_{h+1}) \in \mathscr{F}_h \times \mathscr{A} \times \mathscr{O}$ **do**
    //Update risk belief by Eq. (35)
    $\left[\widehat{\sigma}_{h+1,f_{h+1}}^k\right]_{s_{h+1}} \leftarrow \sum_{s_h} \widehat{\mathbb{T}}_{h,a_h}^k(s_{h+1}|s_h)\widehat{\mathbb{O}}_{h+1}^k(o_{h+1}|s_{h+1}) \cdot \left(\frac{e^{\gamma r_h(s_h,a_h)}}{\mathbb{O}'_{h+1}(o_{h+1}|s_{h+1})}\right) \left[\widehat{\sigma}_{h,f_h}^k\right]_{s_h}$
    //Residue terms
    $\mathsf{t}_h^k(s_h,a_h) \leftarrow \min\left\{1, 3\sqrt{\frac{SH\ln KHSOA/\delta}{\widehat{N}_h^k(s_h,a_h)\vee 1}}\right\}$
    $\mathsf{o}_{h+1}^k(s_{h+1}) \leftarrow \min\left\{1, 3\sqrt{\frac{OH\ln KHSOA/\delta}{\widehat{N}_{h+1}^{k+1}(s_{h+1})\vee 1}}\right\}$
    //Prepare bonus by Eq. (62)
    $\mathsf{b}_h^k(s_h,a_h) \leftarrow \left|e^{\gamma(H-h+1)} - 1\right| \cdot \min\left\{1,\ \mathsf{t}_h^k(s_h,a_h) + \sum_{s_{h+1}} \widehat{\mathbb{T}}_{h,a_h}^k(s_{h+1}|s_h)\mathsf{o}_{h+1}^k(s_{h+1})\right\}$
  end for
  //Backward dynamic programming
  $\widehat{\beta}_{H+1,f_{H+1}}^k \leftarrow \vec{1}_S$
  for $h = H : 1$ **do**
    for $f_h = (a_1, o_2, \cdots, a_h, o_h) \in \mathscr{F}_h,\ a_h \in \mathscr{A}$ **do**
      //Invoke Bellman equation (49) under beta vector representation
      $\widehat{\mathsf{Q}}_h^k(a_h; f_h) \leftarrow \frac{1}{\gamma}\ln \mathbb{E}_{\boldsymbol{O'}\sim\text{Unif}\mathscr{O}}\left\langle \widehat{\sigma}_{h+1,f_{h+1}=(f_h,a_h,\boldsymbol{O'})}^k,\ \widehat{\beta}_{h+1,f_{h+1}=(f_h,a_h,\boldsymbol{O'})}^k \right\rangle$
      $\widehat{\mathsf{V}}_h^k(f_h) \leftarrow \max_{a\in\mathscr{A}} \widehat{\mathsf{Q}}_h^k(a_h; f_h)$
      $\widehat{\pi}_h^k(f_h) \leftarrow \operatorname{argmax}_{a\in\mathscr{A}} \widehat{\mathsf{Q}}_h^k(a; f_h)$ // Obtain the greedy policy
      //Update beta vector by Eq. (63)

$$\widehat{\beta}_{h,f_h}^k(s_h) \leftarrow e^{\gamma r_h(s_h,\widehat{\pi}_h^k(f_h))} \sum_{s_{h+1}} \widehat{\mathbb{T}}_{h,\widehat{\pi}_h^k(f_h)}^k(s_{h+1}|s_h) \sum_{o_{h+1}} \widehat{\mathbb{O}}_{h+1}^k(o_{h+1}|s_{h+1}) \left[\widehat{\beta}_{h+1,f_{h+1}=(f_h,\widehat{\pi}_h^k(f_h),o_{h+1})}^k\right]_{s_{h+1}} + \text{sgn}(\boldsymbol{\gamma}) \cdot \mathsf{b}_{h+1}^k(s_h, \widehat{\pi}_h^k(f_h))$$

      //Control the range of beta vector
      $\widehat{\beta}_{h,f_h}^k \leftarrow \text{Clip}\left(\widehat{\beta}_{h,f_h}^k,\ \left[e^{\gamma^-(H-h+1)}, e^{\gamma^+(H-h+1)}\right]\right)$
    end for
  end for
  //Learning
  Play with the real environment under policy $\{\widehat{\pi}_h^k\}_{h=1}^H$ and collect a trajectory $(\hat{a}_1^k, \hat{o}_2^k, \ldots, \hat{o}_H^k, \hat{a}_H^k)$
  Reveal the hidden states $\{\hat{s}_h^k\}_{h=1}^H$ in the previous H steps to the agent. //Hindsight observation
  // Update the empirical model
  for $h = 1 : H$ **do**
    for $(s,a,o,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{O} \times \mathscr{S}$ **do**
      $\widehat{N}_h^{k+1}(s) \leftarrow \sum_{\kappa=1}^k \mathbb{1}\{\hat{s}_h^\kappa = s\} \qquad \widehat{N}_h^{k+1}(s,a) \leftarrow \sum_{\kappa=1}^k \mathbb{1}\{\hat{s}_h^\kappa = s, \hat{a}_h^\kappa = a\}$
      $\widehat{\mu}_1^k(s) \leftarrow \frac{1}{k}\sum_{\kappa=1}^k \mathbb{1}\{\hat{s}_1^\kappa = s\}$
      $\widehat{\mathbb{T}}_{h,a}^{k+1}(s'|s) \leftarrow \frac{\sum_{\kappa=1}^k \mathbb{1}\{\hat{s}_{h+1}^\kappa = s',\ \hat{s}_h^\kappa = s,\ \hat{a}_h^\kappa = a\}}{\max\{1, \widehat{N}_h^{k+1}(s,a)\}} \quad \widehat{\mathbb{O}}_h^{k+1}(o|s) \leftarrow \frac{\sum_{\kappa=1}^k \mathbb{1}\{\hat{o}_h^\kappa = o,\ \hat{s}_h^\kappa = s\}}{\max\{1, \widehat{N}_h^{k+1}(s)\}}$
    end for
  end for
end for

D. Regret analysis

Our regret analysis takes several steps. First, we study the statistical error of the beta vectors and design a new bonus for our problem, which ensures optimism in value functions. Later, we represent the regret by the beta vectors before we roll down the Bellman equations to calculate the error accumulated during the entire Markov process.

Several facts will be repeatedly used in the proceeding derivations.

D.1. Preparation

The following result captures the empirical error of emission matrix $\mathbb{O}_h(\cdot|s_h)$:

Fact D.1. (Empirical error of the emission matrix) With probability at least $1-\delta$,

$$\|\widehat{\mathbb{O}}_h^k(\cdot|s_h) - \mathbb{O}_h(\cdot|s_h)\|_1 \leq \min\left\{2\,,\ \sqrt{\frac{2O\ln\frac{KHS}{\delta}}{N_h^{k+1}(s_h)\vee 1}}\right\} \tag{57}$$

The proof is similar to Lemma C.1 in (Liang & Luo, 2023). This relation holds without hindsight observability.

The upper and lower bounds for the beta vectors will also be useful.

Fact D.2. (Bounds of the beta vectors) $e^{(H-h+1)\gamma^-} \leq \left[\vec{\beta}_{h,f_h}\right]_{s_h} \leq e^{(H-h+1)\gamma^+}$ for all $f_h \in \mathscr{F}_h$ and $s_h \in \mathscr{S}$.

The proof is done by a simple induction on h.

D.2. Optimism

D.2.1. Statistical Error of The Beta vector

In this section, we quantify the error in beta vectors caused by the inaccurate approximation of the transition and emission probabilities. For the convenience of narration, we will temporarily view the beta vector as a binary function over the space of $\mathscr{S}\times\mathscr{O}$. Since our result holds for any given episode, we omit the subscripts k in the following derivations.

The update rule of the beta vectors in (55) is equivalent to

$$\begin{aligned}\vec{\beta}_{h,f_h}(s_h) =& e^{\gamma r_h(s_h,\pi_h(f_h))}\sum_{\mathbf{s_{h+1}}}\mathbb{T}_{h,\pi_h(f_h)}(\mathbf{s_{h+1}}|s_h)\sum_{\mathbf{o_{h+1}}}\mathbb{O}_{h+1}(\mathbf{o_{h+1}}|\mathbf{s_{h+1}})\vec{\beta}_{h+1,f_{h+1}=(f_h,\pi_h(f_h),\mathbf{o_{h+1}})}(\mathbf{s_{h+1}})\\ =&\mathbb{E}_{\mathbf{S}_{h+1}\sim\mathbb{T}_{h,\pi_h(f_h)}(\cdot|s_h),\mathbf{O_{h+1}}\sim\mathbb{O}_{h+1}(\cdot|\mathbf{s_{h+1}})}\left[e^{\gamma r_h(s_h,\pi_h(f_h))}\vec{\beta}_{h+1,f_{h+1}=(f_h,\pi_h(f_h),\mathbf{O}_{h+1})}(\mathbf{S}_{h+1})\right]\end{aligned}$$

During the update of beta vectors we view $\vec{\beta}_{h,f_h}(s_h)$ as a binary functions over the space of $\mathscr{Z} := \mathscr{S}\times\mathscr{O}$. With a slight abuse of notations, we denote (s_h,o_h) as z_h and write $\vec{\beta}_{h,f_h=(\tau_{h-1},o_h)}(s_h)$ as $\vec{\beta}_h(z_h;\tau_{h-1})$. We also abbreviate $r_h(s_h,\pi_h(f_h))$ as $\mathsf{r}_h(z_h)$. The transition law of the beta vector can be written as a joint distribution over the space of $\mathscr{S}\times\mathscr{O}$: $\mathbb{P}_h^{\pi_h}(z_{h+1}|s_h) := \mathbb{T}_{h,\pi_h(f_h)}(\mathbf{s_{h+1}}|s_h)\mathbb{O}_{h+1}(\mathbf{o_{h+1}}|s_{h+1})$

Using the newly introduced short hands we can write the update rule of beta vectors as

$$\vec{\beta}_h(z_h;\tau_{h-1}) =\mathbb{E}_{\mathbb{P}_h^{\pi_h}(\cdot|s_h)}\left[e^{\gamma\mathsf{r}_h(z_h)}\vec{\beta}_{h+1}(\mathbf{Z}_{h+1};\tau_h)\right] = \mathbb{E}_{\mathbb{P}_h^{\pi_h}}\left[e^{\gamma\left(\mathsf{r}_h(z_h)+\frac{\ln\vec{\beta}_{h+1}(\mathbf{Z}_{h+1};\tau_h)}{\gamma}\right)}\right] = \mathbb{E}_{\mathbb{P}_h^{\pi_h}}\left[e^{\gamma\mathcal{V}_h(\mathbf{Z}_{h+1};\tau_h)}\right]$$

We remind the reader that we have introduced a new function $\mathcal{V}_h(z_{h+1})$ to abbreviate $\mathsf{r}_h(z_h)+\frac{\ln\vec{\beta}_{h+1}(z_{h+1};\tau_h)}{\gamma}$. The function $\mathcal{V}_h$ is bounded in $[1,H-h+1]$ according to Lemma D.2. Next, we will calculate the approximation error occurred when

the beta vector updates. The error is caused by the inaccurate estimate of the joint transition law.

$$
\begin{aligned}
&\left|\left(\mathbb{E}_{\widehat{\mathbb{P}}_h^{\pi_h}(\cdot|s_h)}-\mathbb{E}_{\mathbb{P}_h^{\pi_h}(\cdot|s_h)}\right)\left(e^{\gamma r_h}\vec{\beta}_{h+1}(\cdot;\tau_h)\right)\right| \\
=&\left|\mathbb{E}_{\mathbf{Z}_{h+1}\sim\widehat{\mathbb{P}}_h^{\pi_h}(\cdot|s_h)}[e^{\gamma\mathcal{V}_h(\mathbf{Z}_{h+1};\tau_h)}]-\mathbb{E}_{\mathbf{Z}_{h+1}\sim\mathbb{P}_h^{\pi_h}(\cdot|s_h)}[e^{\gamma\mathcal{V}_h(\mathbf{Z}_{h+1};\tau_h)}]\right| \\
=&\left|\left(\sum_{s_{h+1}}\widehat{\mathbb{T}}_{h,a_h}(s_{h+1}|s_h)\sum_{o_{h+1}}\widehat{\mathbb{O}}_{h+1}(o_{h+1}|s_{h+1})-\sum_{s_{h+1}}\mathbb{T}_{h,a_h}(s_{h+1}|s_h)\sum_{o_{h+1}}\mathbb{O}_{h+1}(o_{h+1}|s_{h+1})\right)e^{\gamma\mathcal{V}_{h+1}(s_{h+1},o_{h+1};\tau_h)}\right| \\
=&\sum_{s_{h+1}}\widehat{\mathbb{T}}_{h,a_h}(s_{h+1}|s_h)\left|\sum_{o_{h+1}}\widehat{\mathbb{O}}_{h+1}(o_{h+1}|s_{h+1})-\mathbb{O}_{h+1}(o_{h+1}|s_{h+1})e^{\gamma\mathcal{V}_{h+1}(s_{h+1},o_{h+1};\tau_h)}\right| \\
+&\left|\sum_{s_{h+1}}\left(\widehat{\mathbb{T}}_{h,a_h}(s_{h+1}|s_h)-\mathbb{T}_{h,a_h}(s_{h+1}|s_h)\right)\cdot\sum_{o_{h+1}}\mathbb{O}_{h+1}(o_{h+1}|s_{h+1})e^{\gamma\mathcal{V}_{h+1}(s_{h+1},o_{h+1};\tau_h)}\right|
\end{aligned}
\tag{58}
$$

Lemma G.6 implies a natural upper bound for the error in Eq.(58):

$$
\begin{aligned}
&\left|\left(\mathbb{E}_{\widehat{\mathbb{P}}_h^{\pi_h}(\cdot|s_h)}-\mathbb{E}_{\mathbb{P}_h^{\pi_h}(\cdot|s_h)}\right)\left(e^{\gamma r_h}\vec{\beta}_{h+1}(\cdot;\tau_h)\right)\right| \\
\le&\left|e^{\gamma^+(H-h+1)}-e^{\gamma^-(H-h+1)}\right|\cdot\|\widehat{\mathbb{P}}_h^{\pi_h}-\mathbb{P}_h^{\pi_h}\|_{tv}\le\left|e^{\gamma(H-h+1)}-1\right|
\end{aligned}
\tag{59}
$$

We can also use Lemma G.10 to derive another upper bound. With probability at least $1-2\delta$,

$$
\begin{aligned}
&\left|\left(\mathbb{E}_{\widehat{\mathbb{P}}_h^{\pi_h}(\cdot|s_h)}-\mathbb{E}_{\mathbb{P}_h^{\pi_h}(\cdot|s_h)}\right)\left(e^{\gamma r_h(\cdot)}\vec{\beta}_{h+1}(\cdot;\tau_h)\right)\right| \\
\le&\left|e^{\gamma(H-h+1)}-1\right|\cdot\left[\min\left\{1,\ 3\sqrt{\frac{S\cdot H\ln\frac{KHSOA}{\delta}}{\widehat{N}_h^k(s_h,\widehat{\pi}_h^k(s_h))\vee 1}}\right\}\right. \\
&\left.+\sum_{s_{h+1}}\widehat{\mathbb{T}}^k_{h,\widehat{\pi}_h^k(s_h)}(s_{h+1}|s_h)\min\left\{1,\ 3\sqrt{\frac{O\cdot H\ln\frac{KHSOA}{\delta}}{\widehat{N}_{h+1}^{k+1}(s_{h+1})\vee 1}}\right\}\right]
\end{aligned}
\tag{60}
$$

The additional $\sqrt{H}$ comes from the inherent history-dependency of the of POMDP:

$$
\sqrt{\ln\frac{KHO(O^hA^h)}{\delta}},\sqrt{\ln\frac{KHSA(O^hA^h)}{\delta}}<\sqrt{H\frac{\ln KHSOA}{\delta}}
$$

.

Putting Eq.(59) and (60) together, we conclude that with probability at least $1-2\delta$:

$$
\begin{aligned}
&\left(\sum_{s_{h+1}}\widehat{\mathbb{T}}^k_{h,\widehat{\pi}_h^k(s_h)}(s_{h+1}|s_h)\sum_{o_{h+1}}\widehat{\mathbb{O}}^k_{h+1}(o_{h+1}|s_{h+1})\right. \\
&\left.-\sum_{s_{h+1}}\mathbb{T}_{h,\pi_h(f_h)}(s_{h+1}|s_h)\sum_{o_{h+1}}\mathbb{O}_{h+1}(o_{h+1}|s_{h+1})\right)e^{\gamma r_h(s_h,\widehat{\pi}_h(s_h))}\vec{\beta}_{h+1,f_{h+1}=(f_h,\pi_h(f_h),o_{h+1})}(s_{h+1}) \\
\le&\left|e^{\gamma(H-h+1)}-1\right|\min\left\{1,\ \min\left\{1,\ 3\sqrt{\frac{S\cdot H\ln\frac{KHSOA}{\delta}}{\widehat{N}_h^k(s_h,\widehat{\pi}_h^k(s_h))\vee 1}}\right\}\right. \\
&\left.+\sum_{s_{h+1}}\widehat{\mathbb{T}}^k_{h,\widehat{\pi}_h^k(s_h)}(s_{h+1}|s_h)\min\left\{1,\ 3\sqrt{\frac{O\cdot H\ln\frac{KHSOA}{\delta}}{\widehat{N}_{h+1}^{k+1}(s_{h+1})\vee 1}}\right\}\right\}
\end{aligned}
\tag{61}
$$

We design the exploration bonus according to Eq.(61).

Definition D.3. (Bonus) The bonuses are a series of real-valued functions $\mathsf{b}_h^k(\cdot,\cdot;\gamma):\mathscr{S}\times\mathscr{A}\to\mathbb{R}_{\geq 0}$ specified as

$$\mathsf{b}_h^k(s_h,a_h;\gamma):=\left|e^{\gamma(H-h+1)}-1\right|\cdot\min\left\{1,\ \min\left\{1,\ 3\sqrt{\frac{S\cdot H\ln\frac{KHSOA}{\delta}}{\widehat{N}_h^k(s_h,a_h)\vee 1}}\right\}\right.$$
$$\left.+\sum_{s_{h+1}}\widehat{\mathbb{T}}^k_{h,\widehat{\pi}_h^k(s_h)}(s_{h+1}|s_h)\min\left\{1,\ 3\sqrt{\frac{O\cdot H\ln\frac{KHSOA}{\delta}}{\widehat{N}_{h+1}^{k+1}(s_{h+1})\vee 1}}\right\}\right\} \tag{62}$$

Previously, we have analyzed the upper bound of the approximation error incurred during the update process. We will utilize the relevant results and define the series of "empirical beta vectors" $\widehat{\beta}_h$ that helps approximate the value function.

Definition D.4. (Empirical beta vectors) Given any episode $k\in[K]$, if we use a_h to abbreviate the action selected by the greedy policy $\widehat{\pi}_h^k(f_h)$ given by algorithm 2, then the empirical beta vector is defined as

$$\begin{aligned}
&\forall f_{H+1}\in\mathscr{F}_{H+1},s_{H+1}\in\mathscr{S}:\quad \widehat{\beta}^k_{H+1,f_{H+1}}(s_{H+1}):=1\\
&\forall h\in[H],f_h\in\mathscr{F}_h,s_h\in\mathscr{S},\\
&\quad\widehat{\beta}^k_{h,f_h}(s_h):=e^{\gamma r_h(s_h,a_h)}\sum_{s_{h+1}\in\mathscr{S}}\widehat{\mathbb{T}}^k_{h,a_h}(s_{h+1}|s_h)\sum o_{h+1}\in\mathscr{O}\widehat{\mathbb{O}}^k_{h+1}(o_{h+1}|s_{h+1})\widehat{\beta}^k_{h+1,f_{h+1}=(f_h,a_h,o_{h+1})}(s_{h+1})\\
&\qquad\qquad+\operatorname{sgn}\gamma\cdot\mathsf{b}_h^k(s_h,a_h;\gamma)
\end{aligned}\tag{63}$$

Remark D.5. (Range of the empirical beta vectors) If we pose no further restrictions on the range of $\widehat{\beta}_h^k$ we can show that the upper bound on $\widehat{\beta}_h^k$ will inevitably depend on $e^{\gamma H^2}$, which will cause an additional factor of H in our regret. To circumvent this issue we have manually clipped the value of empirical beta vector in our algorithm(line 2), so as to force $\widehat{\beta}_h$ to stay in the same range as $\vec{\beta}_h$. As a consequence, the difference between the two beta vectors, which will be called the "beta-vector error", is controlled by

$$\left|\widehat{\beta}_h^k-\vec{\beta}_h\right|\leq\left|e^{\gamma(H-h+1)}-1\right| \tag{64}$$

Apart from the absolute error between the beta vectors, we are also concerned with their size relationship.

Corollary D.6. *(Size relationship between beta vectors)*
For all $s_h\in\mathscr{S}:\quad\widehat{\beta}^k_{h,f_h}(s_h)\geq\vec{\beta}_{h,f_h}(s_h)$ *when* $\gamma>0$ *and* $\widehat{\beta}^k_{h,f_h}(s_h)\leq\vec{\beta}_{h,f_h}(s_h)$ *when* $\gamma<0$.

Proof. We only prove the case when $\gamma<0$. The other way is similar. The statement holds when $h=H+1$ since both vectors are defined to be $\vec{1}_S$. Suppose that it holds at $h+1$, then by the induction hypothesis we have $\widehat{\beta}_h^k\leq\widehat{\mathbb{T}}\widehat{\mathbb{O}}e^{\gamma r_h}\widehat{\beta}^k_{h+1}-\mathsf{b}_h^k$ $=\left[(\widehat{\mathbb{T}}\widehat{\mathbb{O}}-\mathbb{T}\mathbb{O})e^{\gamma r_h}\widehat{\beta}^k_{h+1}-\mathsf{b}_h^k\right]+\mathbb{T}\mathbb{O}\widehat{\beta}^k_{h+1}=\left[(\widehat{\mathbb{T}}\widehat{\mathbb{O}}-\mathbb{T}\mathbb{O})e^{\gamma r_h}\widehat{\beta}^k_{h+1}-\mathsf{b}_h^k\right]+\vec{\beta}_h^k$. Equation (61) implies that the terms in the bracket are less than zero, which helps us complete the proof. □

D.2.2. Optimism in Value Functions

Corollary D.6 directly results in the optimism in value functions: for all $k\in[K]$,

$$V_1^{\pi^\star}\leq\widehat{V}_1^{\widehat{\pi}^k} \tag{65}$$

Proof.

$$V_1^{\pi^\star}-\widehat{V}_1^{\pi^\star}=\frac{1}{\gamma}\ln\langle\vec{\sigma}_1,\vec{\beta}_1\rangle-\frac{1}{\gamma}\ln\langle\widehat{\sigma}_1^k,\widehat{\beta}_1^k\rangle=\frac{1}{\gamma}\ln\langle\vec{\mu}_1,\vec{\beta}_1\rangle-\frac{1}{\gamma}\ln\langle\vec{\mu}_1,\widehat{\beta}_1^k\rangle\leq 0$$

The last step remaining is the fact that $\widehat{V}_1^{\pi^\star}\leq\widehat{V}_1^{\widehat{\pi}^k}$. □

Additional notations For ease of notations we will abbreviate several upper and lower bounds in the following analysis.

$$\text{Upper and lower bounds on the risk measure} \quad \underline{B}_u := e^{-\gamma^-} \le e^{\gamma r_h(s_h,a_h)} \le e^{\gamma^+} := \overline{B_u}$$

$$\text{Upper and lower bounds on the beta-vectors} \quad \underline{B}_{\vec{\beta}_h} := e^{\gamma^-(H-h+1)} \le \vec{\beta}_h \le e^{\gamma^+(H-h+1)} := \overline{B}_{\vec{\beta}_h}$$

$$\text{Upper bound of the bonus} \quad 0 \le \mathsf{b}_h^k(s_h,a_h) \le \left|e^{\gamma(H-h+1)}-1\right| := \overline{B}_{\mathsf{b}_h}$$

$$\text{Upper bound of the "beta vector error"} 0 \le \left|\widehat{\beta}_{h+1}^k - \vec{\beta}_{h+1}\right| \le \left|e^{\gamma(H-h+1)}-1\right| := \overline{B}_{\Delta\vec{\beta}_{h+1}}$$

Remark D.7. According to the update rule in equation (55) and (63), when the risk-sensitivity parameter γ tends to zero, the beta vectors will degenerate to $\vec{1}$. The bonus function, as well as the beta-vector error will vanish with in the rate of H

$$\lim_{\gamma\to 0}\vec{\beta}_h = \lim_{\gamma\to 0}\widehat{\beta}_h = \vec{1} \qquad \lim_{\gamma\to 0}\mathsf{b}_h^k(\cdot,\cdot) = \lim_{\gamma\to 0} B_{\Delta\vec{\beta}_h} = 0 \quad \lim_{\gamma\to 0}\frac{\overline{B}_{\mathsf{b}_h}}{\gamma} = \lim_{\gamma\to 0}\frac{\overline{B}_{\Delta\vec{\beta}_h}}{\gamma} = H \tag{66}$$

which contributes an additional H to our regret.

D.3. Regret Calculation

D.3.1. Represent the Regret by Beta Vectors

The regret can be represented as the approximation error of beta vectors at the initial time step.

$$\begin{aligned}
&\mathrm{Regret}(K;\mathcal{P}) := \sum_{k=1}^K U^{-1}\mathbb{E}_{\mathcal{P}}\left[UV_1^{\pi^\star}\right] - U^{-1}\mathbb{E}_{\mathcal{P}}\left[UV_1^{\widehat{\pi}^k}\right] \\
\le &\sum_{k=1}^K U^{-1}\mathbb{E}_{\mathcal{P}}\left[U\widehat{V}_1^{k,\widehat{\pi}^k}\right] - U^{-1}\mathbb{E}_{\mathcal{P}}\left[UV_1^{\widehat{\pi}^k}\right] \quad \text{//Section D.2.2} \\
= &\sum_{k=1}^K U^{-1}\mathbb{E}_{\mathcal{P}}\left[\langle\widehat{\sigma}_1^k\,,\,\widehat{\beta}_1^{k,\widehat{\pi}^k}\rangle\right] - U^{-1}\mathbb{E}_{\mathcal{P}}\left[\langle\vec{\sigma}_1\,,\,\vec{\beta}_1^{\widehat{\pi}^k}\rangle\right] \quad \text{//Theorem B.26} \\
= &\sum_{k=1}^K \left(U^{-1}\mathbb{E}_{\mathcal{P}}\left[\langle\widehat{\sigma}_1^k\,,\,\widehat{\beta}_1^{k,\widehat{\pi}^k}\rangle\right] - U^{-1}\mathbb{E}_{\mathcal{P}}\left[\langle\vec{\sigma}_1\,,\,\widehat{\beta}_1^{k,\widehat{\pi}^k}\rangle\right]\right) + \left(U^{-1}\mathbb{E}_{\mathcal{P}}\left[\langle\vec{\sigma}_1\,,\,\widehat{\beta}_1^{k,\widehat{\pi}^k}\rangle\right] - U^{-1}\mathbb{E}_{\mathcal{P}}\left[\langle\vec{\sigma}_1\,,\,\vec{\beta}_1^{\widehat{\pi}^k}\rangle\right]\right) \\
= &\sum_{k=1}^K U^{-1}\langle\widehat{\mu}_1^k\,,\,\widehat{\beta}_1^{k,\widehat{\pi}^k}\rangle - U^{-1}\langle\vec{\mu}_1\,,\,\widehat{\beta}_1^{k,\widehat{\pi}^k}\rangle + \sum_{k=1}^K U^{-1}\mathbb{E}_{\mathbf{S}_1\sim\vec{\mu}_1}\left[\widehat{\beta}_1^{k,\widehat{\pi}^k}(\mathbf{S}_1)\right] - U^{-1}\mathbb{E}_{\mathbf{S}_1\sim\vec{\mu}_1}\left[\vec{\beta}_1^{\widehat{\pi}^k}(\mathbf{S}_1)\right] \\
= &\underbrace{\sum_{k=1}^K U^{-1}\mathbb{E}_{\widehat{\mathcal{P}}^k}\left[\widehat{\beta}_1^{k,\widehat{\pi}^k}(\mathbf{S}_1)\right] - U^{-1}\mathbb{E}_{\mathcal{P}}\left[\widehat{\beta}_1^{k,\widehat{\pi}^k}(\mathbf{S}_1)\right]}_{\text{Prior error}} + \underbrace{\sum_{k=1}^K U^{-1}\mathbb{E}_{\mathcal{P}}\left[\widehat{\beta}_1^{k,\widehat{\pi}^k}(\mathbf{S}_1)\right] - U^{-1}\mathbb{E}_{\mathcal{P}}\left[\vec{\beta}_1^{\widehat{\pi}^k}(\mathbf{S}_1)\right]}_{\text{Evolution error}}
\end{aligned} \tag{67}$$

The terms in the last step of Eq. (67) hold significant physical meanings. The first term, referred to as the "prior error", is incurred by the imprecise estimate of the prior distribution of the hidden states. The second term, named "evolution error", represents the accumulation of error throughout the entire horizon of the Markov process, resulting from an inaccurate estimate of the beta vector. In what follows we will try to find upper bounds for the two error terms.

D.3.2. Bound the Prior Error

First, we bound the prior error in Eq. (67). With probability at least $1-\delta$,

$$\begin{aligned}
&\underbrace{\sum_{k=1}^K U^{-1}\mathbb{E}_{\widehat{\mathcal{P}}^k}\left[\widehat{\beta}_1^{k,\widehat{\pi}^k}(\mathbf{S}_1)\right] - U^{-1}\mathbb{E}_{\mathcal{P}}\left[\widehat{\beta}_1^{k,\widehat{\pi}^k}(\mathbf{S}_1)\right]}_{\text{Prior error}} \le \mathsf{K}_\gamma\sum_{k=1}^K\left|\sum_{s_1}\left(\widehat{\mu}_1^k(s_1)-\vec{\mu}_1(s_1)\right)\widehat{\beta}_1^{k,\widehat{\pi}^k}(s_1)\right| \\
\le &\mathsf{K}_\gamma\sum_{k=1}^K\frac{\left|e^{\gamma^+H}-e^{\gamma^-H}\right|}{2}\|\widehat{\mu}_1^k(\cdot)-\vec{\mu}_1(\cdot)\|_1 \le \frac{e^{(-\gamma)^+}}{|\gamma|}\sum_{k=1}^K\frac{\left|e^{\gamma^+H}-e^{\gamma^-H}\right|}{2}\sqrt{\frac{2S}{k}\ln\frac{K}{\delta}} \le \frac{e^{|\gamma|H}-1}{|\gamma|}\sqrt{2KS\ln\frac{K}{\delta}}
\end{aligned} \tag{68}$$

where the second and third inequalities are due to Lemma G.6 and Lemma G.5 respectively.

D.3.3. Control the Evolution Error

Next, we will use the Bellman equations to show that

$$\underbrace{\sum_{k=1}^{K} U^{-1}\mathbb{E}_{\mathcal{P}}\left[\widehat{\beta}_1^{k,\widehat{\pi}^k}(\mathbf{S}_1)\right] - U^{-1}\mathbb{E}_{\mathcal{P}}\left[\vec{\beta}_1^{\widehat{\pi}^k}(\mathbf{S}_1)\right]}_{\text{Evolution error}} \leq \mathcal{O}\left(\frac{e^{|\gamma|H}-1}{|\gamma|\,H}\cdot H^{5/2}\sqrt{KS^2OA}\cdot\ln\frac{KHSOA}{\delta}\right) \tag{69}$$

Proof. By the Lipschitz continuity of U, we have

$$\underbrace{\sum_{k=1}^{K} U^{-1}\mathbb{E}_{\mathcal{P}}\left[\widehat{\beta}_1^{k,\widehat{\pi}^k}(\mathbf{S}_1)\right] - U^{-1}\mathbb{E}_{\mathcal{P}}\left[\vec{\beta}_1^{\widehat{\pi}^k}(\mathbf{S}_1)\right]}_{\text{Evolution error}} = \sum_{k=1}^{K}\frac{1}{\gamma}\ln\mathbb{E}_{\mathcal{P}}\left[e^{\gamma V_1^{\pi^\star}}\right] - \frac{1}{\gamma}\ln\mathbb{E}_{\mathcal{P}}\left[e^{\gamma V_1^{\widehat{\pi}^k}}\right] \leq \mathsf{K}_\gamma\cdot\sum_{k=1}^{K}\left|\mathbb{E}_{\mathcal{P}}\left[\widehat{\beta}_1^k-\beta_1\right]\right| \tag{70}$$

where $\mathsf{K}_\gamma = \frac{e^{(-\gamma)^+H}}{|\gamma|}$. Next, we find the recurrence relation between $\mathbb{E}[\widehat{\beta}_h - \vec{\beta}_h]$ and $\mathbb{E}[\widehat{\beta}_{h+1} - \vec{\beta}_{h+1}]$.

$$\begin{aligned}
&\left|\mathbb{E}_{\mathcal{P}}\left[\widehat{\beta}^{k,\widehat{\pi}^k}_{h,f_h=(\boldsymbol{\tau}_{h-1},\mathbf{O}_h)}(\mathbf{S}_h) - \beta^{\widehat{\pi}^k}_{h,f_h=(\boldsymbol{\tau}_{h-1},\mathbf{O}_h)}(\mathbf{S}_h)\right]\right| \\
=&\left|\ \mathbb{E}_{\mathcal{P}}\left[u(\mathbf{S}_h,\mathbf{A}_h)\sum_{s_{h+1},o_{h+1}}\widehat{\mathbb{T}}_{h,a_h}(s_{h+1}|s_h)\widehat{\mathbb{O}}_{h+1}(o_{h+1}|s_{h+1})\widehat{\beta}_{h+1,(\boldsymbol{F_h},\boldsymbol{A_h},\boldsymbol{O_{h+1}})}(s_{h+1}) + \mathrm{sgn}(\gamma)\mathsf{b}_h^k(\boldsymbol{S_h},\boldsymbol{A_h})\right.\right. \\
&\left.\left. -\ u(\mathbf{S}_h,\mathbf{A}_h)\sum_{s_{h+1},o_{h+1}}\mathbb{T}_{h,a_h}(s_{h+1}|s_h)\mathbb{O}_{h+1}(o_{h+1}|s_{h+1})\beta_{h+1,(\boldsymbol{F_h},\boldsymbol{A_h},\boldsymbol{O_{h+1}})}(s_{h+1})\right]\ \right| \\
=&\left|\mathbb{E}_{\mathcal{P}}\left[u\widehat{\mathbb{T}}\widehat{\mathbb{O}}\widehat{\beta}_{h+1} - u\mathbb{T}\mathbb{O}\vec{\beta}_{h+1} + \mathrm{sgn}(\gamma)\mathsf{b}_h^k\right]\right| \\
:=&\left|\mathbb{E}_{\mathcal{P}}\left[u\widehat{\mathbb{T}}\widehat{\mathbb{O}}\widehat{\beta}_{h+1} - u\mathbb{T}\mathbb{O}\vec{\beta}_{h+1} + \mathrm{sgn}(\gamma)\mathsf{b}_h^k + u\widehat{\mathbb{T}}\widehat{\mathbb{O}}\beta_{h+1} - u\widehat{\mathbb{T}}\widehat{\mathbb{O}}\beta_{h+1} + u\mathbb{T}\mathbb{O}(\widehat{\beta}_{h+1}-\beta_{h+1}) - u\mathbb{T}\mathbb{O}(\widehat{\beta}_{h+1}-\beta_{h+1})\right]\right| \\
=&\left|\underbrace{\mathrm{sgn}(\gamma)\mathbb{E}_{\mathcal{P}}[\mathsf{b}_h^k]}_{\mathrm{I}} + \underbrace{\mathbb{E}_{\mathcal{P}}[(\widehat{\mathbb{T}}\widehat{\mathbb{O}}-\mathbb{T}\mathbb{O})u\vec{\beta}_{h+1}]}_{\mathrm{II}} + \underbrace{\mathbb{E}_{\mathcal{P}}[(\widehat{\mathbb{T}}\widehat{\mathbb{O}}-\mathbb{T}\mathbb{O})(u\widehat{\beta}_{h+1}-u\beta_{h+1})]}_{\mathrm{III}} + \underbrace{\mathbb{E}_{\mathcal{P}}[\mathbb{T}\mathbb{O}(u\widehat{\beta}_{h+1}-u\beta_{h+1})]}_{\mathrm{IV}}\right|
\end{aligned}$$

By Eq.(61) we observe that $|\mathrm{II}| \leq \mathbb{E}_{\mathcal{P}}[\mathsf{b}_h^k]$. Similarly, Eq.(64) implies that $|\mathrm{III}| \leq 2\mathbb{E}_{\mathcal{P}}[\mathsf{b}_h^k]$. As for IV, we always have $|\mathrm{IV}| \leq \overline{B}_u\cdot\mathbb{E}_{\mathcal{P}}\left[\widehat{\beta}_{h+1}-\vec{\beta}_{h+1}\right]$. Putting things together we conclude

$$\left|\mathbb{E}_{\mathcal{P}}[\widehat{\beta}_h-\vec{\beta}_h]\right| \leq |\mathrm{I}|+|\mathrm{II}|+|\mathrm{III}|+|\mathrm{IV}| \leq 4\cdot\mathbb{E}_{\mathcal{P}}[\mathsf{b}_h^k] + \overline{B}_u\cdot\left|\mathbb{E}_{\mathcal{P}}[\widehat{\beta}_{h+1}-\vec{\beta}_{h+1}]\right|$$

Abbreviating the term $\left|\mathbb{E}_{\mathcal{P}}\left[\widehat{\beta}_h^k-\beta_h^k\right]\right|$ as Δ_h^k, we obtain a recursive equation for the beta vector errors:

$$\begin{cases}\Delta_{H+1}^k = 0\\ \Delta_h^k \leq \overline{B}_u\cdot\Delta_{h+1} + 4\mathbb{E}_{\mathcal{P}}\left[\mathsf{b}_h^k\right]\ ,\ \forall h = H:1\end{cases} \tag{71}$$

Recall that in Eq.(70) we have shown the regret can be controlled by the "initial beta vector errors":

$$\mathrm{Regret}(K;\mathcal{P},\gamma) \leq \mathsf{K}_\gamma\cdot\sum_{k=1}^{K}\mathbb{E}_{\mathcal{P}}\left[\widehat{\beta}_1^k-\beta_1^k\right] = \mathsf{K}_\gamma\sum_{k=1}^{K}\Delta_1^k$$

Using Lemma G.15, we roll down Eq.(71) to solve Δ_1^k, after which we bound the regret by the bonus functions:

$$\begin{aligned}\text{Regret}(K;\mathcal{P},\gamma) \leq & \mathsf{K}_\gamma \sum_{k=1}^{K} \mathbb{E}_{\mathcal{P}}[\widehat{\beta}_1^k - \beta_1^k] \leq 4\mathsf{K}_\gamma \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{P}}[\mathsf{b}_h^k] \prod_{t=1}^{h-1} \overline{B}_u \\ < & 4\mathsf{K}_\gamma \sum_{h=1}^{H} \overline{B}_u^{h-1} \underbrace{\left[\sum_{k=1}^{K} \left(\mathbb{E}_{\mathcal{P}}[\mathsf{b}_h^k(\cdot,\cdot;\gamma)] - \mathsf{b}_h^k(\hat{s}_h^k,\hat{a}_h^k;\gamma)\right)\right.}_{\text{Sample bias}} + \underbrace{\left.\sum_{k=1}^{K} \mathsf{b}_h^k(\hat{s}_h^k,\hat{a}_h^k;\gamma)\right]}_{\text{Bonus sample}}\end{aligned} \tag{72}$$

We invoke Lemma G.9 to bound the terms in the first curly bracket. With probability at least $1-\delta$,

$$\underbrace{\sum_{k=1}^{K} \left(\mathbb{E}_{\mathcal{P}}[\mathsf{b}_h^k(\cdot,\cdot;\gamma)] - \mathsf{b}_h^k(\hat{s}_h^k,\hat{a}_h^k;\gamma)\right)}_{\text{Sample bias}} \leq \left|e^{\gamma(H-h+1)} - 1\right| \sqrt{\frac{K}{2} \cdot \frac{\ln HSA}{\delta}}$$

The terms in the second curly bracket will be controlled by the pigeon hole lemma G.12. Recall that for the empirical data $(\hat{s}_h^k, \hat{a}_h^k)$ collected during the learning process (line 2 in algorithm 2), the bonus function picks the value of

$$\mathsf{b}_h^k(\hat{s}_h^k,\hat{a}_h^k;\gamma) = \left|e^{\gamma(H-h+1)} - 1\right| \cdot \min\left\{1\,,\ \mathsf{t}_h^k(\hat{s}_h^k,\hat{a}_h^k) + \sum_{s_{h+1}} \widehat{\mathbb{T}}_{h,a_h}^k(s_{h+1}|\hat{s}_h^k)\mathsf{o}_{h+1}^k(s_{h+1})\right\} \tag{73}$$

In what follows we calculate the summation over the residue terms. With probability at least $1-2\delta$:

$$\begin{aligned}&\sum_{k=1}^{K} \mathsf{t}_h^k(\hat{s}_h^k,\hat{a}_h^k) + \sum_{s_{h+1}} \widehat{\mathbb{T}}_{h,\hat{a}_h^k}^k(s_{h+1}|\hat{s}_h^k)\mathsf{o}_{h+1}^k(s_{h+1}) \\ =&\sum_{k=1}^{K}\left[\mathsf{t}_h^k(\hat{s}_h^k,\hat{a}_h^k) + \left(\sum_{s_{h+1}} \widehat{\mathbb{T}}_{h,\hat{a}_h^k}^k(s_{h+1}|\hat{s}_h^k)\mathsf{o}_{h+1}^k(s_{h+1}) - \sum_{s_{h+1}} \mathbb{T}_{h,\hat{a}_h^k}(s_{h+1}|\hat{s}_h^k)\mathsf{o}_{h+1}^k(s_{h+1})\right)\right. \\ &\left. + \sum_{s_{h+1}} \mathbb{T}_{h,\hat{a}_h^k}(s_{h+1}|\hat{s}_h^k)\mathsf{o}_{h+1}^k(s_{h+1})\right] \\ \leq&\sum_{k=1}^{K}\left[\mathsf{t}_h^k(\hat{s}_h^k,\hat{a}_h^k) + \underbrace{1\cdot\mathsf{t}_h^k(\hat{s}_h^k,\hat{a}_h^k)}_{\text{transition error}} + \underbrace{\left(\mathbb{E}_{s_{h+1}\sim\mathbb{T}_{h,\hat{a}_h^k}(\cdot|\hat{s}_h^k)}\mathsf{o}_{h+1}^k(s_{h+1}) - \mathsf{o}_{h+1}^k(\hat{s}_{h+1}^k)\right)}_{\text{Sample bias}} + \mathsf{o}_{h+1}^k(\hat{s}_{h+1}^k)\right] \ \text{//Lemma G.10} \\ \leq& 2\left(\sum_{k=1}^{K} \mathsf{t}_h^k(\hat{s}_h^k,\hat{a}_h^k)\right) + \underbrace{2\cdot\sqrt{\frac{K}{2}\cdot\ln HSA/\delta}}_{\text{Sample bias}} + \sum_{k=1}^{K}\mathsf{o}_{h+1}^k(\hat{s}_{h+1}^k) \quad \text{//Lemma G.9}\end{aligned} \tag{74}$$

According to the definition of the residue terms in Eq.(62), the pigeon-hole Lemma G.12 suggests that

$$\begin{aligned}\sum_{k=1}^{k} \mathsf{t}_h^k(\hat{s}_h^k,\hat{a}_h^k) \equiv \sum_{k=1}^{k} \min\left\{1,\ 3\sqrt{\frac{S\cdot H\ln\frac{KHSOA}{\delta}}{\widehat{N}_h^k(\hat{s}_h^k,\hat{a}_h^k)\vee 1}}\right\} \leq \left(3\sqrt{SH}\cdot\iota\right)\cdot 2\sqrt{K\cdot SA} \\ \sum_{k=1}^{K} \mathsf{o}_{h+1}^k(\hat{s}_{h+1}^k) \equiv \sum_{k=1}^{K} \min\left\{1,\ 3\sqrt{\frac{O\cdot H\ln\frac{KHSOA}{\delta}}{\widehat{N}_{h+1}^{k+1}(\hat{s}_{h+1}^k)\vee 1}}\right\} \leq \left(3\sqrt{OH}\cdot\iota\right)\cdot 2\sqrt{K\cdot S}\end{aligned} \tag{75}$$

where $\ln KHSOA/\delta$ is abbreviated as ι. Then with probability at least $1-2\delta$,

$$\begin{aligned}\underbrace{\sum_{k=1}^{K}\mathsf{b}_h^k(\hat{s}_h^k,\hat{a}_h^k;\gamma)}_{\text{Bonus sample}} \le& \left|e^{\gamma(H-h+1)}-1\right|\cdot\left(12\sqrt{K\cdot SA\cdot S\cdot H}+6\sqrt{K\cdot S\cdot O}+\sqrt{K/2}\right)\cdot\sqrt{\ln\left(\frac{KHSOA}{\delta}\right)}\\ <&12\underbrace{\left|e^{\gamma(H-h+1)}-1\right|}_{\text{Risk measure}}\cdot\left(\sqrt{K}\underbrace{\sqrt{SA}\cdot\sqrt{S}}_{\text{Hidden state error}}+\sqrt{K}\underbrace{\sqrt{S}\cdot\sqrt{O}}_{\text{Observation error}}+\underbrace{\sqrt{K}}_{\text{MDS}}\right)\cdot\underbrace{\sqrt{H}\cdot\sqrt{\ln\left(\frac{KHSOA}{\delta}\right)}}_{\text{History-dependency of POMDP}}\end{aligned} \tag{76}$$

Putting things together we can safely state that with probability at least $1-3\delta$,

$$\begin{aligned}&\underbrace{\sum_{k\in[K]}\mathbb{E}_{\mathcal{P}}[\mathsf{b}_h^k(\cdot,\cdot;\gamma)]-\mathsf{b}_h^k(\hat{s}_h^k,\hat{a}_h^k;\gamma)}_{\text{Sample bias}}+\underbrace{\sum_{k=1}^{K}\mathsf{b}_h^k(\hat{s}_h^k,\hat{a}_h^k;\gamma)}_{\text{Bonus samples}}\\ \le& 12\cdot\underbrace{\left|e^{\gamma(H-h+1)}-1\right|}_{\text{Risk measure}}\cdot\left(\underbrace{\sqrt{KS^2A}}_{\text{Hidden state error}}+\underbrace{\sqrt{KSO}}_{\text{Observation error}}+\underbrace{\sqrt{K}}_{\text{MDS}}\right)\cdot\underbrace{\sqrt{H}\cdot\sqrt{\ln\left(\frac{KHSOA}{\delta}\right)}}_{\text{History-dependency of POMDP}}\end{aligned} \tag{77}$$

Bringing Eq.(77)back to Eq.(72), we conclude that with probability at least $1-3\delta$,

$$\begin{aligned}&\underbrace{\sum_{k=1}^{K}U^{-1}\mathbb{E}_{\mathcal{P}}\left[\widehat{\beta}_1^{k,\widehat{\pi}^k}(\mathbf{S}_1)\right]-U^{-1}\mathbb{E}_{\mathcal{P}}\left[\widehat{\beta}_1^{\widehat{\pi}^k}(\mathbf{S}_1)\right]}_{\text{Evolution error}}\\ \le& 48\cdot\underbrace{\frac{e^{|\gamma|H}-1}{|\gamma|\,H}\cdot H}_{\text{Risk and bonus}}\cdot\left(\underbrace{H\sqrt{KS^2A}}_{\text{Hidden state error}}+\underbrace{H\sqrt{KSO}}_{\text{Observation error}}+\underbrace{H\sqrt{K}}_{\text{Sample bias}}\right)\cdot\underbrace{\sqrt{H}\sqrt{\ln\left(\frac{KHSOA}{\delta}\right)}}_{\text{History-dependency of POMDP}}\end{aligned} \tag{78}$$

□

D.4. Result and Discussion

We summarize our previous analysis in Section D.3.3 and D.3.2 into the following theorem, which characterizes the upper bound of the regret given by the algorithm of Beta Vector Value Iteration.

D.4.1. The Main Theorem

Theorem D.8. *(Regret of Beta Vector Value Iteration) Given a POMDP model $\mathcal{P}$, risk-sensitive parameter $\gamma\in\mathbb{R}\setminus\{0\}$ and the number of episodes $K\in\mathbb{Z}_+$, the regret after running algorithm 2 can be controlled by the following upper bound with probability at least $1-4\delta$:*

$$\begin{aligned}\textit{Regret}(K;\mathcal{P},\gamma)\le&48\cdot\underbrace{\frac{e^{|\gamma|H}-1}{|\gamma|}}_{\textit{Risk and bonus}}\cdot\left(\underbrace{\sqrt{KS}}_{\textit{Prior error}}+\underbrace{H\sqrt{KS^2A}}_{\textit{Transition error}}+\underbrace{H\sqrt{KSO}}_{\textit{Emission error}}+\underbrace{H\sqrt{K}}_{\textit{Sample bias}}\right)\cdot\underbrace{\sqrt{H\cdot\ln\left(\frac{KHSOA}{\delta}\right)}}_{\textit{History-dependency of POMDP}}\\ \le&\mathcal{O}\Bigg(\underbrace{\frac{e^{|\gamma|H}-1}{|\gamma|\,H}}_{\textit{Risk-awareness}}\cdot\underbrace{H^2\sqrt{KS^2AO}}_{\textit{Statistical error}}\cdot\underbrace{\sqrt{H}\sqrt{\ln\frac{KHSOA}{\delta}}}_{\textit{History-dependency}}\Bigg)\end{aligned} \tag{79}$$

In what follows, we will provide a technical analysis of the composition of our regret, which is of particular interest to the possible improvements to our algorithm. We will also compare our results with other classical bounds in the related fields of reinforcement learning.

D.4.2. COMPOSITION OF THE REGRET

The factor $\sqrt{K}$ is brought by the pigeon-hole lemma when we sum up several terms across the episodes. Similarly, one factor of H is brought by the summation over the horizon $\sum_{h=1}^{H}$. Another H is brought by the bonus on the beta vector, as is demonstrated in remark D.7. The other $\sqrt{H}$ is incurred by the history-dependency of the POMDP model, as is shown in Eq.(60).

The factor $\sqrt{O}$ and one of the $\sqrt{S}$ is brought by the coverage of an Epsilon net when we try to bound the beta vector function using Lemma G.10. The other $\sqrt{SA}$ and $\sqrt{S}$ come from the pigeon-hole lemma, when we sum the residue terms across $k \in [K]$.

D.4.3. SAMPLE COMPLEXITY

Based on Theorem D.8 we can use the online-to-PAC conversion argument (cf.Appendix G.1.5) to obtain the sample complexity of algorithm 2.

Corollary D.9. *(Sample complexity of Beta Vector Value Iteration) For any* $K \gtrsim \frac{1}{\epsilon^2\delta^2}\left(\frac{e^{|\gamma|H}-1}{|\gamma|H}\right)^2 H^5S^2OA \cdot \ln\left(\frac{KHSOA}{\delta}\right)$*, with probability at least* $1-\delta$*, the uniform mixture of the output policies of algorithm 1 is* ϵ *optimal:*

$$\frac{1}{K}\sum_{k=1}^{K} V_1^{\star} - V_1^{\widehat{\pi}^k} < \epsilon$$

D.4.4. COMPARISON WITH OTHER STUDIES

In this section, we will try to study what the bound of our regret will be in the risk-neutral case and/or the completely observable scenario, so that we can test whether our algorithm is still provably efficient after the degeneration into simpler settings.

Comparison with risk-neutral HOMDP When we take the limit $\gamma \to 0$ to Eq.(79), our regret bound degenerates into

$$\lim_{\gamma\to 0} \text{Regret}(K;\mathcal{P},\gamma) = \tilde{O}(H\sqrt{KSAO}\cdot\sqrt{O^2+H^3S}) \tag{80}$$

which characterizes the performance of our algorithm in the risk-neutral setting.

(Lee et al., 2023) studied risk-neutral POMDP under hindsight observability, whose setting differs from ours only in the risk-sensitivity of the agent. The regret of their algorithm "HOP-B" is provided in theorem C.1 of (Lee et al., 2023), which is

$$\begin{aligned}\sum_{k\in[K]} v\left(\pi^{\star}\right) - v\left(\hat{\pi}_k\right) \lesssim & \underbrace{\sqrt{H^5K\log(2/\delta)}}_{\text{Azuma-Hoeffding}} + \underbrace{\sqrt{OSH^5K\iota}}_{\text{Emission error}} + \underbrace{\sqrt{SAH^4K\iota} + H^4S^2A\iota(1+\log(K))}_{\text{Transition error}} \\ & + \underbrace{H^3S\sqrt{O\iota} + HSA\sqrt{H^3\iota}}_{\text{Residual pigeonhole error}},\end{aligned} \tag{81}$$

We see that our regret improves their result in the order of H, S and A.

The sample complexity the BVVI algorithm also nearly reaches the information-theoretic lower bound for the tabular HOMDPs, which is provided in theorem 5.1 of (Lee et al., 2023):

$$K = \Omega\left(SO/\epsilon^2\right)$$

Comparison with risk-sensitive MDP Due to the significant difference between the formulation of POMDP and MDP, several adjustments to our algorithm are necessary to adapt to a fully observable environment. We will no longer approximate the emission process and drop the terms in the bonus that is relevant with the emission residue. The confidence level will also be increased for O-times. In the end our algorithm will degenerate into the RSVI2 algorithm proposed by (Fei et al., 2021a). After some revision in the proofs, our regret will drop the terms relevant with O, as well as the additional $\sqrt{H}$ brought by the history-dependency of POMDP. The regret will take the form of

$$\text{Regret}(K;\boldsymbol{M},\gamma) = \frac{e^{|\gamma|H}-1}{|\gamma|H}\cdot\tilde{O}(H^2\sqrt{KS^2A}) \tag{82}$$

which matches the result of (Fei et al., 2021a):

$$\text{Regret}(K) \lesssim \frac{e^{|\gamma|H}-1}{|\gamma|H}\sqrt{H^4S^2AK\log^2(HSAK/\delta)}. \tag{83}$$

Consequently, in the MDP case, our regret reaches the lower bound for risk-sensitive RL using the entropic risk(Fei et al., 2020) in terms of K and γH:

$$\text{Regret}(K) \gtrsim \frac{e^{|\gamma|H/2}-1}{|\gamma|\,H}\cdot H^{3/2}\sqrt{K\ln KH} \tag{84}$$

E. Experimental Details

In this section, we elaborate on the details of our experiment presented in Section 9. We also provide additional information on the empirical performance of BVVI.

We designed a POMDP model with two actions ($A = 2$), three states ($S = 3$) and three observations ($O = 3$). The length of horizon H is set to be 4 and the agent interacts with the POMDP for $K = 2,000$ episodes. We chose confidence level $\delta = 0.2$ and set the risk sensitivity parameter γ to 1.0.

For simplicity, in the POMDP, the environment starts deterministically at state 1 and evolves in a time-homogeneous fashion. At each step, the environment transitions from state 1 to state $1, 2$ and 3 with probabilities $0.03, 0.95$, and 0.02, respectively. When starting from state 2, it transitions to states 1, 2, and 3 with probabilities $0.04, 0.02$, and 0.94. The state-transition probabilities become $0.89, 0.10$, and 0.01 when starting from state 3. In each state, the environment emits three possible observations. In state 1, the agents receives observations $1, 2$ or 3 with probabilities $0.83, 0.08$, and 0.09, respectively. The observation distributions become $0.05, 0.79, 0.06$ or $0.02, 0.09, 0.89$ when in states 2 or 3, respectively. The agent receives a reward of 1 when she takes action 1 in state 1, action 2 in state 2, or action 1 in state 3. In other cases, the agent gains a reward of 0.

We train the agent using the BVVI algorithm in the default environment specified above. We then repeat the training procedure in an fully observable environment, which is obtained by replacing the emission kernel of the POMDP with the identity matrix, resulting in Figure 1. We also compare the episodic return of the agent in both POMDP and MDP environments (see Figure 3). Our experiment demonstrates that BVVI converges faster in the MDP environment and that in a POMDP, the agent suffers from higher variance in returns. This can be attributed to the uncertainty of observations confounding the decision-making process.

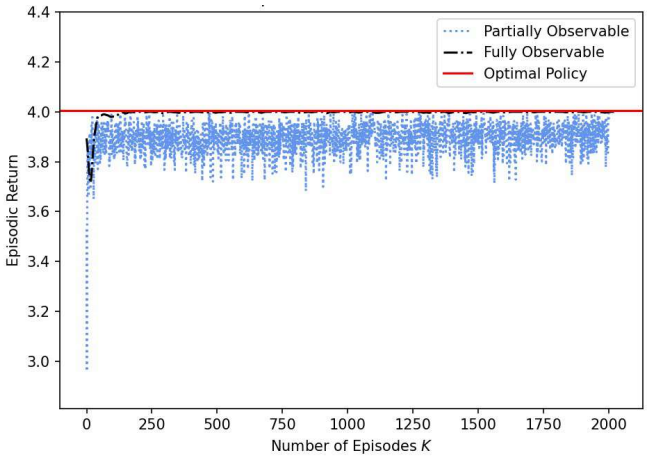

Figure 3: Episodic return $J(\pi^k; \mathcal{P}, \gamma)$ of BVVI(Algorithm 1) in MDP and POMDP.

We also test the performance of BVVI in a POMDP with varying degrees of risk-awareness, as illustrated in Figure 2. Our experiment furthers shows that the agent successfully learns the optimal policy when γ is chosen in $\{-5.0, -3.0, -1.0, 1.0, 3.0, 5.0\}$, consistent with Corollary D.9. The findings are demonstrated in Figure 4.

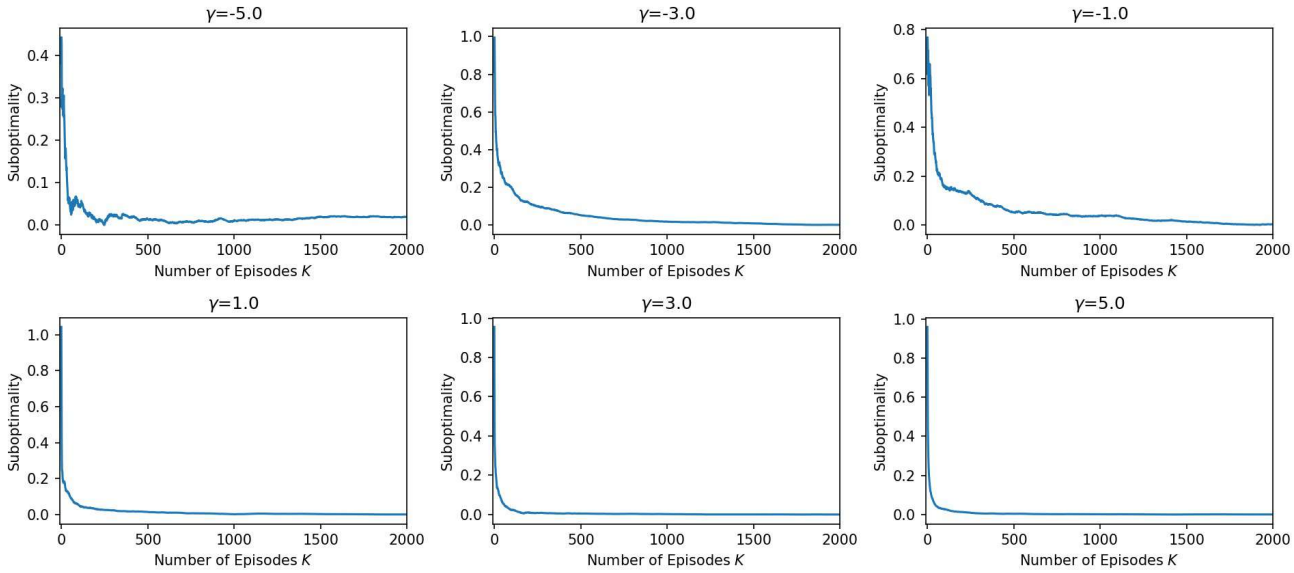

Figure 4: Suboptimality $J(\pi^\star;\mathcal{P},\gamma) - \frac{1}{K}\sum_{k=1}^{K} K(\pi^k;\mathcal{P},\gamma)$ of BVVI(Algorithm 1) with various risk-sensitivity parameters. We pick the risk-level γ from $\{-5.0,-3.0,-1.0,1.0,3.0,5.0\}$.

.

F. More Related Work

Risk-sensitive RL The analytical properties of the risk measures adopted in risk-sensitive RL has been extensively studied in (Boda & Filar, 2006; Sereda et al., 2010; Righi, 2018; Hau et al., 2023). Combined with statistical learning theory, previous works have developed provably efficient algorithms for risk-sensitive RL in both the tabular case (Fei et al., 2020) and the function-approximation setting (Fei et al., 2021b). Entropic risk (Fei et al., 2021a) and CVaR (Du et al., 2022) are among the most popular risk measures adopted in RL.

Intractability of general POMDP Studies (Papadimitriou & Tsitsiklis, 1987; Jin et al., 2020) have shown that seeking the exact solution to the planning or learning problem of general POMDPs is intractable. For this reason, the planning process in algorithm 2, as well as other exact algorithms for POMDP are inefficient in terms of computation complexity. To obtain polynomial sample complexity, recent research follows two directions: they either assume the structure of the POMDP model could leak certain information about the hidden states, or suggest that the training process will offer more knowledge to the agent. We will introduce the two lines of work in what follows.

Learning a POMDP with structural assumptions The first line of research considers sub-classes of POMDP with additional structural properties, such as (Golowich et al., 2022a)(γ-observable POMDPs) and (Liu et al., 2022a)(α-weak-revealing POMDPs). In the tabular case, (Jin et al., 2020) assumed that the emission matrix posses a full column rank. In the continuous setting, (Cai et al., 2022) asks the emission kernel to have a left inverse. Once the assumptions fails to be satisfied, their regret bounds could become vacuous (Liu et al., 2022a). It also remains unclear whether these assumptions are acceptable in the application scenarios.

Learning a POMDP with hindsight observation Another line of research concerning with partially observable RL does not pose structural assumptions on the POMDP model. They consider a friendlier training setting in which the agent could review the sample path of the hidden states at the end of each episode. The new formulation for the POMDP, also referred to as the "Hindsight Observation Markov Decision Process(HOMDP)", is proposed by (Lee et al., 2023) and echoed by (Sinclair et al., 2023; Shi et al., 2023; Guo et al., 2023). The concept of "hindsight observation" is reasonable both theoretically and empirically. Supported by at least six examples in (Lee et al., 2023), reinforcement learning in a partially observable environment offers hindsight information in various application scenarios. Furthermore, they also showed that the lower bound of sample complexity of learning an HOMDP is polynomial for the sizes of the spaces. For these reasons, we follow the second direction of research.

We remind the reader that there is a significant difference between our work and that of (Lee et al., 2023). We propose

the analytical tool of beta vector, which is different from the risk-neutral counterpart alpha vector. We also deploy a change-of-measure technique. We also adopt different analysis of the statistical errors and improve their regret bound in terms of H, S and A.

Risk-sensitive planning with partial information Risk-aware decision-making in partially observable environments has been studied theoretically: (Whittle, 1991) derived an approximate solution to the continuous-time partially-observed risk-sensitive optimal control problem. (Elliott et al., 1996) developed a risk-sensitive Viterbi algorithm for the hidden Markov models. (Baras & James, 1997) also studied similar control-theoretic problems for the finite-state machines. (James et al., 1994; Cavazos-Cadena & Hernández-Hernández, 2005; Bäauerle & Rieder, 2017) considered risk-sensitive planning of a POMDP with the entropy or the utility risk measures. However, these studies did not consider the learning problem of POMDP, not to mention sample complexity.

Our framework is built upon the study of (James et al., 1994). However, we should notice that there fundamental differences between the two works. The study of (James et al., 1994) posed strong assumptions on the POMDP model: they set the initial state as a Gaussian random variable and required the transition law of the states and observations to be i.i.d. Gaussian distributions. We generalize their result to accommodate transition matrices beyond Gaussian distributions. The study of (James et al., 1994) did not consider the learning problem as they assumed the transition matrices are fully known, neither have they carried out a regret analysis which thoroughly discusseed in this work. We also propose new concepts such as the beta vectors and the partially observable risk-sensitive Q functions not considered by the work of (James et al., 1994). We also devise a novel bonus function.

G. Supplementary Materials

G.1. Technical Lemmas

In this section we provide the technical lemmas adopted in this work. We will give the references for existing results and provide a proof for the lemmas developed in this work.

G.1.1. Results from Real and Functional Analysis

Theorem G.1. *(Lebesgue-Radon-Nikodyn theorem, theorem 6.10 in (Rankin, 1968)) Let $\mathscr{M}$ be a $\sigma-$ algebra of a set X. let λ be a measure on $\mathscr{M}$ and μ be a positive $\sigma-$ finite measure on $\mathscr{M}$. There is a unique pair of measures λ_a and λ_s on $\mathscr{M}$ such that $\lambda = \lambda_a + \lambda_s$, where λ is absolutely continuous with respect to μ ($\lambda \mu$) while λ_s and λ_a are concentrated on disjoint sets ($\lambda_s \perp \mu$). Moreover, there is a unique $h \in L^1(\mu)$ such that $\lambda_a(E) = \int_E h d\mu$ for every $E \in \mathscr{M}$. We call h the Radon-Nikodyn derivative of the measure λ_a with respect to μ and we may express the derivative as $h = \frac{d\lambda_a}{d\mu}$.*

Theorem G.2. *(Hilbert-adjoint operator, theorem 3.9-2 and 3.10-2 in (Kreyszig, 1991)) Let H_1 and H_2 be two Hilbert spaces and $T : H_2 \to H_1$ be a bounded linear operator. There exists a unique linear bounded operator $T^\star$ with the same norm of T such that for all $\vec{x} \in H_1$ and $\vec{y} \in H_2$, $\langle \vec{x}, T^\star \vec{y} \rangle = \langle T\vec{x}, \vec{y} \rangle$. If H_1 and H_2 have finite dimensions so that T could be represented by some matrix, then $T^\star$ will be represented by the complex conjugate transpose of that matrix.*

Lemma G.3. *(Dirac function and the expectation)*

$$\int_{\mathcal{X}} dx \mathbb{E}[f(\mathbf{X}, \mathbf{Y})\delta(\mathbf{X} - x)|\mathbf{Z}] = \mathbb{E}[f(\mathbf{X}, \mathbf{Y})|\mathbf{Z}]$$

Proof.

$$\begin{aligned} LHS &= \int_{\mathcal{X}} dx \int_{\mathcal{X}} d\zeta \int_{\mathcal{Y}} d\eta f(\zeta, \eta)\delta(\zeta - x) p_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}(\zeta, \eta|z) = \int_{\mathcal{X}} d\zeta \int_{\mathcal{X}} dx \delta(x - \zeta) \int_{\mathcal{Y}} d\eta f(\zeta, \eta) p_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}(\zeta, \eta|z) \\ &= \int_{\mathcal{X}} d\zeta \int_{\mathcal{Y}} d\eta f(\zeta, \eta) p_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}(\zeta, \eta|z) = RHS \end{aligned} \tag{85}$$

□

Lemma G.4. *(Dirac function, inner product and the expectation)*

$$\langle\, \mathbb{E}[\delta(\mathbf{X} = \cdot\,) F(\mathbf{X}, \mathbf{Y})]\,,\ f(\cdot)\,\rangle = \mathbf{E}[f(\mathbf{X}) \cdot F(\mathbf{X}, \mathbf{Y}))] \tag{86}$$

Proof.

$$\begin{aligned}LHS &= \int_{\mathcal{X}} dx f(x)\mathbf{E}[\delta(\mathbf{X}-x)F(\mathbf{X},\mathbf{Y})]p_{\mathbf{X},\mathbf{Y}}(\zeta,\eta) = \int_{\mathcal{X}} dx f(x)\int_{\mathcal{Y}} d\eta \int_{\mathcal{X}} d\zeta\delta(\zeta-x)\ F(\zeta,\eta)p_{\mathbf{X},\mathbf{Y}}(\zeta,\eta)\\ &= \int_{\mathcal{X}} d\zeta\left[\int_{\mathcal{X}} dx f(x)\delta(x-\zeta)\int_{\mathcal{Y}} d\eta F(\zeta,\eta)\right]p_{\mathbf{X},\mathbf{Y}}(\zeta,\eta) = \int_{\mathcal{X}} d\zeta f(\zeta)\int_{\mathcal{Y}} d\eta F(\zeta,\eta)p_{\mathbf{X},\mathbf{Y}}(\zeta,\eta)\\ &= \mathbf{E}[f(\mathbf{X})\cdot F(\mathbf{X},\mathbf{Y})] = RHS\end{aligned}$$

□

G.1.2. Concentration Inequalities

Lemma G.5. *(Concentration in the ℓ_1 norm, fact 4 of (Liang & Luo, 2023), adapted from (Weissman et al., 2003)) Let P be a probability distribution over a finite discrete measurable space $(\mathcal{X},\Sigma)$. Let $\widehat{P}_n$ be the empirical distribution of P estimated from n samples. Then with probability at least $1-\delta$,*

$$\|\widehat{P}_n - P\|_1 \le \sqrt{\frac{2\,|\mathcal{X}|}{n}\ln\frac{1}{\delta}} \tag{87}$$

Fact G.6. (Naive upper bound) For any bounded function $f(\cdot):\mathcal{X}\to[a,b]$ and two probability measures $\mathbb{P}',\mathbb{P}\in\Delta(\mathcal{X})$, the difference in the expectation can be controlled by the range of the function and the total variance distance between the probability measures.

$$|\mathbb{E}_{X\sim\mathbb{P}'}[f(X)] - \mathbb{E}_{X\sim\mathbb{P}}[f(X)]| \le \frac{(b-a)}{2}\cdot\|\mathbb{P}'(\cdot)-\mathbb{P}(\cdot)\|_1 \tag{88}$$

Proof. By the fact that all probability measures normalize to 1,

$$\begin{aligned}LHS &= \left|\mathbb{E}_{X\sim\mathbb{P}'}\left(f(X)-\frac{b-a}{2}\right) - \mathbb{E}_{X\sim\mathbb{P}}\left(f(X)-\frac{b-a}{2}\right)\right|\\ &= \left|\sum_{x\in\mathcal{X}}\left(f(X)-\frac{b-a}{2}\right)\cdot(\mathbb{P}'(x)-\mathbb{P}(x))\right| \le \sup_{x\in\mathcal{X}}\left|f(x)-\frac{b-a}{2}\right|\cdot\|\mathbb{P}'(x)-\mathbb{P}(x))\|_1 = RHS\end{aligned}$$

□

Remark G.7. The upper bound provided in this lemma is tight for deterministic variable $\boldsymbol{X}$, which will be particularly useful when we study how the regret behaves when the risk-sensitivity parameters tends to zero.

Lemma G.8. *(Hoeffding inequality for random variables, adapted from theorem 2.3 in (Boucheron et al., 2003)) Let $\{\mathbf{Y}_t\}_{t=1}^n$ be a finite set of independent random variables. Suppose that there exists two constant real numbers $\underline{Y}<\overline{Y}$ such that $\underline{Y}\le\mathbf{Y}_t\le\overline{Y}$ holds almost surely for any $\mathbf{Y}_t$, then with probability at least $1-\delta$,*

$$\left|\frac{1}{n}\sum_{i=1}^n Y_i - \frac{1}{n}\sum_{i=1}^n\mathbb{E}[Y_i]\right| \le (\overline{Y}-\underline{Y})\sqrt{\frac{1}{2n}\ln\frac{2}{\delta}} \tag{89}$$

Lemma G.9. *(Azuma-Hoeffding inequality for martingale difference sequences, theorem 2.16 in (Bercu et al., 2015)) Let $\{Y_t\}_{t=1}^\infty$ be a martingale difference sequence with respect to some other stochastic process $\{Xt\}_{t=1}^\infty$. Suppose that there exists two constants $a<b$ such that $a\le Y_t\le b$ almost sure for any $t\in\mathbb{Z}_+$, then for any $n\in\mathbb{Z}_+$ the following relation holds with probability at least $1-\delta$:*

$$\sum_{t=1}^n Y_t < (b-a)\sqrt{\frac{n}{2}\ln\frac{1}{\delta}} \tag{90}$$

Lemma G.10. *(Hoeffding inequality for the function of random variables,extended from lemma 12 of (Bai & Jin, 2020)) Let $\mathbf{X}'$ be a random variable supported on $\mathcal{X}$ that follows an unknown distribution $\mathbb{P}$. Let $f(\cdot)$ be any bounded function that maps $\mathcal{X}$ to $[a,b]$. We draw N i.i.d. samples from $\mathbb{P}$ to construct the empirical distribution $\widehat{\mathbb{P}} := \frac{1}{N}\sum_{i=1}^N\mathbb{1}\{\widehat{x}_i'=x'\}$. Denote Θ as the set of all the parameters that may distinguish the samples. Then with probability at least $1-\delta$,*

$$\left|\mathbb{E}_{\mathbf{X}'\sim\widehat{\mathbb{P}}(\cdot)}[f(\mathbf{X}')] - \mathbb{E}_{\mathbf{X}'\sim\mathbb{P}(\cdot)}[f(\mathbf{X}')]\right| \le (b-a)\cdot\min\left\{1,\ 3\cdot\sqrt{|\mathcal{X}|}\cdot\sqrt{\frac{1}{N}\ln\frac{|\Theta|}{\delta}}\right\} \tag{91}$$

Remark G.11. The factor of $\sqrt{|\mathcal{X}|}$ comes from the epsilon coverage of the range of $f(\cdot)$, since $|\mathcal{X}| = \ln \frac{1}{\epsilon}^{|\mathcal{X}|}$.

Lemma G.12. *(The pigeon-hole lemma, extended from (Azar et al., 2017)) Fix constant h. Suppose that $\{\widehat{z}_h^t\}_{t=1}^K$ are i.i.d. samples drawn from a distribution $\mathbb{P}$ over the finite set $\mathcal{Z}$. For any $k = 0, 1, \cdots, K$, let $N_h^{k+1}(\cdot) : \mathcal{Z} \to [K]$ be defined as the counter function $N_h^{k+1}(z) := \sum_{t=1}^k \mathbb{1}\{\widehat{z}_h^t = z\}$ that records number of occurrences of z within the first k samples. Let $f(\cdot) : \mathcal{Z} \to \mathbb{R}$ be any function that receives integer input. The following relations always hold:*

$$(1)\ \sum_{z\in\mathcal{Z}} N_h^{k+1}(z) = k \quad (2)\ \sum_{k=1}^{K} f(N_h^{k+1}(\widehat{z}_h^k)) = \sum_{z\in\mathcal{Z}} \sum_{i=1}^{N_h^{K+1}(z)} f(i) \quad (3)\ \sum_{k=1}^{K} \frac{1}{\sqrt{\max\{1, N_h^{k+1}(\widehat{z}_h^k)\}}} < 2\sqrt{K\cdot|\mathcal{Z}|} \tag{92}$$

Remark G.13. The third equation in Lemma G.12 shows that the pigeon-hole upper bound depends on the size of the space $\mathbf{Z}$. For MDPs, $\mathcal{Z}$ will be replaced by $\mathcal{S}\times\mathcal{A}$, which is polynomial in the relevant parameters. However in a POMDP, since decision-making depends on the entire history, $\mathcal{Z}$ will be replaced with $\mathscr{O}^h\mathscr{A}^{h-1}$, which causes the regret to be at least of order $O^H A^H$.

Lemma G.14. *(Linearization of Utility Function, Fact 1(a) in Appendix A of (Fei et al., 2020))*

$$\begin{aligned} &\textit{When } \gamma > 0, \quad \textit{for all } 1 < y < x < e^{\gamma H}, \quad \textit{we have } 0 < \frac{1}{\gamma}\ln x - \frac{1}{\gamma}\ln y < \frac{1}{\gamma}(x-y) \\ &\textit{When } \gamma < 0, \quad \textit{for all } e^{\gamma H} < x < y < 1, \quad \textit{we have } 0 < \frac{1}{\gamma}\ln x - \frac{1}{\gamma}\ln y < \frac{e^{|\gamma| H}}{|\gamma|}(y-x) \end{aligned} \tag{93}$$

Lemma G.14 implies that the differences between the entropic risk measures can be bounded by linear functions of the differences between their variables.

G.1.3. Value functions using utility risk

Exponential risks belong to a special class of risk criteria, named the utility function, [14] which fits into the definition of both static and dynamic risk measures. For simplicity, we demonstrate this property in the MDP case when actions are deterministic and the initial state s_1 is fixed. We use $U\circ$ to represent a utility function. For entropic risk, $U = \frac{1}{\gamma}e^{\gamma(\cdot)}$.

(Static risk-measure narration) Assume that the initial state is fixed,

$$\begin{aligned} J(\pi; M) =& U^{-1}\circ\mathbb{E}_M^\pi U\left(\sum_{h=1}^H r_h(\mathbf{S}_h, \mathbf{A}_h)\right) \\ V_{H+1,\text{static}}^\pi =& 0 \quad V_{h,\text{static}}^\pi(s_h) = U^{-1}\mathbb{E}_M^\pi U\left[\sum_{t=h}^H r_t(\mathbf{S}_t, \mathbf{A}_t)\middle| s_h\right] \end{aligned}$$

(Dynamic risk-measure narration) Assume that the initial state is fixed,

$$\begin{aligned} V_{H+1}^\pi =& 0 \\ Q_h^\pi(s_h, a_h) =& r_h(s_h, a_h) + \left(U^{-1}\circ\mathbb{E}_{s_{h+1}\sim P_{h+1}(\cdot|s_h,a_h)}U\right)\left(V_{h+1}^\pi(s_{h+1})\right) \\ V_h^\pi(s_h) =& \mathbb{E}_{a_h\sim\pi_h(\cdot|s_h)}Q_h^\pi(s_h, a_h) \\ J(\pi; M) =& V_1^\pi(s_1) \end{aligned}$$

Next, we show that static and dynamic narrations are equivalent when the policy is deterministic and the initial state is fixed. Precisely speaking, $\forall h \in [H+1], s_h \in \mathscr{S}$:

$$V_h^\pi(s_h) = V_{h,\text{static}}^\pi(s_h)$$

[14] Readers may refer to (Von Neumann & Morgenstern, 1944; Bäuerle & Rieder, 2011; Bäauerle & Rieder, 2017) for details.

Proof. We prove by an induction on h. First, the statement holds obviously at h=H+1. Then

$$\begin{aligned}
V_h^\pi(s_h) =&\mathbb{E}_{a_h}^\pi\left[Q_h^\pi(s_h,a_h)\middle|s_h\right] = \mathbb{E}_{a_h}^\pi\left[U^{-1}\mathbb{E}_{s_{h+1}}\left[U\left[r_h(s_h,a_h)+V_{h+1}^\pi(s_{h+1})\right]\middle|s_h,a_h\right]s_h\right]\\
=&\mathbb{E}_{a_h}^\pi\left[U^{-1}\mathbb{E}_{s_{h+1}}\left[U\left[r_h(s_h,a_h)+U^{-1}\mathbb{E}_M^\pi\left[U\sum_{t=h+1}^H r_t(\mathbf{S}_t,\mathbf{A}_t)\mid s_{h+1}\right]\right]\middle|s_h,a_h\right]s_h\right] \text{ (Induction hypothesis)}\\
=&\mathbb{E}_{a_h}^\pi\left[U^{-1}\mathbb{E}_{s_{h+1}}\left[U\left[U^{-1}\mathbb{E}_M^\pi\left[U\sum_{t=h}^H r_t(\mathbf{S}_t,\mathbf{A}_t)\middle|s_{h+1},s_h,a_h\right]\right]\middle|s_h,a_h\right]s_h\right] \text{ (Markov property)}
\end{aligned}$$

$$\begin{aligned}
=&\mathbb{E}_{a_h}^\pi\left[U^{-1}\mathbb{E}_{s_{h+1}}\left[\mathbb{E}_M^\pi\left[U\sum_{t=h}^H r_t(\mathbf{S}_t,\mathbf{A}_t)\mid s_{h+1},s_h,a_h\right]\middle|s_h,a_h\right]s_h\right]\\
=&\mathbb{E}_{a_h}^\pi\left[U^{-1}\mathbb{E}_M^\pi\left[U\sum_{t=h}^H r_t(\mathbf{S}_t,\mathbf{A}_t)\mid s_h,a_h\right]s_h\right]\\
=&U^{-1}\mathbb{E}_M^\pi\left[U\sum_{t=h}^H r_t(\mathbf{S}_t,\pi_t(\mathbf{S}_t))\middle|s_h\right] = V_{h,\text{static}}^\pi(s_h) \quad \text{(Deterministic policy)}
\end{aligned}$$

□

We stress that the coincidental equivalence between the two narrations arises from the fact that $r_h(s_h,a_h) = U^{-1}\mathbb{E}_M^\pi[U\circ r_h(s_h,a_h)|s_h,a_h]$ and $\mathbb{E}_M^\pi[U\sum_{t\geq h+1} r_t(\mathbf{S}_t,\mathbf{A}_t)|s_{h+1}] = \mathbb{E}_M^\pi[U\sum_{t\geq h+1} r_t(\mathbf{S}_t,\mathbf{A}_t)|s_{h+1}s_h,a_h]$.

G.1.4. The Auto-regressive Equation

Lemma G.15. *The solution to the initial value problem of the equations* $x_N = 0 \quad x_n = A_n x_{n+1} + C_n$ *is* $x_1 = \sum_{\tau=1}^{N-1} A_{1:\tau-1}C_\tau$

We can obtain this result by an induction argument.

Remark G.16. If we restrict $\mathbf{x}_t, A_t \in \mathbb{R}_{\geq 0}$, similarly we can prove that that if $x_N = 0,\ x_n \leq A_n x_{n+1} + C_n$ we have $x_1 \leq \sum_{\tau=1}^{N-1} A_{1:\tau-1}C_\tau$

G.1.5. Online-to-PAC Conversion

The relationship between the regret of an online learning algorithm and its sample complexity is studied by (Cesa-Bianchi et al., 2004) and (Jin et al., 2018). In section 3.1 of (Jin et al., 2018) the authors used Markov's inequality to show that if we choose the output policies $\{\widehat{\pi}^k\}$ of an online learning algorithm uniformly at random, then to ensure these policies are provably approximately correct, i.e.

$$\mathbb{P}\left(\sum_{k=1}^K V^\star - V^{\widehat{\pi}^k} \leq \epsilon\right) \geq 1-\delta$$

one only needs to ensure that the number of episodes K will make the average expected regret lower than $\epsilon\delta$

$$\overline{\text{Regret}}(K) := \frac{1}{K}\sum_{k=1}^K V^\star - V^{\widehat{\pi}^k} \leq \epsilon\delta$$

This technique is frequently used in reinforcement learning, such as in the derivation of corollary 5 in (Liu et al., 2022a) and theorem 6.3 of (Lee et al., 2023) and we have invoked this relation in the derivation of Corollary D.9. For an elementary introduction to the conversion argument please refer to (Tirinzoni et al., 2022).