Provable Risk-Sensitive Distributional Reinforcement Learning with General Function Approximation

Yu Chen [*1] **Xiangcheng Zhang** [*1] **Siwei Wang** [2] **Longbo Huang** [1]

Abstract

In the realm of reinforcement learning (RL), accounting for risk is crucial for making decisions under uncertainty, particularly in applications where safety and reliability are paramount. In this paper, we introduce a general framework on Risk-Sensitive Distributional Reinforcement Learning (RS-DisRL), with static Lipschitz Risk Measures (LRM) and general function approximation. Our framework covers a broad class of risk-sensitive RL, and facilitates analysis of the impact of estimation functions on the effectiveness of RSRL strategies and evaluation of their sample complexity. We design two innovative meta-algorithms: `RS-DisRL-M`, a model-based strategy for model-based function approximation, and `RS-DisRL-V`, a model-free approach for general value function approximation. With our novel estimation techniques via Least Squares Regression (LSR) and Maximum Likelihood Estimation (MLE) in distributional RL with augmented Markov Decision Process (MDP), we derive the first $\widetilde{\mathcal{O}}(\sqrt{K})$ dependency of the regret upper bound for RSRL with static LRM, marking a pioneering contribution towards statistically efficient algorithms in this domain.

1. Introduction

Reinforcement learning (RL) (Sutton & Barto, 2018) has emerged as a powerful framework for sequential decision-making in dynamic and uncertain environments. While traditional RL methods, predominantly focused on maximizing the expected return, have seen significant advancements through approaches such as Q-learning (Mnih et al., 2015; Jin et al., 2018) and policy gradients (Kakade, 2001; Cai et al., 2020), they often fall short in real-world scenarios demanding strict risk control, such as financial investment (Bielecki et al., 2000) , medical treatment (Ernst et al., 2006), and automous driving (Candela et al., 2023).

The significance of comprehending risk management in RL has led to the emergence of Risk-Sensitive RL (RSRL). Unlike risk-neutral RL, which primarily focuses on maximizing expected returns, RSRL seeks to optimize risk metrics, such as entropy risk measures (ERM) (Fei et al., 2020; 2021) or conditional value-at-risk (CVaR) (Wang et al., 2023a), of the possible *cumulative reward* which emphasizes its distributional characteristics. However, traditional RL framework based on Q-learning which typically considers the mean of reward-to-go and corresponding Bellman equation, cannot efficiently capture the characteristics of the cumulative reward's distribution. Therefore, there has been an upsurge of interest in Distributional RL (DisRL) due to its capacity to understand the intrinsic distributional attributes of cumulative rewards, which has already achieved significant empirical success in risk-sensitive tasks (Bellemare et al., 2017; Dabney et al., 2018; Keramati et al., 2020; Urpí et al., 2020; Lim & Malik, 2022).

However, there remains a dearth of comprehensive theoretical insights into the sample complexity of Risk-Sensitive Distributional RL (RS-DisRL), particularly in scenarios encompassing general risk measures and function approximation. Previous theory works of RS-DisRL have primarily been constrained to tabular MDPs (Bastani et al., 2022; Liang & Luo, 2022) which fail in extending to infinite-state settings, or have been confined to specific risk measures such as CVaR or ERM (Stanko & Macek, 2019).

In this paper, we delve into the RS-DisRL with *static Lipschitz risk measures* (LRM), a general risk measure class that includes various well-known risk measures such as coherent risk, convex risk, CVaR, and ERM. In order to address the challenges posed by extremely large or infinite state spaces, we consider two distinct general function approximation scenarios: model-based function approximation and value function approximation. The model-based function approximation, as extensively used in prior works such as Sun et al. (2019); Liu et al. (2022; 2023), typically assumes that the agent has access to a model class that contains

[*]Equal contribution [1]Tsinghua University, Beijing, China [2]Microsoft Research Asia. Correspondence to: Longbo Huang <longbohuang@tsinghua.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

the true transition model. On the other hand, the general value function approximation, as explored in Wang et al. (2020); Jin et al. (2021); Agarwal et al. (2023); Wu et al. (2023); Wang et al. (2023b), offers the agent a distributionally Bellman-complete value function class with the true value distribution.

Under these settings, we introduce general model-based and model-free meta-algorithms, respectively, and employ estimate techniques including Least Squares Regression (LSR) and Maximum Likelihood Estimation (MLE), achieving a sublinear regret upper bound with respect to the number of episodes. Importantly, our work establishes the first statistically efficient framework for RS-DisRL with static Lipschitz risk measures.

Challenges Significant gaps persist in our quest to achieve statistically efficient RS-DisRL with static LRM. (i) The utilization of static LRM in RSRL presents unique complexities. Unlike standard RL, where the focus is on maximizing the expected cumulative reward, RSRL with static LRM involves optimizing the entire distribution of cumulative rewards. This distinct characteristic makes the traditional Q-learning methods, which typically consider the mean of reward-to-go and the corresponding Bellman equation for mean value functions, inadequate for characterizing policy performance and the reward's distribution. (ii) In RSRL with static LRM, the optimal policy is non-Markovian, dependent not only on the current state but also on the rewards received thus far. Thus, it is hard to extend previous works for learning a Markovian policy within polynomial sample complexity.

Technical Contribution To surmount these obstacles, our approach involves several novel technical components. (i) We integrate rigorous distribution analysis techniques from DisRL into the RSRL framework and design novel distributional learning in augmented MDPs, which help us better understand the distributional characteristics of the problem objective. (ii) We pioneer the application of LSR in the exploration process of distributional RL with augmented MDP, incorporating our innovative regression technique tailored for cumulative distribution functions (CDFs) (see Sections 6.1 and 7.1). (iii) Furthermore, we extend traditional MLE methods to DisRL within the augmented MDP framework, supported by a novel connection technique: the augmented simulation lemma (Lemma D.2). In these manners, we present the first statistically efficient algorithms for RSRL with static LRM in this paper.

We summarize Table 1 to present the technical results in this paper, and discuss our detailed contribution as follows:

(i) We provide a general framework for RSRL with static LRM and the general function approximation, which covers a broad class of RSRL studies with various popular static risk measures, such as ERM (Fei et al., 2020; 2021), CVaR (Wang et al., 2023a; Zhao et al., 2023), and spectral risk (Bastani et al., 2022). The framework facilitates analysis of the impact of estimation functions on the effectiveness of RSRL strategies and evaluation of their sample complexity.

Table 1. Summary of results in this paper, where D represent the structural complexity determined by specific methods.

Algorithm	Regret
Model-based Framework `RS-DisRL-M` (Algorithm 1)	
LSR Approach: `M-Est-LSR` (Algorithm 4)	$\widetilde{\mathcal{O}}(L_\infty(\rho)H\,\mathrm{D}_1\sqrt{K})$ (Theorem 6.5)
MLE Approach: `M-Est-MLE` (Algorithm 5)	$\widetilde{\mathcal{O}}(L_\infty(\rho)\mathrm{poly}(H)\,\mathrm{D}_2\sqrt{K})$ (Theorem 6.6)
Model-free Framework `RS-DisRL-V` (Algorithm 2)	
LSR Approach: `V-Est-LSR` (Algorithm 8)	$\widetilde{\mathcal{O}}(L_\infty(\rho)H\,\mathrm{D}_3\sqrt{K})$ (Theorem 7.5)
MLE Approach: `V-Est-MLE` (Algorithm 9)	$\widetilde{\mathcal{O}}(L_\infty(\rho)\mathrm{poly}(H)\,\mathrm{D}_4\sqrt{K})$ (Theorem 7.6)

(ii) For model-based function approximation, we propose a novel meta-algorithm named `RS-DisRL-M` (Algorithm 1), with a general regret upper bound $\mathcal{O}(L_\infty(\rho)\xi(\texttt{M-Est}))$, where $L_\infty(\rho)$ represents the Lipschitz constant of the LRM ρ, and $\xi(\texttt{M-Est})$ is the effectiveness determined by the model-based estimation function. Based on the meta-algorithm, we also obtain the first analysis on the model-based LSR and MLE approaches in distributional RL with augmented MDPs, with effectiveness $\xi = \widetilde{\mathcal{O}}(\mathrm{poly}(H)\,\mathrm{D}\sqrt{K})$, where H is the horizon length, K is the number of episodes, and D is the structural complexity (specified in Theorems 6.5 and 6.6).

(iii) For general value function approximation, we present a new model-free framework `RS-DisRL-V` (Algorithm 2), with general regret upper bound $\mathcal{O}(L_\infty(\rho)\zeta(\texttt{V-Est}))$ where $\zeta(\texttt{V-Est})$ describes the effectiveness of the model-free estimation approach. We also provide novel analysis of LSR and MLE in distributional RL with augmented MDPs and theoretical guarantees of $\zeta = \widetilde{\mathcal{O}}(\mathrm{poly}(H)\,\mathrm{D}\sqrt{K})$, where the dimension term D is specified in Theorem 7.5 for the LSR case and Theorem 7.6 for the MLE case.

2. Related Works

RSRL In RSRL studies, there are two types of risk measures. One is to consider the iterated risk measure, i.e., computing the risk value iteratively. For example, Du et al. (2022); Chen et al. (2023) considers iterated CVaR risk measures, and Liang & Luo (2023) considers iterated LRM risk measures. The other is to consider the static risk measure, i.e., a risk measure of the cumulative reward. For example, Fei et al. (2020; 2021) focus on RSRL with ERM, Wang et al. (2023a); Zhao et al. (2023) investigate the static CVaR risk

measures, and Bastani et al. (2022) studies the static spectral risk measures. In this paper, we consider the static LRM, a static risk measure that encompasses the ERM, static CVaR and spectral risk, and give the theoretical learning analysis.

DisRL Many previous works (Rowland et al., 2018; 2023; Bellemare et al., 2017) develop asymptotic covergence analysis for DisRL. With MLE approaches, Wu et al. (2023) discusses the statistical complexity bounds for off-policy DisRL and Wang et al. (2023b) considers the small-loss bounds for DisRL. Bastani et al. (2022) is the first to give the sample complexity bounds for RSRL, while it only studies the static spectral risk measure within the tabular MDPs. Compared to these results, our work focuses on a more general risk-sensitive target (LRM) and enables to use the general function approximation.

3. Notations

For a positive integer N, we let $[N] = \{1, 2, \cdots, N\}$. Denote $\int_s f(s) := \int_S f(s)ds$ if we integrate f over the universal set of $s \in S$. For a function $f : \mathcal{X} \to \mathbb{R}$, we define the ℓ_1-norm $\|f\|_1 = \int_{\mathcal{X}} |f(x)|dx$ and ℓ_∞-norm $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. Denote $\Delta(\mathcal{X})$ as the distribution over space $\mathcal{X}$. We use the standard $\mathcal{O}(\cdot)$ to hide universal constant factors, and $\widetilde{\mathcal{O}}(\cdot)$ to further suppress the polylogarithmic factors in $\mathcal{O}(\cdot)$.

4. Problem Formulation

Episodic Markov Decision Process In this study, we examine an episodic Markov Decision Process (MDP) denoted as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, K, H, \{\mathbb{P}_h\}_{h=1}^H, \{\mathbb{R}_h\}_{h=1}^H)$ characterized by state space $\mathcal{S}$, action space $\mathcal{A}$, finite episode number K, finite time horizon length H, transition probabilities $\mathbb{P}_h(\cdot|s, a) \in \Delta(\mathcal{S})$ and *distributional* reward $\mathbb{R}_h(\cdot|s, a) \in \Delta([0, 1])$[1] for step $h \in [H]$. At the outset of each episode $k \in [K]$, we start with an initial state $s_{k,1}$ chosen by the MDP. At each step $h \in [H]$, the agent receives a historical record $\mathcal{H}_{k,h} = \{s_{k,1}, a_{k,1}, r_{k,1}, \cdots, s_{k,h}\}$ and select an action $a_{k,h} \sim \pi_h^k(\cdot|\mathcal{H}_{k,h})$ by a *history-dependent* policy $\boldsymbol{\pi}_h^k = \{\pi_h^k\}_{h=1}^H$ Then, the MDP will return a reward $r_{k,h} \sim \mathbb{R}_h(\cdot|s_{k,h}, a_{k,h})$ and transfer into next state $s_{k,h+1} \sim \mathbb{P}_h(\cdot|s_{k,h}, a_{k,h})$. This episode will end in step $H + 1$. Throughout this paper, we assume that the agent lacks knowledge of the transition probabilities. For a fixed history-dependent policy $\boldsymbol{\pi}$, the cumulative reward for an episode played with policy $\boldsymbol{\pi}$ is a bounded real-valued random variable $Z^{\boldsymbol{\pi}} = \sum_{h=1}^H r_h$, where $r_h \sim \mathbb{R}_h(\cdot|s_h, a_h)$.

Lipschitz Risk Measures *Lipschitz Risk Measures* (LRM) are quantified by a function $\rho : \mathcal{Z} \to \mathbb{R}$ mapping random variables $Z \in \mathcal{Z}$ to real numbers, distinguished by two critical properties. **C1. Law invariance:** If $Z, W \in \mathcal{Z}$ have the same distribution functions, $F_Z = F_W$, then $\rho(Z) = \rho(W)$. **C2. Lipschitz continuity:** Consider the supremum norm $\|\cdot\|_\infty$ over the set of all distribution functions of the random variable class $\mathcal{Z}$. There exists a Lipschitz constant $L_\infty(\rho)$ such that $|\rho(Z) - \rho(W)| \leq L_\infty(\rho)\|F_Z - F_W\|_\infty$ holds for any $Z, W \in \mathcal{Z}$.

The law invariance condition, foundational in risk measure studies (Kusuoka, 2001; Frittelli & Gianin, 2005; Liang & Luo, 2023), indicates that identical distribution functions result in equal risk measures. LRM encompass a broad spectrum of general risk assessments, including coherent risk (Artzner et al., 1999), monetary risk (Jia et al., 2020), and convex risk measures (Föllmer & Knispel, 2012), highlighting the versatility of LRM. Popular metrics like Entropy Risk Measures (ERM) and Conditional Value-at-Risk (CVaR) also align with these conditions, with Lipschitz constants $L_\infty(\mathrm{ERM}_\beta) = \frac{\exp(|\gamma|H)-1}{|\gamma|}$ and $L_\infty(\mathrm{CVaR}_\alpha) = \frac{H}{a}$ (Liang & Luo, 2023).

RSRL with Static LRM In this paper, we delve into the **R**isk-**S**ensitive **R**einforcement **L**earning (RSRL) with static **L**ipschitz **R**isk **M**easures (LRM), focusing on optimizing risk-sensitive rewards via *history-dependent* policies. The objective is to find an optimal policy $\boldsymbol{\pi}^*$ that maximizes the LRM-defined cumulative reward $\boldsymbol{\pi}^* = \mathrm{argmax}_{\boldsymbol{\pi}} \rho(Z^{\boldsymbol{\pi}})$, then we define our regret as the difference between the cumulative rewards of the optimal policy and the policy deployed at each episode: $\mathrm{Regret}(K) = \sum_{k=1}^K \rho(Z^{\boldsymbol{\pi}^*}) - \rho(Z^{\boldsymbol{\pi}^k})$.

5. Augmented MDPs and Distributional Bellman Equation

The key of our analysis revolves around the distributional Bellman equation applied to augmented MDPs. Recognizing that learning an optimal history-dependent policy π^* can be prohibitively sample-intensive, previous works (Bäuerle & Ott, 2011; Bastani et al., 2022) leverage the concept of augmented MDP in risk-sensitive conditions, where any history-dependent policy $\tilde{\boldsymbol{\pi}}$ corresponds to a *Markov* policy $\boldsymbol{\pi}^\dagger$ in the augmented MDP. This equivalence allows for facilitating effective risk-sensitive policy learning without sacrificing computational tractability.

Augmented MDPs We embrace the notion of augmented MDPs, originally introduced by Bäuerle & Ott (2011) and widely used in RSRL with static risk measures (Bastani et al., 2022; Wang et al., 2023a; Zhao et al., 2023). In the context of augmented MDPs, the state space is expanded to $\mathcal{S}^\dagger = \{s_h^\dagger = (s_h, y_h)\}$, where $y_h = \sum_{h=1}^H r_h$ represents the cumulative reward accumulated up to time step h.We denote the augmented MDP as $\mathcal{M}^\dagger = (\mathcal{S}^\dagger, \mathcal{A}, K, H, \{\mathbb{P}_h\}_{h=1}^H, \{\mathbb{R}_h\}_{h=1}^H)$. To capture the

[1]Without loss of generality, we assume the reward $r \in [0, 1]$ for each step. Additionally, we assume that the agent has knowledge of the reward distribution, a common assumption in prior research (Liu et al., 2022; 2023; Wang et al., 2023b).

augmented dynamics, we introduce the augmented transition operator as follows:

$$\mathbb{T}_h(s_{h+1}^\dagger|s_h^\dagger, a_h) := \mathbb{P}_h(s_{h+1}|s_h, a_h)\mathbb{R}_h(y_{h+1}-y_h|s_h, a_h)$$

Let $\Pi^\dagger$ denote the set of the Markov policies within the augmented MDP $\mathcal{M}^\dagger$. Theorem 3.1 in Bastani et al. (2022) shows that for any history-dependent policy $\tilde{\pi}$ in original MDP, there exists a Markov policy in the augmented MDP $\boldsymbol{\pi} \in \Pi^\dagger$, such that $F_{Z^{\boldsymbol{\pi}}} = F_{Z^{\tilde{\boldsymbol{\pi}}}}$. This result underscores the equivalence between the distribution of cumulative rewards under a history-dependent policy in original MDP and a corresponding Markov policy in the augmented MDP.

Distributional Bellman Equation Similar to prior works such as Bellemare et al. (2017); Bastani et al. (2022); Wang et al. (2023b); Wu et al. (2023), we integrate the distributional Bellman equation within an augmented MDP framework. For any policy $\boldsymbol{\pi} \in \Pi^\dagger$ and $h \in [H]$, we denote the random variable $Z_h^{\boldsymbol{\pi}}(s_h^\dagger, a_h)$ as the cumulative return $\sum_{h'=h}^H r_{h'}$ after taking action a_h in state $s_h^\dagger$ via policy $\boldsymbol{\pi} \in \Pi^\dagger$.

Definition 5.1 (Distributional Bellman Equation (Wang et al., 2023b; Bastani et al., 2022))**.** Let $F_h^{\boldsymbol{\pi}}(\cdot|s_h^\dagger, a_h)$ be the cumulative distribution function (CDF) of random variable $Z_h^{\boldsymbol{\pi}}(s_h^\dagger, a_h)$, and let $f_h^{\boldsymbol{\pi}}(\cdot|s_h^\dagger, a_h)$ be its probability density function (PDF). We define the augmented distributional Bellman operator $\mathcal{T}_{h,\boldsymbol{\pi}}^\dagger : \Delta(\mathbb{R}) \to \Delta(\mathbb{R})$ as:

$$\mathcal{T}_{h,\boldsymbol{\pi}}^\dagger f_{h+1}^{\boldsymbol{\pi}}(x|s_h^\dagger, a_h) := \int_{s_{h+1}^\dagger, a_{h+1}} \mathbb{T}_h(s_{h+1}^\dagger|s_h^\dagger, a_h) \cdot \pi_{h+1}(a_{h+1}|s_{h+1}^\dagger) f_{h+1}^{\boldsymbol{\pi}}(x-(y_{h+1}-y_h)|s_{h+1}^\dagger, a_{h+1}),$$

$$\mathcal{T}_{h,\boldsymbol{\pi}}^\dagger F_{h+1}^{\boldsymbol{\pi}}(x|s_h^\dagger, a_h) := \int_{s_{h+1}^\dagger, a_{h+1}} \mathbb{T}_h(s_{h+1}^\dagger|s_h^\dagger, a_h) \cdot \pi_{h+1}(a_{h+1}|s_{h+1}^\dagger) F_{h+1}^{\boldsymbol{\pi}}(x-(y_{h+1}-y_h)|s_{h+1}^\dagger, a_{h+1}).$$

By the definition of $Z_h^{\boldsymbol{\pi}}$ and $Z_{h+1}^{\boldsymbol{\pi}}$, we have $f_h^{\boldsymbol{\pi}} = \mathcal{T}_{h,\boldsymbol{\pi}}^\dagger f_{h+1}^{\boldsymbol{\pi}}$ and $F_h^{\boldsymbol{\pi}} = \mathcal{T}_{h,\boldsymbol{\pi}}^\dagger F_{h+1}^{\boldsymbol{\pi}}$. Generally, we can write the distributional Bellman equation in random variable version as $Z_h^{\boldsymbol{\pi}} = \mathcal{T}_{h,\boldsymbol{\pi}}^\dagger Z_{h+1}^{\boldsymbol{\pi}}$.

6. Model-Based Meta-Algorithm `RS-DisRL-M`

This section introduces the meta-algorithm `RS-DisRL-M` for model-based function approximation in RS-DisRL, alongside its theoretical underpinnings. It further delves into two pivotal estimation techniques: LSR and MLE, formulating statistically efficient algorithms for RSRL with LRM.

We describe the model-based function approximation framework utilized for our analysis, drawing on the methodologies previously explored by Fei et al. (2021); Liu et al. (2022; 2023); Chen et al. (2023). Specifically, we model each transition probability $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ using a parametric form $\mathbb{P}_\theta$, with the true transition model for each decision epoch h represented by θ_h^*. The set of true models across all epochs is denoted by $\boldsymbol{\theta}^* = \{\theta_h^*\}_{h=1}^H$. We then establish a standard realizability assumption for the model-based function approximation, ensuring that our model accurately reflects the dynamics of the environment

Assumption 6.1 (Model-based realizability (Fei et al., 2021; Liu et al., 2022; 2023; Chen et al., 2023))**.** The agent is given a model class $\boldsymbol{\Theta} = \{\Theta_h : h \in [H]\}$ which specifies a class of transition probabilities $\{\mathbb{P}_{\theta_h} : \theta_h \in \Theta_h, h \in [H]\}$ and satisfies $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$.[2]

Algorithm 1 `RS-DisRL-M`

1: **Input:** Model class $\boldsymbol{\Theta}$, confidence radius β.
2: **Initialize:** $\widehat{\boldsymbol{\Theta}}_1 \leftarrow \boldsymbol{\Theta}$.
3: **for** $k = 1$ **to** K **do**
4: $(\boldsymbol{\pi}^k, \widehat{\boldsymbol{\theta}}_k) = \operatorname{argmax}_{\boldsymbol{\pi}\in\boldsymbol{\Pi}^\dagger, \boldsymbol{\theta}\in\widehat{\boldsymbol{\Theta}}_k} \rho(Z_{\boldsymbol{\theta}}^{\boldsymbol{\pi}})$. //Optimistic planning
5: Execute policy $\boldsymbol{\pi}^k$, add the collected data $\boldsymbol{\tau}_k = \{(s_{k,h}, a_{k,h}, r_{k,h})\}_{h=1}^H$ and $\boldsymbol{\pi}^k$, $\widehat{\boldsymbol{\theta}}_k$ into history $\mathcal{H}_k = \mathcal{H}_{k-1} \cup \{(\boldsymbol{\tau}_k, \boldsymbol{\pi}^k, \widehat{\boldsymbol{\theta}}_k)\}$. //Data collection
6: $\widehat{\boldsymbol{\Theta}}_{k+1} = \texttt{M-Est}\,(\boldsymbol{\Theta}, \mathcal{H}_k, \beta)$. //Confidence set construction
7: **end for**

We introduce the meta framework `RS-DisRL-M` (Algorithm 1) for **R**isk-**S**ensitive **Dis**tributional **RL** with **M**odel-**B**ased function approximation. `RS-DisRL-M` is a model-based algorithm which takes a model class $\boldsymbol{\Theta}$ as an input, and operates in three main steps.

(i) Optimistic planning (Line 4): the algorithm computes the optimistic model $\widehat{\boldsymbol{\theta}}_k$ and corresponding augmented policy $\boldsymbol{\pi}^k$ in the model confidence set $\widehat{\boldsymbol{\Theta}}_k$. Here the random variable $Z_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}$ denotes the cumulative reward collected with policy $\boldsymbol{\pi}$ in augmented MDPs modeled by $\boldsymbol{\theta}$. **(ii) Data collection** (Line 5): the algorithm executes the optimal policy $\boldsymbol{\pi}^k$ planned from step (i) and collects the trajectory $\boldsymbol{\tau}_k$. **(iii) Confidence set construction** (Line 6): the algorithm estimates the models and constructs the new confidence set for the next episode based on a **M**odel-based **Est**imation function (`M-Est`), model class $\boldsymbol{\Theta}$, and confidence radius β. The estimation function `M-Est` can be designed by various estimation methods, such as LSR or MLE, depending on the specific structure of the MDP or the model class.

This framework encapsulates the essence of leveraging model-based approaches for efficient learning and adaptation in RSRL, aligning with strategies explored in recent literature (Bastani et al., 2022; Liu et al., 2023; Chen et al., 2023).

[2]We assume this structure of $\boldsymbol{\Theta}$ to simplify the notations in analysis. In fact, our analysis works as long as $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$.

Theoretical Guarantees The theoretical guarantees for the RS-DisRL-M algorithm are anchored on two critical conditions related to the estimation function M-Est.

Condition 6.2 (Concentration)**.** For $\delta \in (0,1]$, with probability at least $1-\delta$, $\boldsymbol{\theta}^* \in \widehat{\boldsymbol{\Theta}}_k$ holds for every $k \in [K]$.

The concentration condition, common in theoretical RL analysis (Agarwal et al., 2019; Jin et al., 2018; Ayoub et al., 2020), ensures that the true transition model is consistently included within the algorithm's confidence set throughout the learning process.

Condition 6.3 (General elliptical potential)**.** For $\delta \in (0,1]$, with probability at least $1-\delta$, the supremum distance between the CDF of the chosen estimated model $\widehat{\boldsymbol{\theta}}^k$ and real model $\boldsymbol{\theta}^*$ under policy $\boldsymbol{\pi}^k$ can be bounded by $\sum_{k=1}^K \left\| F_{Z^{\boldsymbol{\pi}^k}_{\widehat{\boldsymbol{\theta}}_k}} - F_{Z^{\boldsymbol{\pi}^k}_{\boldsymbol{\theta}^*}} \right\|_\infty \leq \xi(K,H,\boldsymbol{\Theta},\beta,\delta)$, where $\xi(K,H,\boldsymbol{\Theta},\beta,\delta)$ is the complexity bound determined by the estimation function M-Est.

Intuitively, Condition 6.3 bounds the estimation error by controlling the discrepancy between the CDF of the estimated and real models under the chosen policy. This condition resembles the pigeonhole principle for tabular MDPs (Jin et al., 2018) and the elliptical potential lemma for linear and linear mixture MDPs (Jin et al., 2020; Zhou et al., 2021). However, Condition 6.3 demands to bound the supremum difference of the CDF during the learning process, which is natrually different from the previous.

Conditions 6.2 and 6.3 together establish the reliability and effectiveness of the estimation function in the RS-DisRL-M framework. Adherence to these conditions signifies that the estimation function can facilitate efficient learning in RS-DisRL-M.

Theorem 6.4. *Under Assumption 6.1, if the estimation function M-Est satisfies Conditions 6.2 and 6.3, then the regret of RS-DisRL-M (Algorithm 1) can be bounded by* $\mathrm{Regret}(K) \leq L_\infty(\rho)\xi(K,H,\boldsymbol{\Theta},\beta,\delta)$.

This theorem integrates and extends results from previous research on RSRL with static risk measures, offering a comprehensive view that includes notable theorems from Bastani et al. (2022); Wang et al. (2023a); Fei et al. (2021). The primary challenge lies in satisfying the concentration and elliptical potential conditions for the estimation function M-Est and managing the complexity bound $\xi(K,H,\boldsymbol{\Theta},\beta,\delta)$.

Below, we introduce LSR (M-Est-LSR, Algorithm 4) and MLE (M-Est-MLE, Algorithm 5) based estimation functions. These functions meet the necessary conditions and demonstrate an effective bound $\xi(K,H,\boldsymbol{\Theta},\beta,\delta) = \widetilde{\mathcal{O}}(\mathrm{poly}(H)\,\mathrm{D}\sqrt{\beta K/\delta})$ with dimension term D specified in Theorem 6.5 for LSR and Theorem 6.6 for MLE, giving the $\sqrt{K}$ dependency for meta-algorithm RS-DisRL-M and achieving minimax-optimal in terms of K in tabular MDPs for ERM (Fei et al., 2020) and CVaR (Wang et al., 2023a).

6.1. Estimation by Model-Based LSR Approach

Least Squares Regression (LSR), a well-established estimation technique, has been effectively utilized in linear and linear mixture MDPs (Jin et al., 2020; Zhou et al., 2021). Its common application involves regression on combinations of the transition model with bounded functions. In risk-neutral scenarios, it's often paired with mean value functions (Jin et al., 2020), while in risk-sensitive settings, utility functions are preferred (Chen et al., 2023; Xu et al., 2023). This section explores a novel approach in distributional RL by combining the transition model with mixed distribution functions based on the distributional Bellman equation for CDFs. The newly proposed M-Est-LSR algorithm represents a statistically efficient LSR method tailored for this context.

The intuition of LSR is to approximate the Bellman operator with the empirical sample. Different from previous stuides (Jin et al., 2020; Fei et al., 2021; Chen et al., 2023), we have to analyze the augmented distributional Bellman equation for transition model $\boldsymbol{\theta}$. Denote $F_h^{\boldsymbol{\pi},\boldsymbol{\theta}}(x|s_h^\dagger) := \int_{a_h} \pi_h(a_h|s_h^\dagger) F_h^{\boldsymbol{\pi},\boldsymbol{\theta}}(x|s_h^\dagger,a_h)$, we have the following Bellman equation

$$F_h^{\boldsymbol{\pi},\boldsymbol{\theta}}(x|s_h^\dagger) = \int_{(a_h,s_{h+1}^\dagger)} \pi_h(a_h|s_h^\dagger)\mathbb{T}_{\theta_h}(s_{h+1}^\dagger|s_h^\dagger,a_h) \cdot F_{h+1}^{\boldsymbol{\pi},\boldsymbol{\theta}}\left(x-(y_{h+1}-y_h)\middle|s_{h+1}^\dagger\right).$$

Notice that the transition $\mathbb{T}_{\theta_h}$ for augmented MDP compresses the real transition $\mathbb{P}_{\theta_h}$ in original MDP and the reward distribution $\mathbb{R}_h$. However, the only empirical observation available to the agent is the transfer sample $(s_{k,h},a_{k,h}) \to s_{k,h+1}$ and the reward sample $r_{k,h} \sim \mathbb{R}_h(s_{k,h},a_{k,h})$. Therefore, we have to decompose the augmented Bellman operator for estimating the transition models of original MDP. To do so, we design a mixed distribution function

$$\widehat{F}_{k,h+1}(s) := \int_{y_{k,h}}^1 \mathbb{R}_h(r|s_{k,h},a_{k,h}) F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x_{k,h}-r|s,y_{k,h}+r)dr\,,$$

where $\boldsymbol{\pi}^k$ and $\widehat{\boldsymbol{\theta}}_k$ are given by the optimistic planning (Line 4 in Algorithm 1) based on the information before episode k. Here $x_{k,h}$ is defined as Eq.(2), which maximizes the diameter of $\widehat{\Theta}_{k,h}$ with mixed CDFs, intuitively contributing to the exploration direction by maximizing the uncertainty. For simplicity, we denote the combination form $[\mathbb{P}_{\theta_h}\widehat{F}_{k,h+1}](s_h,a_h) := \int_{s_{h+1}} \mathbb{P}_{\theta_h}(s_{h+1}|s_h,a_h)\widehat{F}_{k,h+1}(s_{h+1})$, and the combination set $\mathcal{W}_{\boldsymbol{\Theta}} := \{[\mathbb{P}_{\theta_h}F] : \mathcal{S}\times\mathcal{A} \to [0,1] : \boldsymbol{\theta} \in \boldsymbol{\Theta}, F : \mathcal{S} \to [0,1]\}$.

We are now ready to present the procedure of the estimation function M-Est-LSR. Due to space limitations, we defer

the formal pseudocode to the appendix (see Algorithm 4). First, M-Est-LSR estimate a model $\theta_{k,h}^{\mathrm{LSR}}$ for step h at episode k by LSR based on the information before episode $k-1$, which serves as the center of the confidence set:

$$\theta_{k,h}^{\mathrm{LSR}} \leftarrow \operatorname*{argmin}_{\theta_h \in \Theta_h} \sum_{i=1}^{k-1} \left(\widehat{F}_{i,h+1}(s_{i,h+1}) - [\mathbb{P}_{\theta_h} \widehat{F}_{i,h+1}](s_{i,h}, a_{i,h}) \right)^2 .$$

Then, the confidence set for each step $h \in [H]$ is constructed by previous history

$$\widehat{\Theta}_{k,h} := \left\{ \theta_h \in \Theta_h : \mathrm{Dist}_{\mathcal{H}_{k-1},h}^{\mathrm{LSR}}(\theta_h || \theta_{k,h}^{\mathrm{LSR}}) \leq \beta^{\mathrm{LSR}} \right\} .$$

where the Euclidean-type distance function is defined as

$$\mathrm{Dist}_{\mathcal{H}_k,h}^{\mathrm{LSR}}(\theta^1 || \theta^2) := \sum_{i=1}^{k} \left([\mathbb{P}_{\theta^1} \widehat{F}_{i,h+1}](s_{i,h}, a_{i,h}) - [\mathbb{P}_{\theta^2} \widehat{F}_{i,h+1}](s_{i,h}, a_{i,h}) \right)^2$$

Finally, the function returns the confidence set $\widehat{\boldsymbol{\Theta}}_k = \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} : \theta_h \in \widehat{\Theta}_{k,h}, h \in [H] \right\}$.

Theorem 6.5. *Let the confidence radius for LSR* $\beta^{\mathrm{LSR}} := 8 \log(2H^2 \mathcal{N}_C(\boldsymbol{\Theta}, 1/K, \|\cdot\|_1)/\delta) + 4\sqrt{\log(4HK^2/\delta)}$. *The estimation function* M-Est-LSR *satisfies Conditions 6.2 and 6.3 with* $\xi^{\mathrm{LSR}}(K, H, \boldsymbol{\Theta}, \beta^{\mathrm{LSR}}, \delta) = \widetilde{\mathcal{O}}\left(H \sqrt{K \dim_E(\mathcal{W}_{\boldsymbol{\Theta}}, \frac{1}{\sqrt{K}}) \log\left(\mathcal{N}_C(\boldsymbol{\Theta}, \frac{1}{K}, \|\cdot\|_1)/\delta\right)} \right)$, *where* $\dim_E$ *represents the eluder dimension, and* $\mathcal{N}_C$ *denotes the covering number.*

The formal proof is detailed in Appendix C. By applying Theorem 6.5, the meta-algorithm RS-DisRL-M equipped with the M-Est-LSR estimation function achieves a sublinear regret upper bound of $\widetilde{\mathcal{O}}(L_\infty(\rho) H \, \mathrm{D} \sqrt{K})$, where $\mathrm{D} = \sqrt{\dim_E(\mathcal{W}_{\boldsymbol{\Theta}}, 1/\sqrt{K}) \log\left(\mathcal{N}_C(\boldsymbol{\Theta}, 1/K, \|\cdot\|_1)/\delta\right)}$ represents the structural dimension. Here $\dim_E(\mathcal{W}_{\boldsymbol{\Theta}}, \sqrt{K})$ characterizes the eluder dimension of the combination set $\mathcal{W}_{\boldsymbol{\Theta}}$, and $\mathcal{N}_C(\boldsymbol{\Theta}, 1/K, \|\cdot\|_1)$ denotes the covering number of the model class, both of which are commonly employed in the analysis of model-based LSR (Ayoub et al., 2020; Fei et al., 2021; Chen et al., 2023).

Compared to previous works studying RSRL with ERM, the result of Theorem 6.5 improves upon the findings of Fei et al. (2020; 2021). Notably, our regret bound does not include the additional $e^{|\gamma| H^2}$ term in the Lipschitz constant. This improvement is attributed to the distributional analysis, which avoids the $e^{|\gamma| H}$ factor while back propagating the Bellman error in Lemma 3 of Fei et al. (2020). Furthermore, when transitioning to the risk-neutral setting, the result established in Theorem 6.5 aligns with the regret bound presented by Ayoub et al. (2020) up to H factors.

6.2. Estimation by Model-Based MLE Approach

We develop the MLE methods to give a refined estimation for the transition models in augmented MDPs. Our method is inspired by the generic model-based MLE method OMLE in Liu et al. (2023).

Similarly to Algorithm 1 in Liu et al. (2023), we construct the estimation algorithm M-Est-MLE, where we provide the detailed pseudocode in Algorithm 5 for space limitation. Employing a standard MLE analysis (Geer, 2000), we effectively bound the total squared total variation (TV) distance between our estimated model and the true model by: $\sum_{i=1}^{k-1} \sum_{h=1}^{H} \mathbb{E}_{\nu_{\boldsymbol{\theta}^*,h}^{\boldsymbol{\pi}^i}} \left[\left\| \left(\mathbb{P}_{\hat{\theta}_{k,h}} - \mathbb{P}_{\theta_h^*} \right)(s_h, a_h) \right\|_1^2 \right] \leq \mathcal{O}(\beta^{\mathrm{MLE}})$, where $\nu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}^i}$ denotes the visitation measure for $\boldsymbol{\pi}^i$ under the real transition kernel.

However, the standard simulation lemma (e.g., Lemma 10 in Sun et al. (2019)) fails in analyzing the efficiency of M-Est-MLE since the policy learned in our meta-algorithm is non-Markovian for standard episodic MDPs. Instead, we proposed a novel augmented simulation lemma (see Lemma D.2) connecting the Total Variation (TV) distance between model difference with the ℓ_∞ distance between the CDFs of the cumulative return random variable: $\left\| F_{Z_{\hat{\boldsymbol{\theta}}_k}^{\boldsymbol{\pi}^k}} - F_{Z_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}^k}} \right\|_\infty \leq \sum_{h=1}^{H} \mathbb{E}_{\nu_{\boldsymbol{\theta}^*,h}^{\boldsymbol{\pi}^k}} \left[\left\| \left(\mathbb{P}_{\hat{\theta}_{k,h}} - \mathbb{P}_{\theta_h^*} \right)(s_h, a_h) \right\|_1 \right]$. To limit the ℓ_∞ distance of CDFs via the estimated error above, we adopt the witness rank $\mathrm{d}_{\mathrm{wit}}$ defined in Definition D.7, which is a common structural complexity measure used for model-based RL (Sun et al., 2019; Huang et al., 2022; Chen et al., 2022; Zhong et al., 2022; Liu et al., 2023). By this way, we have the following theoretical guarantees:

Theorem 6.6 (Estimation by Model-Based MLE Approach)**.** *Let* $\beta^{\mathrm{MLE}} := H \log\left(eK\mathcal{N}_{[\cdot]}(\boldsymbol{\Theta}, 1/K, \|\cdot\|_1)/\delta\right)$, *where* $\mathcal{N}_{[\cdot]}(\boldsymbol{\Theta}, 1/K, \|\cdot\|_1)$ *is the bracketing number (see Definition A.2). The estimation function* M-Est-MLE *satisfies the Conditions 6.2 and 6.3 with* $\xi^{\mathrm{MLE}}(K, H, \boldsymbol{\Theta}, \beta^{\mathrm{MLE}}, \delta) = \widetilde{\mathcal{O}}\left(\mathrm{poly}(H) \left(\sqrt{K \, \mathrm{d}_{\mathrm{wit}} \, \beta^{\mathrm{MLE}}} \right) \right)$, *where* $\mathrm{d}_{\mathrm{wit}}$ *is the witness rank of the MDP model. (Definition D.7).*

We present the formal proof in Appendix D. In the risk neutral setting where $L_\infty(\mathbb{E}) = 1$, our Theorem 6.6, in conjunction with Theorem 6.4, presents a regret upper bound that aligns closely with Liu et al. (2023)'s result. We also reference Zhao et al. (2023)'s exploration of RSRL in static CVaR measures and low-rank MDPs, a special subcase of the MDPs with low V-type witness rank (Sun et al., 2019; Agarwal et al., 2020; Uehara et al., 2021). Our analysis extends to V-type witness ranks, offering a more favorable dependence on d, the rank of the transition matrix, compared to Zhao et al. (2023)'s approach. This distinction is elaborated in Appendix D.6, demonstrating our method's

broader applicability and efficiency.

7. Model-Free Meta-Algorithm RS-DisRL-V

In this section, we expand the scope of RSRL to include general value function approximation. We begin by establishing the foundational assumptions for general value function approximation. Then, we present a meta-algorithm **R**isk-**S**ensitive **Dis**tributional **RL** with general **V**alue function approximation (RS-DisRL-V, Algorithm 2). This algorithm's theoretical guarantees and performance are then discussed.

At first, we introduce a generic function class $\boldsymbol{\mathcal{Z}} = \{\mathcal{Z}_h : h \in [H]\}$, with each element $Z_h(s_h^\dagger, a_h) \in \mathcal{Z}_h$ representing a probability distribution, which serves as an estimator candidate for the random variable of the optimal cumulative reward $Z_h^{\boldsymbol{\pi}^*}$. Then we make the foundation assumption of the general value function approximation.

Assumption 7.1 (General value function approximation (Wu et al., 2023; Wang et al., 2023b))**.** For each $h \in [H]$, we have $Z_h^{\boldsymbol{\pi}^*} \in \mathcal{Z}_h$, and the given function set $\mathcal{Z}_h$ satisfy the distributional bellman completeness, such that for any $Z_{h+1} \in \mathcal{Z}_{h+1}$, we have $\mathcal{T}_{h,\boldsymbol{\pi}}^\dagger Z_{h+1} \in \mathcal{Z}_h$.

While we assume the agent has access to a class of random variables containing the value distribution, in practice, we often estimate the random variable through its Cumulative Distribution Function (CDF) or Probability Density Function (PDF). The practical estimation approach will be elaborated upon in the specific settings outlined in Section 7.1 and Section 7.2.

Algorithm 2 RS-DisRL-V

1: **Input:** Function class $\boldsymbol{\mathcal{Z}} = \mathcal{Z}_1 \times \mathcal{Z}_2 \cdots \mathcal{Z}_H$, confidence radius γ.
2: **Initialize:** $\widehat{\mathcal{Z}}_{1,\boldsymbol{\pi}} \leftarrow \mathcal{Z}$.
3: **for** $k = 1$ **to** K **do**
4: $(\boldsymbol{\pi}_k, \widehat{Z}^k) = \operatorname{argmax}_{\boldsymbol{\pi} \in \Pi^\dagger, Z \in \widehat{\mathcal{Z}}_{k,\boldsymbol{\pi}}} \rho(Z_1)$. //Optimistic planning
5: Execute policy $\boldsymbol{\pi}^k$, add the collected data $\boldsymbol{\tau}_k = \{(s_{k,h}, a_{k,h}, r_{k,h})\}_{h=1}^H$ and $\boldsymbol{\pi}^k$, $\widehat{\boldsymbol{\theta}}_k$ into history $\mathcal{H}_k = \mathcal{H}_{k-1} \cup \{(\boldsymbol{\tau}_k, \boldsymbol{\pi}^k, \widehat{Z}^k)\}$. //Data collection
6: $\widehat{\mathcal{Z}}_{k+1,\boldsymbol{\pi}} = \texttt{V-Est}(\mathcal{H}_k, \boldsymbol{\mathcal{Z}}, \boldsymbol{\pi}, \gamma)$. //Confidence set construction
7: **end for**

The meta-algorithm RS-DisRL-V (Algorithm 8) similar in structure to RS-DisRL-M, adopts value-type optimistic planning and confidence set construction. It uniquely constructs a version space for each augmented policy using the V-Est estimation algorithm, incorporating the actual optimal cumulative reward distribution. For case when augmented policy set $\Pi^\dagger$ is infinite, we employ the *policy covering argument*, discretizing the policy set using a normalized lower bracketing set $\underline{\Pi}$ defined in Definition A.2. This approach, commonly used in prior studies (Kallus et al., 2022; Wang et al., 2023b; Zhou et al., 2023; Huang et al., 2022), ensures the practicality and scalability of the algorithm, especially in complex policy environments (see Appendix E.1 for detailed discussion).

Theoretical Guarantees Similar to the Section 6, we introduce two sufficient conditions which describe the validity of the estimation function V-Est to establish the theoretical result for RS-DisRL-V.

Condition 7.2 (Concentration)**.** With probability at least $1-\delta$, $\delta \in (0,1]$, for all policy $\boldsymbol{\pi} \in \Pi^\dagger$, we have that the actual random variable of the cumulative reward collected by policy $\boldsymbol{\pi}$ is in the confidence set with high probability, i.e., $Z^{\boldsymbol{\pi}} \in \widehat{\mathcal{Z}}_{k,\boldsymbol{\pi}}$.

Condition 7.3 (General elliptical potential)**.** For $\delta \in (0,1]$, the ℓ_∞ distance between the CDF of the optimistic reward distribution $\widehat{Z}^k$ and the actual reward distribution $Z^{\boldsymbol{\pi}^k}$ under policy $\boldsymbol{\pi}^k$ can be bounded by $\sum_{k=1}^K \left\| F_{\widehat{Z}^k} - F_{Z^{\boldsymbol{\pi}^k}} \right\|_\infty \leq \zeta(K, H, \mathcal{Z}, \Pi^\dagger, \gamma, \delta)$ with probability at least $1-\delta$.

These two conditions for RS-DisRL-V, paralleling those in RS-DisRL-M, encapsulate the efficiency of the estimation function V-Est. The theoretical result for RS-DisRL-V is presented below.

Theorem 7.4. *Under Assumption 7.1, if the estimation function V-Est satisfies Conditions 7.2 and 7.3, then the regret of RS-DisRL-V can be bounded by* $\mathrm{Regret(K)} \leq L_\infty(\rho)\zeta(K, H, \mathcal{Z}, \Pi^\dagger, \gamma, \delta)$.

The regret bound for the RS-DisRL-V algorithm is characterized by a form similar to that in the model-based case (Theorem 6.4). The effectiveness bound $\zeta(K, H, \mathcal{Z}, \Pi^\dagger, \gamma, \delta)$, sharing the same form as $\widetilde{\mathcal{O}}(\mathrm{poly}(H)\,\mathrm{D}\sqrt{\gamma K/\delta})$ for both LSR (V-Est-LSR, Algorithm 8) and MLE (V-Est-MLE, Algorithm 9) approaches, exhibiting a dependency of $\sqrt{K}$, indicating a sublinear complexity in terms of episodes. Here D is the structural complexity specified in dependency Theorem 7.5 for LSR and Theorem 7.6 for MLE.

7.1. Estimation by Value-Based LSR Approach

In this section, we design a novel LSR approach V-Est-LSR for random-variable estimation through CDFs in the augmented MDP. To the best of our knowledge, we are the first to present the statistically efficient LSR estimation for DisRL with general value function approximation.

Denote $F_h(\cdot|s^\dagger, a)$ as the CDF of $Z_h(\cdot|s^\dagger, a) \in \mathcal{Z}_h$. The estimation function V-Est-LSR focuses on estimating the Bellman operator with real transition probability. For a given target CDF $\widetilde{F}_{h+1}(\cdot|s^\dagger, a)$ and $\boldsymbol{\pi}$, the data collection process in episode k gives us an empir-

ical sample of $\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}\widetilde{F}_{h+1}(x|s^{\dagger}_{k,h},a_{k,h})$ since we observe the transfer from $(s^{\dagger}_{k,h},a_{k,h})$ to $(s^{\dagger}_{k,h+1})$ with policy $\boldsymbol{\pi}^k$. Therefore, we can use $\pi^{\top}_{h+1}\widetilde{F}_{h+1}(x-r_{k,h}|s^{\dagger}_{k,h+1}) := \int_{a_{h+1}}\pi_{h+1}(a_{h+1}|s^{\dagger}_{k,h+1})\widetilde{F}_{h+1}(x-r_{k,h}|s^{\dagger}_{k,h+1},a_{h+1})$ to perform an unbiased estimate of $\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}\widetilde{F}_{h+1}(x|s^{\dagger}_{k,h},a_{k,h})$:

$$\widehat{F}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi},\widetilde{F}} = \operatorname*{argmin}_{F_h\in\mathcal{Z}_h}\sum_{i=1}^{k-1}\Big(F_h(x^{\boldsymbol{\pi},\widetilde{F}}_{k,h}|s^{\dagger}_{k,h},a_{k,h}) - \pi^{\top}_{h+1}\widetilde{F}_{h+1}(x^{\boldsymbol{\pi},\widetilde{F}}_{k,h}-r_{k,h}|s^{\dagger}_{k,h+1})\Big)^2,$$

where $x^{\boldsymbol{\pi},\widetilde{F}}_{i,h}$, defined in Eq.(11), intuitively leads the exploration direction with maximal uncertainty. With this novel estimator, we can prove that with high probability.

$$\sum_{i<k}\Big(\mathcal{T}_{h,\boldsymbol{\pi}}\widetilde{F}_{h+1}(x^{\boldsymbol{\pi},\widetilde{F}}_{i,h}|s^{\dagger}_{i,h},a_{i,h}) - \widehat{F}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi},\widetilde{F}}(x^{\boldsymbol{\pi},\widetilde{F}}_{i,h}|s^{\dagger}_{i,h},a_{i,h})\Big)^2 \le \widetilde{\mathcal{O}}(\gamma^{\mathrm{LSR}}).$$

By a union bound over the covering $\underline{\Pi}^{\dagger}$ and $\underline{\boldsymbol{\mathcal{Z}}}$ (detailed in Appendix E.1), this property derives the concentration condition (Condition 7.2). With the concentration bound, we can easily establish the general elliptical potential condition (Condition 7.3) utilizing the similar argument in model-based LSR method (Section 6.1). We propose the complete pseudocode of `V-Est-LSR` in Algorithm 8. Theoretical guarantees are provided below.

Theorem 7.5. *Let* $\gamma^{\mathrm{LSR}} := 16\log(HK^2/\delta) + \log(\mathcal{N}_C(\Pi^{\dagger},1/K,\|\cdot\|_1)) + \log(\mathcal{N}_C(\boldsymbol{\mathcal{Z}},1/K,\|\cdot\|_1))$. *The estimation function* `V-Est-LSR` *satisfies the Conditions 7.2 and 7.3 with* $\zeta^{\mathrm{LSR}} = \widetilde{\mathcal{O}}\left(\mathrm{poly}(H)\sqrt{K\gamma^{\mathrm{LSR}}\dim_E(\boldsymbol{\mathcal{Z}},\sqrt{K})}\right)$, *where* $\dim_E$ *represents the eluder dimension.*

The formal proof of Theorem 7.5 is presented in Appendix F. The `RS-DisRL-V` algorithm, when implemented with the `V-Est-LSR` estimation function attains a significant regret upper bound $\widetilde{\mathcal{O}}(L_\infty(\rho)\mathrm{poly}(H)\sqrt{K\,\mathrm{D}_{\mathrm{cov}}\dim_E(\boldsymbol{\mathcal{Z}},\sqrt{K})})$ in RSRL with static LRM, where the covering dimension $\mathrm{D}_{\mathrm{cov}} = \log(\mathcal{N}_C(\Pi^{\dagger},1/K,\|\cdot\|_1)) + \log(\mathcal{N}_C(\boldsymbol{\mathcal{Z}},1/K,\|\cdot\|_1))$. This bound, characterized by a $\sqrt{K}$ dependency signifies the first sample-efficient RSRL with general value function approximation and static LRM. Furthermore, when degenerating to the risk-neutral and tabular case, this result aligns with optimal dependencies on K as demonstrated in Wang et al. (2020).

7.2. Estimation by Value-Based MLE Approach

For DisRL with general value function approximation, a standard estimation method adopted is estimating the candidate of true cumulative reward $Z^{\boldsymbol{\pi}}$ by its density function with MLE, which is powerful in theoretical studies (Wang et al., 2023b; Wu et al., 2023) and practice (Hessel et al., 2018; Bellemare et al., 2017). Inspired by previous studies (Wang et al., 2023b; Wu et al., 2023), we combine the standard MLE method with our general risk-sensitive model-free framework and provide the estimation function `V-Est-MLE` which performs efficient estimation in the augmented MDP and risk-sensitive target.

Assume $f_h(x|s^{\dagger}_h,a_h)$ is the PDF of $Z_h(s^{\dagger}_h,a_h)\in\mathcal{Z}_h$. Inspired by the MLE method utilized in risk-neutral DisRL (Wu et al., 2023; Wang et al., 2023b), we estimate by maximizing the log likelihood function: $\log f_h(z^{f,\pi}_{k,h}|s_{k,h},a_{k,h})$ where $z^{f,\pi}_{k,h}$ is sampled from $f_{h+1}(\cdot|s^{\dagger}_{k,h+1},\pi(s_{k,h+1})) + r_{k,h}$. The details of `V-Est-MLE` are presented in Algorithm 9 in appendix due to the space limitation.

The following theorem addresses the efficiency of the MLE approach in risk-sensitive case. We denote d_{BE} as the Bellman eluder dimension (Definition G.3) that aligns with the approaches discussed in prior studies by Jin et al. (2021); Wang et al. (2023b).

Theorem 7.6. *With* $\gamma^{\mathrm{MLE}} = \log(\mathcal{N}_{[\cdot]}(\boldsymbol{\mathcal{Z}},\epsilon,\|\cdot\|_1)) + \log(\mathcal{N}_{[\cdot]}(\Pi^{\dagger},\epsilon,\|\cdot\|_1)) + \log(KH/\delta)$, *estimation function* `V-Est-MLE` *satisfies Conditions 7.2 and 7.3 with* $\zeta^{\mathrm{MLE}} = \widetilde{\mathcal{O}}\left(\mathrm{poly}(H)\sqrt{\mathrm{d}_{\mathrm{BE}}\gamma^{\mathrm{MLE}}K}\right)$. *Here* d_{BE} *represents the Bellman eluder dimension, which is a common structural complexity studied in Jin et al. (2021); Wang et al. (2023b).*

This result enables us to establish a regret upper bound for `RS-DisRL-V` of $\widetilde{\mathcal{O}}(L_\infty(\rho)\mathrm{poly}(H)\sqrt{K\,\mathrm{d}_{\mathrm{BE}}\,\mathrm{D}_{\mathrm{cov}}})$, with the covering dimension $\mathrm{D}_{\mathrm{cov}} = \log(\mathcal{N}_{[\cdot]}(\Pi^{\dagger},1/K,\|\cdot\|_1)) + \log(\mathcal{N}_{[\cdot]}(\boldsymbol{\mathcal{Z}},1/K,\|\cdot\|_1))$. Notably, this bound aligns closely with results from existing research (Jin et al., 2021) in the risk-neutral domain, demonstrating its relevance and applicability in a wide range of reinforcement learning contexts.

8. Conclusion

We give a comprehensive discussion of RS-DisRL with static LRM and general function approximation. We propose the model-based meta-algorithm `RS-DisRL-M` (Algorithm 1) for model-based function approximation and the model-free meta-algorithm `RS-DisRL-V` (Algorithm 2) for the general approximation of value functions. Equipped with our novel LSR or MLE estimation approaches, both meta-algorithms achieve the $\widetilde{\mathcal{O}}(\sqrt{K})$ dependency of the regret upper bound, giving the first statistically efficient algorithms for RSRL with static LRM. Additionally, we establish a computationally tractable and statistically efficient algorithm in the specific setting with static CVaR risk measure and linear function approximation. In this case, we provide numerical experiments to validate the theoretical results (see Appendix I.6).

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

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32, 2019.

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Agarwal, A., Jin, Y., and Zhang, T. VOQL: Towards Optimal Regret in Model-free RL with Nonlinear Function Approximation. In *Proceedings of Thirty Sixth Conference on Learning Theory*, pp. 987–1063. PMLR, July 2023.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. Coherent Measures of Risk. *Mathematical Finance*, 9(3):203–228, 1999. ISSN 1467-9965. doi: 10.1111/1467-9965.00068.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Bastani, O., Ma, J. Y., Shen, E., and Xu, W. Regret Bounds for Risk-Sensitive Reinforcement Learning. *Advances in Neural Information Processing Systems*, 35:36259–36269, December 2022.

Bäuerle, N. and Ott, J. Markov Decision Processes with Average-Value-at-Risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379, December 2011. ISSN 1432-5217. doi: 10.1007/s00186-011-0367-0.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. PMLR, 2017.

Bielecki, T. R., Pliska, S. R., and Sherris, M. Risk sensitive asset allocation. *Journal of Economic Dynamics and Control*, 24(8):1145–1177, 2000.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably Efficient Exploration in Policy Optimization. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1283–1294. PMLR, November 2020.

Candela, E., Doustaly, O., Parada, L., Feng, F., Demiris, Y., and Angeloudis, P. Risk-aware controller for autonomous vehicles using model-based collision prediction and reinforcement learning. *Artificial Intelligence*, 320:103923, 2023.

Chen, Y., Du, Y., Hu, P., Wang, S., Wu, D., and Huang, L. Provably efficient iterated cvar reinforcement learning with function approximation. *arXiv preprint arXiv:2307.02842*, 2023.

Chen, Z., Li, C. J., Yuan, A., Gu, Q., and Jordan, M. I. A general framework for sample-efficient function approximation in reinforcement learning. *arXiv preprint arXiv:2209.15634*, 2022.

Dabney, W., Ostrovski, G., Silver, D., and Munos, R. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pp. 1096–1105. PMLR, 2018.

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*, September 2022.

Ernst, D., Stan, G.-B., Goncalves, J., and Wehenkel, L. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 667–672. IEEE, 2006.

Fei, Y., Yang, Z., Chen, Y., Wang, Z., and Xie, Q. Risk-Sensitive Reinforcement Learning: Near-Optimal Risk-Sample Tradeoff in Regret. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22384–22395. Curran Associates, Inc., 2020.

Fei, Y., Yang, Z., and Wang, Z. Risk-Sensitive Reinforcement Learning with Function Approximation: A Debiasing Approach. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 3198–3207. PMLR, July 2021.

Föllmer, H. and Knispel, T. Convex risk measures: Basic facts, law-invariance and beyond, asymptotics for large portfolios. In *Handbook of the Fundamentals of Financial Decision Making*, volume Volume 4 of *World Scientific Handbook in Financial Economics Series*, pp. 507–554. WORLD SCIENTIFIC, June 2012. ISBN 978-981-4417-34-1. doi: 10.1142/9789814417358_0030.

Frittelli, M. and Gianin, E. R. Law invariant convex risk measures. In Kusuoka, S. and Yamazaki, A. (eds.), *Advances in Mathematical Economics*, Advances in Mathematical Economics, pp. 33–46. Springer, Tokyo, 2005. ISBN 978-4-431-27233-5. doi: 10.1007/4-431-27233-X_2.

Geer, S. A. *Empirical Processes in M-Estimation*, volume 6. Cambridge University Press, January 2000. ISBN 978-0-521-65002-1.

Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., and Silver, D. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Huang, B., Lee, J. D., Wang, Z., and Yang, Z. Towards general function approximation in zero-sum markov games. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=sA4qIu3zv6v.

Jia, G., Xia, J., and Zhao, R. Monetary Risk Measures, December 2020.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-Learning Provably Efficient? In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

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

Jin, C., Liu, Q., and Miryoosefi, S. Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13406–13418. Curran Associates, Inc., 2021.

Kakade, S. M. A Natural Policy Gradient. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

Kallus, N., Mao, X., Wang, K., and Zhou, Z. Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*, pp. 10598–10632. PMLR, 2022.

Keramati, R., Dann, C., Tamkin, A., and Brunskill, E. Being optimistic to be conservative: Quickly learning a cvar policy. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 4436–4443, 2020.

Kusuoka, S. On law invariant coherent risk measures. In Kusuoka, S. and Maruyama, T. (eds.), *Advances in Mathematical Economics*, Advances in Mathematical Economics, pp. 83–95. Springer Japan, Tokyo, 2001. ISBN 978-4-431-67891-5. doi: 10.1007/978-4-431-67891-5_4.

Liang, H. and Luo, Z.-Q. Bridging Distributional and Risk-sensitive Reinforcement Learning with Provable Regret Bounds, October 2022.

Liang, H. and Luo, Z.-q. Regret Bounds for Risk-sensitive Reinforcement Learning with Lipschitz Dynamic Risk Measures, June 2023.

Lim, S. H. and Malik, I. Distributional Reinforcement Learning for Risk-Sensitive Policies. *Advances in Neural Information Processing Systems*, 35:30977–30989, December 2022.

Liu, Q., Chung, A., Szepesvari, C., and Jin, C. When Is Partially Observable Reinforcement Learning Not Scary? In *Proceedings of Thirty Fifth Conference on Learning Theory*, pp. 5175–5220. PMLR, June 2022.

Liu, Q., Netrapalli, P., Szepesvari, C., and Jin, C. Optimistic MLE: A Generic Model-Based Algorithm for Partially Observable Sequential Decision Making. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, pp. 363–376, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 978-1-4503-9913-5. doi: 10.1145/3564246.3585161.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529–533, February 2015. ISSN 1476-4687. doi: 10.1038/nature14236.

Rowland, M., Bellemare, M., Dabney, W., Munos, R., and Teh, Y. W. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 29–37. PMLR, 2018.

Rowland, M., Munos, R., Azar, M. G., Tang, Y., Ostrovski, G., Harutyunyan, A., Tuyls, K., Bellemare, M. G., and Dabney, W. An analysis of quantile temporal-difference learning. *arXiv preprint arXiv:2301.04462*, 2023.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Stanko, S. and Macek, K. Risk-averse distributional reinforcement learning: A cvar optimization approach. In *IJCCI*, pp. 412–423, 2019.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based RL in Contextual Decision Processes: PAC bounds and Exponential Improvements over Model-free Approaches. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pp. 2898–2933. PMLR, June 2019.

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

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

Urpí, N. A., Curi, S., and Krause, A. Risk-Averse Offline Reinforcement Learning. In *International Conference on Learning Representations*, October 2020.

Wang, K., Kallus, N., and Sun, W. Near-minimax-optimal risk-sensitive reinforcement learning with cvar. *arXiv preprint arXiv:2302.03201*, 2023a.

Wang, K., Zhou, K., Wu, R., Kallus, N., and Sun, W. The benefits of being distributional: Small-loss bounds for reinforcement learning. *arXiv preprint arXiv:2305.15703*, 2023b.

Wang, R., Salakhutdinov, R. R., and Yang, L. Reinforcement Learning with General Value Function Approximation: Provably Efficient Approach via Bounded Eluder Dimension. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6123–6135. Curran Associates, Inc., 2020.

Wu, R., Uehara, M., and Sun, W. Distributional Offline Policy Evaluation with Predictive Error Guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 37685–37712. PMLR, July 2023.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Xu, W., Gao, X., and He, X. Regret bounds for markov decision processes with recursive optimized certainty equivalents. *arXiv preprint arXiv:2301.12601*, 2023.

Zhao, Y., Zhan, W., Hu, X., Leung, H.-f., Farnia, F., Sun, W., and Lee, J. D. Provably efficient cvar rl in low-rank mdps. *arXiv preprint arXiv:2311.11965*, 2023.

Zhong, H., Xiong, W., Zheng, S., Wang, L., Wang, Z., Yang, Z., and Zhang, T. Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond. *arXiv preprint arXiv:2211.01962*, 2022.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly Minimax Optimal Reinforcement Learning for Linear Mixture Markov Decision Processes. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pp. 4532–4576. PMLR, July 2021.

Zhou, Z., Athey, S., and Wager, S. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023.

A. Notations

Define Ω as the measurable space containing all the augmented trajectories $\boldsymbol{\tau} = \left\{s_1^\dagger, a_1, \cdots, s_H^\dagger, a_H\right\}$. We consider the probability space $(\Omega, \Sigma, \mathbb{P})$, where Σ is the σ-algebra and $\mathbb{P}$ is the productive probability measure combine the transition distribution and reward distribution.

Let $Z_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}$ be the random variable of $\sum_{h=1}^{H} r_h$ defined on the σ-algebra Σ of Ω. Let $F_Z(x)$ be its cumulative distribution function (CDF). For an augmented trajectory $\boldsymbol{\tau} \in \Omega$ we denote $\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(\boldsymbol{\tau})$ as the probability measure on $\boldsymbol{\tau}$ by following policy $\boldsymbol{\pi}$ under transition model $\boldsymbol{\theta}$, i.e., for any augmented state action pair $s^\dagger, a$,

$$\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(s^\dagger, a) = \int_{\boldsymbol{\tau}} \mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(\boldsymbol{\tau}) \mathbf{1}(\tau_h = (s_h^\dagger, a_h))$$

Since our policy $\boldsymbol{\pi}$ is Markov on the augmented MDP, the visitation μ admits a factorized structure:

$$\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(\boldsymbol{\tau}) = \prod_{h=1}^{H} \pi_h(a_h|s_h^\dagger) \mathbb{T}_{\theta_h}(s_{h+1}^\dagger|s_h^\dagger, a_h)$$

We further define $\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(s^\dagger, a) = \mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(s^\dagger)\boldsymbol{\pi}(a|s^\dagger)$.

We also denote ν as a probability measure defined on the original state action pairs (s, a):

$$\nu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(s, a) = \int_{\boldsymbol{\tau}} \mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(\boldsymbol{\tau}) \mathbf{1}(s, a \in \boldsymbol{\tau}) = \int_{y} \mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}((s, y), a)$$

However, we remark that ν can not be factorized since our policy depends not only on the state s.

Then we introduce the standard concepts of the covering and bracketing numbers for a function class, which are widely employed in the analysis of general function approximation (Ayoub et al., 2020; Liu et al., 2022; 2023; Wang et al., 2023b).

Definition A.1 (Covering Number)**.** The ϵ-covering number of a set $\mathcal{V}$ with metric ρ, denoted as $\mathcal{N}_C(\mathcal{V}, \epsilon, \rho)$, is the minimum integer n such that there exists a subset $\mathbb{V}_o \subset \mathbb{V}$ with cardinality n, for every $x \in \mathbb{V}$, there exists a $y \in \mathbb{V}_o$, with $\rho(x, y) \leq \epsilon$

Definition A.2 (Bracketing Number)**.** Let $\mathcal{G}$ be a set of functions mapping $\mathcal{X} \to \mathbb{R}$. Given $l, u \in \mathcal{G}$ such that $l(x) \leq u(x)$ for all $x \in \mathcal{X}$. We say that the *bracket* $[l, u]$ is the set of functions $g \in \mathcal{G}$ such that $l(x) \leq g(x) \leq u(x)$. for all $x \in \mathbb{X}$. We call $[l, u]$ and ϵ-bracket if $\|u - l\| \leq \epsilon$. Then the ϵ-bracketing number of $\mathcal{G}$ with respect to $\|\cdot\|$ denoted by $\mathcal{N}_{[\cdot]}(\mathcal{G}, \epsilon, \|\cdot\|)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{G}$. And we denote $\mathcal{G}^\downarrow$ as the set of the lower bracket functions l of this ϵ-brackets covering, i.e., $\left|\mathcal{G}^\downarrow\right| = \mathcal{N}_{[\cdot]}(\mathcal{G}, \epsilon, \|\cdot\|)$

Another important concept is the eluder dimension, which will be used to measure the structural complexity in the following LSR analysis. To introduce the eluder dimension, we first define the concept of ε-independence.

Definition A.3 (ε-dependence (Russo & Van Roy, 2013))**.** For $\varepsilon > 0$ and function class $\mathcal{Z}$ whose elements are with domain $\mathcal{X}$, an element $x \in \mathcal{X}$ is ε-dependent on the set $\mathcal{X}_n := \{x_1, x_2, \cdots, x_n\} \subset \mathcal{X}$ with respect to $\mathcal{Z}$, if any pair of functions $z, z' \in \mathcal{Z}$ with $\sqrt{\sum_{i=1}^{n} (z(x_i) - z'(x_i))^2} \leq \varepsilon$ satisfies $z(x) - z'(x) \leq \varepsilon$. Otherwise, x is ε-independent on $\mathcal{X}_n$ if it does not satisfy the condition.

Definition A.4 (Eluder dimension (Russo & Van Roy, 2013))**.** For any $\varepsilon > 0$, and a function class $\mathcal{Z}$ whose elements are in domain $\mathcal{X}$, the Eluder dimension $\dim_E(\mathcal{Z}, \varepsilon)$ is defined as the length of the longest possible sequence of elements in $\mathcal{X}$ such that for some $\varepsilon' \geq \varepsilon$, every element is ε'-independent of its predecessors.

B. General Model-based framework: Algorithm `RS-DisRL-M`

In our model-based framework for Risk-Sensitive Distributional Reinforcement Learning (RS-DisRL), we focus on estimating the transition model, denoted as $\widehat{\boldsymbol{\theta}}_k$ for each episode k. This involves leveraging historical data up to episode $k-1$, represented as $\mathcal{H}_{k-1}$,to construct a confidence set $\widehat{\boldsymbol{\Theta}}_k$. The construction of this set is crucial for guiding the selection of actions, as it is based on a specified confidence radius β, which helps in balancing exploration and exploitation by considering the uncertainty in our model estimates

To ensure the regret bound, we require the following conditions: Condition 6.2 and Condition 6.3.

Algorithm 3 RS-DisRL-M

1: **Input:** Model class $\boldsymbol{\Theta}$, confidence radius β.
2: **Initialize:** $\widehat{\boldsymbol{\Theta}}_1 \leftarrow \boldsymbol{\Theta}$.
3: **for** $k = 1$ **to** K **do**
4: $(\boldsymbol{\pi}^k, \widehat{\boldsymbol{\theta}}_k) = \operatorname{argmax}_{\boldsymbol{\pi}\in\boldsymbol{\Pi}^\dagger, \boldsymbol{\theta}\in\widehat{\boldsymbol{\Theta}}_k} \rho(Z^{\boldsymbol{\pi}}_{\boldsymbol{\theta}})$. //Optimistic planning
5: Execute policy $\boldsymbol{\pi}^k$, add the collected data $\boldsymbol{\tau}_k = \{(s_{k,h}, a_{k,h}, r_{k,h})\}_{h=1}^H$ and $\boldsymbol{\pi}^k$, $\widehat{\boldsymbol{\theta}}_k$ into history $\mathcal{H}_k = \mathcal{H}_{k-1} \cup \{(\boldsymbol{\tau}_k, \boldsymbol{\pi}^k, \widehat{\boldsymbol{\theta}}_k)\}$. //Data collection
6: $\widehat{\boldsymbol{\Theta}}_{k+1} = \texttt{M-Est}\,(\boldsymbol{\Theta}, \mathcal{H}_k, \beta)$. //Confidence set construction
7: **end for**

Condition B.1 (Concentration condition)**.** For $\delta \in (0, 1]$, we have $\boldsymbol{\theta}^* \in \widehat{\boldsymbol{\Theta}}_k$ holds for any $k \in [K]$, with probability at least $1 - \delta$.

Condition B.2 (Elliptical potential condition)**.** If for any $k \in [K]$, we have for any given $\{\widehat{\boldsymbol{\theta}}_k\}_k \subset \{\widehat{\boldsymbol{\Theta}}_k\}_k$ and corresponding greedy policy $\boldsymbol{\pi}^k = \operatorname{argmax}_{\boldsymbol{\pi}\in\Pi^\dagger} \rho(Z^{\boldsymbol{\pi}}_{\boldsymbol{\theta}_k})$, the L_p-norm of the difference of reward-to-gos' CDFs for chosen model $\boldsymbol{\theta}_k$ and true model $\boldsymbol{\theta}^*$ can be bound by

$$\sum_{k=1}^{K} \left\| F_{Z^{\boldsymbol{\pi}_k}_{\boldsymbol{\theta}_k}} - F_{Z^{\boldsymbol{\pi}_k}_{\boldsymbol{\theta}^*}} \right\| \leq \xi\left(K, H, d_\Theta, \beta, \delta\right)$$

with probability at least $1 - \delta$, $\delta > 0$.

This general framework emphasizes that the key to efficiently learning the MDP with a static Lipschitz risk measure is centered on constructing a confidence set for the transition model. This construction leverages the elliptical potential principle for cumulative distribution functions within the augmented MDP. Combined with above conditions, we can establish the following theoretical result

Theorem B.3. *Under Assumption 6.1, if the estimation function* `M-Est` *satisfies Conditions 6.2 and 6.3, then the regret of* `RS-DisRL-M` *(Algorithm 1) can be bounded by* $\mathrm{Regret}(K) \leq L_\infty(\rho)\xi(K, H, \boldsymbol{\Theta}, \beta, \delta)$ *with probability at least* $1 - 2\delta$.

C. Model Based Estimation by LSR Approach

In this section, we design a Least Squares Regression (LSR) based estimation method to construct the confidence set of the model at each episode, and theoretically demonstrate that our algorithm `M-Est-LSR` satisfies the Conditions 6.2 and 6.3.

First we introduce some notations for simplicity. We define $Z_h^{\boldsymbol{\pi},\boldsymbol{\theta}}(s_h, y_h) = \sum_{i=h}^H r_i$ as the random variable of the reward-to-go from step h, where s_h is the starting state, and y_h is the previous cumulative reward from step 1 to $h-1$. Moreover, we denote $F_h^{\boldsymbol{\pi},\boldsymbol{\theta}}(x|s_h, y_h)$ as the CDF of $Z_h^{\boldsymbol{\pi},\boldsymbol{\theta}}$. Our analysis for LSR approach in model-based function approximation is inspired by Chen et al. (2023). However, Chen et al. (2023) focus on Iterated CVaR risk measure and analyse the model with value function in its general function approximation algorithm. In this paper, we develop novel technique for distribution function analysis for general model-based function approximation and augmented MDP.

C.1. Algorithm `M-Est-LSR`

where the distance function is defined by

$$\begin{aligned}\mathrm{Dist}^{\mathrm{LSR}}_{\mathcal{H}_k}(\theta_1||\theta_2) = \sum_{i=1}^{k} \Big(&\int_{s'} \mathbb{P}_{\theta_1}(s'|s_{k,h}, a_{k,h}) \int_r \mathbb{R}(r|s_{i,h}, a_{i,h}) F^{\widehat{\boldsymbol{\pi}}^i, \widehat{\boldsymbol{\theta}}_i}_{h+1}(x_{i,h} - r|s', y_{i,h} + r) \\ &- \int_{s'} \mathbb{P}_{\theta_2}(s'|s_{k,h}, a_{k,h}) \int_r \mathbb{R}(r|s_{i,h}, a_{i,h}) F^{\widehat{\boldsymbol{\pi}}^i, \widehat{\boldsymbol{\theta}}_i}_{h+1}(x_{i,h} - r|s', y_{i,h} + r)\Big)^2,\end{aligned} \tag{1}$$

Algorithm 4 M-Est-LSR($\boldsymbol{\Theta}, \mathcal{H}_{k-1}, \beta^{\mathrm{LSR}}$)

Input: History information $\mathcal{H}_{k-1}$, Model class $\boldsymbol{\Theta}$, and confidence radius β^{LSR}.
Estimate the transition model for every $h \in [H]$

$$\theta_{k,h}^{\mathrm{LSR}} = \operatorname*{argmin}_{\theta_h \in \Theta_h} \sum_{i=1}^{k-1} \Big(\int_r \mathbb{R}(r|s_{i,h}, a_{i,h}) F_{h+1}^{\widehat{\boldsymbol{\pi}}^i, \widehat{\boldsymbol{\theta}}_i}(x_{i,h} - r | s_{i,h+1}, y_{i,h} + r) - \int_{s_{h+1}^\dagger} \mathbb{T}_{\theta_h}(s_{h+1}^\dagger | s_{i,h}^\dagger, a_{i,h}) F_{h+1}^{\widehat{\boldsymbol{\pi}}^i, \widehat{\boldsymbol{\theta}}_i}(x_{i,h} - (y_{h+1} - y_{i,h}) | s_{h+1}, y_{h+1}) \Big)^2 .$$

Construct Confidence set:

$$\widehat{\Theta}_{k,h} := \left\{ \theta_h \in \Theta_h : \mathrm{Dist}_{\mathcal{H}_{k-1},h}^{\mathrm{LSR}}(\theta_h || \theta_{k,h}^{\mathrm{LSR}}) \le \beta^{\mathrm{LSR}} \right\}$$

$$\widehat{\boldsymbol{\Theta}}_k = \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} : \theta_h \in \widehat{\Theta}_{k,h}, h \in [H] \right\}$$

Return $\widehat{\boldsymbol{\Theta}}_k$

$x_{k,h}$ is defined by:

$$x_{i,h} = \operatorname*{argmax}_{x \in \mathbb{R}} \sup_{\theta_h^1 \in \widehat{\Theta}_{k,h}} \int_{s_{h+1}^\dagger} \mathbb{T}_{\theta_h^1}(s_{h+1}^\dagger | s_{k,h}^\dagger, a_{k,h}) F_{h+1}^{\boldsymbol{\pi}^k, \widehat{\boldsymbol{\theta}}_k}(x - (y_{h+1} - y_{k,h}) | s_{h+1}^\dagger) - \inf_{\theta_h^2 \in \widehat{\Theta}_{k,h}} \int_{s_{h+1}^\dagger} \mathbb{T}_{\theta_h^2}(s_{h+1}^\dagger | s_{k,h}^\dagger, a_{k,h}) F_{h+1}^{\boldsymbol{\pi}^k, \widehat{\boldsymbol{\theta}}_k}(x - (y_{h+1} - y_{k,h}) | s_{h+1}^\dagger) , \tag{2}$$

which represents the direction of maximum uncertainty in confidence set $\widehat{\Theta}_{k,h}$, we can obtain the least-squares estimate as:

$$\begin{aligned}
\theta_{k,h}^{\mathrm{LSR}} &= \operatorname*{argmin}_{\theta_h \in \Theta_h} \sum_{i=1}^{k-1} \Bigg(\int_r \mathbb{R}_h(r|s_{i,h}, a_{i,h}) F_{h+1}^{\widehat{\boldsymbol{\pi}}^i, \widehat{\boldsymbol{\theta}}_i}(x_{i,h} - r | s_{i,h+1}, y_{i,h} + r) \\
&\qquad - \int_{s_{h+1}^\dagger} \mathbb{T}_{\theta_h}(s_{h+1}^\dagger | s_{i,h}^\dagger, a_{i,h}) F_{i,h+1}^{\widehat{\boldsymbol{\pi}}^i, \widehat{\boldsymbol{\theta}}_i}(x_{i,h} - (y_{h+1} - y_{i,h}) | s_{h+1}, y_{h+1}) \Bigg)^2 \\
&= \operatorname*{argmin}_{\theta_h \in \Theta_h} \sum_{i=1}^{k-1} \Bigg(\int_r \mathbb{R}_h(r|s_{i,h}, a_{i,h}) F_{h+1}^{\widehat{\boldsymbol{\pi}}^i, \widehat{\boldsymbol{\theta}}_i}(x_{i,h} - r | s_{i,h+1}, y_{i,h} + r) \\
&\qquad - \int_{s_{h+1}} \mathbb{P}_{\theta_h}(s_{h+1} | s_{i,h}, a_{i,h}) \int_r \mathbb{R}_h(r|s_{i,h}, a_{i,h}) F_{h+1}^{\widehat{\boldsymbol{\pi}}^i, \widehat{\boldsymbol{\theta}}_i}(x_{i,h} - r | s_{h+1}, y_{i,h} + r) \Bigg)^2 .
\end{aligned}$$

In the following proof, we show that with $\beta^{\mathrm{LSR}} = 8\log(2H^2\mathcal{N}_C(\boldsymbol{\Theta}, 1/K, \|\cdot\|_1)/\delta) + 4\sqrt{\log(4HK^2/\delta)}$, we have the concentration condition holds with probability at least $1-\delta$. And the elliptical potential condition holds for

$$\xi^{\mathrm{LSR}}(K, H, \Theta, \beta^{\mathrm{LSR}}, \delta) = O(H\sqrt{K} \cdot \sqrt{1 + \dim_E(\mathcal{W}_{\boldsymbol{\Theta}}, 1/\sqrt{K}) + \dim_E(\mathcal{W}_{\boldsymbol{\Theta}}, 1/\sqrt{K})\beta^{\mathrm{LSR}} \log K} + H\sqrt{2K\log(1/\delta)})$$

where $\mathcal{W}_{\boldsymbol{\Theta}} := \{\mathbb{P}_{\theta_h} F : \mathcal{S} \times \mathcal{A} \to [0,1] : \boldsymbol{\theta} \in \boldsymbol{\Theta}, F : \mathcal{S} \to [0,1]\}$ and $\dim_E$ represents the eluder dimension.

C.2. Least Squares Form for Estimation

Notice that we first calculate the estimator kernel

$$\begin{aligned}\theta_{k,h}^{\mathrm{LSR}} =& \operatorname*{argmin}_{\theta_h\in\Theta_h}\sum_{i=1}^{k-1}\Bigg(\int_r \mathbb{R}_h(r|s_{i,h},a_{i,h})F_{h+1}^{\widehat{\boldsymbol{\pi}}^i,\widehat{\boldsymbol{\theta}}_i}(x_{i,h}-r|s_{i,h+1},y_{i,h}+r)\\ &-\int_{s_{h+1}^\dagger}\mathbb{T}_{\theta_h}(s_{h+1}^\dagger|s_{i,h}^\dagger,a_{i,h})F_{i,h+1}^{\widehat{\boldsymbol{\pi}}^i,\widehat{\boldsymbol{\theta}}_i}(x_{i,h}-(y_{h+1}-y_{i,h})|s_{h+1},y_{h+1})\Bigg)^2\\ =& \operatorname*{argmin}_{\theta_h\in\Theta_h}\sum_{i=1}^{k-1}\Bigg(\int_r \mathbb{R}_h(r|s_{i,h},a_{i,h})F_{h+1}^{\widehat{\boldsymbol{\pi}}^i,\widehat{\boldsymbol{\theta}}_i}(x_{i,h}-r|s_{i,h+1},y_{i,h}+r)\\ &-\int_{s_{h+1}}\mathbb{P}_{\theta_h}(s_{h+1}|s_{i,h},a_{i,h})\int_r \mathbb{R}_h(r|s_{i,h},a_{i,h})F_{h+1}^{\widehat{\boldsymbol{\pi}}^i,\widehat{\boldsymbol{\theta}}_i}(x_{i,h}-r|s_{h+1},y_{i,h}+r)\Bigg)^2.\end{aligned}$$

which takes the least-square regression form. If we denote the mixed contribution function

$$\widehat{F}_{i,h}(s) := \int_r \mathbb{R}_h(r|s_{i,h},a_{i,h})F_{h+1}^{\widehat{\boldsymbol{\pi}}^i,\widehat{\boldsymbol{\theta}}_i}(x_{i,h}-r|s,y_{i,h}+r)\,.$$

Thus we can simplify the least squares regression as

$$\theta_{k,h}^{\mathrm{LSR}} = \operatorname*{argmin}_{\theta_h\in\Theta_h}\sum_{i=1}^{k-1}\left(\widehat{F}_{i,h}(s_{i,h+1})-\int_{s_{h+1}}\mathbb{P}_{\theta_h}(s_{h+1}|s_{i,h},a_{i,h})\widehat{F}_{i,h}(s_{h+1})\right)^2.$$

We can further define

$$[\mathbb{P}_{\theta_h}\widehat{F}_{i,h}](s_h,a_h) := \int_{s_{h+1}}\mathbb{P}_{\theta_h}(s_{h+1}|s_h,a_h)\widehat{F}_{i,h}(s_{h+1})$$

and $\mathcal{W}_{\boldsymbol{\Theta}} := \{\mathbb{P}_{\theta_h}F : \mathcal{S}\times\mathcal{A}\to[0,1] : \boldsymbol{\theta}\in\boldsymbol{\Theta}, F:\mathcal{S}\to[0,1]\}$.

Then the distance function and our constructed confidence sets can be expressed by

$$\mathrm{Dist}_{\mathcal{H}_k,h}^{\mathrm{LSR}}(\theta_1||\theta_2) = \sum_{i=1}^{k}\left([\mathbb{P}_{\theta_1}\widehat{F}_{i,h}](s_h,a_h)-[\mathbb{P}_{\theta_2}\widehat{F}_{i,h}](s_h,a_h)\right)^2,$$

$$\widehat{\Theta}_{k,h} := \left\{\theta_h\in\Theta_h : \mathrm{Dist}_{\mathcal{H}_{k-1},h}^{\mathrm{LSR}}(\theta_h||\widehat{\theta}_{k,h}^{\mathrm{LSR}})\le\beta^{\mathrm{LSR}}\right\},$$

$$\widehat{\boldsymbol{\Theta}}_k = \left\{\boldsymbol{\theta}\in\boldsymbol{\Theta} : \theta_h\in\widehat{\Theta}_{k,h}, h\in[H]\right\}.$$

C.3. Concentration Condition for LSR approach

Lemma C.1 (LSR concentration)**.** *The LSR-type construction algorithm* `M-Est-LSR` *satisfies Condition 6.2. That is, for* $\delta\in(0,1]$*, with probability at least* $1-\delta$*, we have* $\theta_h^*\in\widehat{\boldsymbol{\Theta}}_k$ *for all* $k\in[K]$*,*

Proof. Recall that we calculate the estimation kernel $\theta_{k,h}^{\mathrm{LSR}}$ by least squares regression as follows:

$$\theta_{k,h}^{\mathrm{LSR}} = \operatorname*{argmin}_{\theta_h\in\Theta_h}\sum_{i=1}^{k-1}\left(\widehat{F}_{i,h}(s_{i,h+1})-[\mathbb{P}_{\theta_h}\widehat{F}_{i,h}](s_{i,h},a_{i,h})\right)^2.$$

Notice that the $\widehat{F}_{i,h}(s_{i,h+1})$ is $\sigma_{i,h+1}$-measurable by definition and $[\mathbb{P}_{\theta_h}\widehat{F}_{i,h}](s_{i,h},a_{i,h})$ is $\sigma_{i,h}$-measurable, with $\sigma_{i,h}$ be the filtration containing history up to the h step in episode i. We have

$$\mathbb{E}\left[\widehat{F}_{i,h}(s_{i,h+1})\middle|\sigma_{i,h}\right] = [\mathbb{P}_{\theta_h^*}\widehat{F}_{i,h}](s_{i,h},a_{i,h})\,.$$

Based on concentration lemma H.2, we have the following holds with probability at least $1-\delta/H$

$$\sum_{i=1}^{k}\left([\mathbb{P}_{\theta_h^*}\widehat{F}_{i,h}](s_{i,h},a_{i,h})-[\mathbb{P}_{\theta_{k,h}^{\mathrm{LSR}}}\widehat{F}_{i,h}](s_{i,h},a_{i,h})\right)^2 \le 8\log(2H\mathcal{N}(\mathcal{W}_{\boldsymbol{\Theta}},1/K,\|\cdot\|_\infty)/\delta)+4\sqrt{\log(4K^2/\delta)}\,.$$

Moreover, we have for any $\boldsymbol{\theta}^1,\boldsymbol{\theta}^2\in\boldsymbol{\Theta}$, we can bound the supremum distance of $\mathbb{P}_{\theta_h^1}F,\mathbb{P}_{\theta_h^2}F\in\mathcal{W}_{\boldsymbol{\Theta}}$ for any $h\in[H]$ by

$$\begin{aligned}\|\mathbb{P}_{\theta_h^1}F-\mathbb{P}_{\theta_h^2}F\|_\infty &= \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|\int_{s'}(\mathbb{P}_{\theta_h^1}(s'|s,a)F(s')-\mathbb{P}_{\theta_h^2}(s'|s,a)F(s'))\right|\\ &= \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|\int_{s'}(\mathbb{P}_{\theta_h^1}(s'|s,a)-\mathbb{P}_{\theta_h^2}(s'|s,a))F(s')\right|\\ &\le \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\mathbb{P}_{\theta_h^1}(s,a)-\mathbb{P}_{\theta_h^2}(s,a)\|_1\\ &\le \|\boldsymbol{\theta}^1-\boldsymbol{\theta}^2\|_1\,.\end{aligned}$$

Thus we have $\mathcal{N}_C(\mathcal{W}_{\boldsymbol{\Theta}},1/K,\|\cdot\|_\infty)\le\mathcal{N}_C(\boldsymbol{\Theta},1/K,\|\cdot\|_1)$. Recall the definition of $\mathrm{Dist}^{\mathrm{LSR}}_{\mathcal{H}_{k-1},h}$. Taking union bound over $h\in[H]$, we have

$$\mathrm{Dist}^{\mathrm{LSR}}_{\mathcal{H}_{k-1},h}(\theta_h^*||\theta_{k,h}^{\mathrm{LSR}})\le 8\log(2H^2\mathcal{N}_C(\boldsymbol{\Theta},1/K,\|\cdot\|_1)/\delta)+4\sqrt{\log(4HK^2/\delta)}=\beta^{\mathrm{LSR}}\,,$$

which shows that $\theta_h^*\in\widehat{\Theta}_{k,h}$ for every $k,h\in[K]\times[H]$ with probability at least $1-\delta$ and implies that $\boldsymbol{\theta}^*\in\widehat{\boldsymbol{\Theta}}_k$ for every $k\in[K]$ with probability at least $1-\delta$. □

C.4. Elliptical Potential Condition for LSR Approach

We have the bellman equation for distributional function

$$F_h^{\boldsymbol{\pi},\boldsymbol{\theta}}(x|s_h,y_h)=F_h^{\boldsymbol{\pi},\boldsymbol{\theta}}(x|s_h^\dagger)=\int_{s_{h+1}^\dagger}\int_{a_h}\pi_h(a_h|s_h^\dagger)\mathbb{T}_{\theta_h}(s_{h+1}^\dagger|s_h^\dagger,a_h)F_{h+1}^{\boldsymbol{\pi},\boldsymbol{\theta}}(x-(y_{h+1}-y_h)|s_{h+1},y_{h+1})\,.$$

$$\mathcal{T}_{\theta_h,\boldsymbol{\pi}}F_{h+1}^{\boldsymbol{\pi},\boldsymbol{\theta}}(x|s_h^\dagger)=\int_{s_{h+1}^\dagger}\int_{a_h}\pi_h(a_h|s_h^\dagger)\mathbb{T}_{\theta_h}(s_{h+1}^\dagger|s_h^\dagger,a_h)F_{h+1}^{\boldsymbol{\pi},\boldsymbol{\theta}}(x-(y_{h+1}-y_h)|s_{h+1},y_{h+1})\,.$$

Lemma C.2. *With probability at least $1-\delta$, we have*

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\sup_{\theta_h^1\in\widehat{\Theta}_{k,h}}[\mathbb{P}_{\theta_h^1}\widehat{F}_{k,h}](s_{k,h},a_{k,h})-\inf_{\theta_h^2\in\widehat{\Theta}_{k,h}}[\mathbb{P}_{\theta_h^2}\widehat{F}_{k,h}](s_{k,h},a_{k,h})\right)^2\\ \le& H+H\dim_E(\mathcal{W}_{\boldsymbol{\Theta}},1/\sqrt{K})+4H\beta^{\mathrm{LSR}}\dim_E(\mathcal{W}_{\boldsymbol{\Theta}},1/\sqrt{K})(\log(K)+1)\,.\end{aligned}$$

Proof. This proof is almost same with the elliptical potential lemma for general function approximation given in Lemma 9 of Chen et al. (2023). □

Lemma C.3 (LSR elliptical potential)**.** *The algorithm* `M-Est-LSR` *satisfies Condition 6.3 with*

$$\xi^{\mathrm{LSR}}(K,H,\boldsymbol{\Theta},\beta^{\mathrm{LSR}},\delta)=\widetilde{\mathcal{O}}\left(H\sqrt{K\beta^{\mathrm{LSR}}\dim_E(\mathcal{W}_{\boldsymbol{\Theta}},1/\sqrt{K})}\right)\,.$$

Proof. We have

$$
\begin{aligned}
&\sum_{k=1}^{K} \sup_{x\in[0,H]} \left| F_{Z^{\boldsymbol{\pi}^k}_{\widehat{\boldsymbol{\theta}}_k}}(x) - F_{Z^{\boldsymbol{\pi}^k}_{\boldsymbol{\theta}^*}}(x) \right| \\
=&\sum_{k=1}^{K} \sup_{x\in[0,H]} \left| F_1^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x|s_{k,1},0) - F_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^*}(x|s_{k,1},0) \right| \\
=&\sum_{k=1}^{K} \sup_{x\in[0,H]} \left| \mathcal{T}_{\widehat{\theta}_{k,1},\boldsymbol{\pi}^k} F_2^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x|s_{k,1},0) - \mathcal{T}_{\theta_1^*,\boldsymbol{\pi}^k} F_2^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^*}(x|s_{k,1},0) \right| \\
\le&\sum_{k=1}^{K} \sup_{x\in[0,H]} \left| \mathcal{T}_{\widehat{\theta}_{k,1},\boldsymbol{\pi}^k} F_2^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x|s_{k,1}^\dagger) - \mathcal{T}_{\theta_1^*,\boldsymbol{\pi}^k} F_2^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x|s_{k,1}^\dagger) \right| \\
&+\sum_{k=1}^{K} \sup_{x\in[0,H]} \left| \mathcal{T}_{\theta_1^*,\boldsymbol{\pi}^k} F_2^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x|s_{k,1}^\dagger) - \mathcal{T}_{\theta_1^*,\boldsymbol{\pi}^k} F_2^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^*}(x|s_{k,1}^\dagger) \right| \\
\le&\sum_{k=1}^{K} \sup_{x\in[0,H]} \Bigg| \int_{s_2^\dagger} \mathbb{T}_{\widehat{\theta}_{k,1}}(s_2^\dagger|s_{k,1}^\dagger,a_{k,1}) F_2^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x-(y_2-y_{k,1})|s_2^\dagger) \\
&\qquad - \int_{s_2^\dagger} \mathbb{T}_{\theta_1^*}(s_2^\dagger|s_{k,1}^\dagger,a_{k,1}) F_2^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x-(y_2-y_{k,1})|s_2^\dagger) \Bigg| \\
&+\sum_{k=1}^{K} \sup_{x\in[0,H]} \left| F_2^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x|s_{k,2}^\dagger) - F_2^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^*}(x|s_{k,2}^\dagger) \right| \\
&+\sum_{k=1}^{K} \Delta_{k,1}^{(1)} + \Delta_{k,1}^{(2)} + \Delta_{k,1}^{(3)} + \Delta_{k,1}^{(4)},
\end{aligned}
$$

where the sequence $\Delta_{k,h}^{(i)}$ for $i=1,2,3,4$ is defined as follows.

$$
\begin{aligned}
\Delta_{k,h}^{(1)} =& \sup_{x\in[0,H]} \left| \int_{s_{h+1}^\dagger} \mathbb{T}_{\widehat{\theta}_{k,h}}(s_{h+1}^\dagger|s_{k,h}^\dagger,a_{k,h}) F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x-(y_{h+1}-y_{k,h})|s_{h+1}^\dagger) - \mathcal{T}_{\widehat{\theta}_{k,h},\boldsymbol{\pi}^k} F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x|s_{k,h}^\dagger) \right| \\
=& \sup_{x\in[0,H]} \Bigg| \int_{s_{h+1}^\dagger} \mathbb{T}_{\widehat{\theta}_{k,h}}(s_{h+1}^\dagger|s_{k,h}^\dagger,a_{k,h}) F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x-(y_{h+1}-y_{k,h})|s_{h+1}^\dagger) \\
&\qquad - \int_{a_h} \pi_h^k(a_h|s_{k,h}^\dagger) \int_{s_{h+1}^\dagger} \mathbb{T}_{\widehat{\theta}_{k,h}}(s_{h+1}^\dagger|s_{k,h}^\dagger,a_h) F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x-(y_{h+1}-y_{k,h})|s_{h+1}^\dagger,y_{h+1}) \Bigg|,
\end{aligned}
$$

$$
\begin{aligned}
\Delta_{k,h}^{(2)} =& \sup_{x\in[0,H]} \left| \int_{s_{h+1}^\dagger} \mathbb{T}_{\theta_h^*}(s_{h+1}^\dagger|s_{k,h}^\dagger,a_{k,h}) F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x-(y_{h+1}-y_{k,h})|s_{h+1}^\dagger) - \mathcal{T}_{\theta_h^*,\boldsymbol{\pi}^k} F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x|s_{k,h}^\dagger) \right| \\
=& \sup_{x\in[0,H]} \Bigg| \int_{s_{h+1}^\dagger} \mathbb{T}_{\theta_h^*}(s_{h+1}^\dagger|s_{k,h}^\dagger,a_{k,h}) F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x-(y_{h+1}-y_{k,h})|s_{h+1}^\dagger) \\
&\qquad - \int_{a_h} \pi_h^k(a_h|s_{k,h}^\dagger) \int_{s_{h+1}^\dagger} \mathbb{T}_{\theta_h^*}(s_{h+1}^\dagger|s_{k,h}^\dagger,a_h) F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x-(y_{h+1}-y_{k,h})|s_{h+1}^\dagger) \Bigg|,
\end{aligned}
$$

$$
\Delta_{k,h}^{(3)} = \sup_{x\in[0,H]} \left| \mathcal{T}_{\theta_h^*,\boldsymbol{\pi}^k} F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x|s_{k,h}^\dagger) - F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x-r_{k,h}|s_{k,h+1}^\dagger) \right|,
$$

$$
\Delta_{k,h}^{(4)} = \sup_{x\in[0,H]} \left| \mathcal{T}_{\theta_h^*,\boldsymbol{\pi}^k} F_{h+1}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^*}(x|s_{k,h}^\dagger) - F_{h+1}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^*}(x-r_{k,h}|s_{k,h+1}^\dagger) \right|.
$$

Thus we have $\mathbb{E}[\Delta_{k,h}^{(i)}|\sigma_{k,h}] = 0$ for any $k \in [K], h \in [H]$ and $i \in \{1,2,3,4\}$. Thus we have $\{\Delta_{k,h}^{(i)}\}_{i=1}^{k}$ is a martingale difference sequence. Repeat the above method for h steps,

$$
\begin{aligned}
&\sum_{k=1}^{K} \left\| F_{Z_{\widehat{\boldsymbol{\theta}}_k}^{\boldsymbol{\pi}^k}} - F_{Z_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}^k}} \right\|_\infty \\
\leq &\sum_{k=1}^{K}\sum_{h=1}^{H} \sup_{x\in[0,H]} \Bigg| \int_{s_{h+1}^\dagger} \mathbb{T}_{\widehat{\theta}_{k,h}}(s_{h+1}^\dagger|s_{k,h}^\dagger, a_{k,h}) F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x-(y_{h+1}-y_{k,h})|s_{h+1}^\dagger) \\
&\qquad \underbrace{- \int_{s_{h+1}^\dagger} \mathbb{T}_{\theta_h^*}(s_{h+1}^\dagger|s_{k,h}^\dagger, a_{k,h}) F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x-(y_{h+1}-y_{k,h})|s_{h+1}^\dagger) \Bigg|}_{I} \\
&+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \Delta_{k,h}^{(1)} + \Delta_{k,h}^{(2)} + \Delta_{k,h}^{(3)} + \Delta_{k,h}^{(4)}}_{J} .
\end{aligned}
$$

Applying the standard Azuma-Hoeffding inequality to the martingale difference sequence, we have

$$J \leq \widetilde{\mathcal{O}}(H\sqrt{K})$$

The main challenge falls in bounding the term I. By Lemma C.1, we have with probability at least $1-\delta$, $\theta_h^* \in \widehat{\Theta}_{k,h}$ holds for all $k \in [K]$ and $h \in [H]$. Therefore,

$$
\begin{aligned}
I \leq &\sum_{k=1}^{K}\sum_{h=1}^{H} \sup_{x\in[0,H]} \sup_{\theta_h^1 \in \widehat{\Theta}_{k,h}} \int_{s_{h+1}^\dagger} \mathbb{T}_{\theta_h^1}(s_{h+1}^\dagger|s_{k,h}^\dagger, a_{k,h}) F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x-(y_{h+1}-y_{k,h})|s_{h+1}^\dagger) \\
&\qquad - \inf_{\theta_h^2 \in \widehat{\Theta}_{k,h}} \int_{s_{h+1}^\dagger} \mathbb{T}_{\theta_h^2}(s_{h+1}^\dagger|s_{k,h}^\dagger, a_{k,h}) F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x-(y_{h+1}-y_{k,h})|s_{h+1}^\dagger) \\
= &\sum_{k=1}^{K}\sum_{h=1}^{H} \sup_{\theta_h^1 \in \widehat{\Theta}_{k,h}} \int_{s_{h+1}^\dagger} \mathbb{T}_{\theta_h^1}(s_{h+1}^\dagger|s_{k,h}^\dagger, a_{k,h}) F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x_{k,h}-(y_{h+1}-y_{k,h})|s_{h+1}^\dagger) \\
&\qquad - \inf_{\theta_h^2 \in \widehat{\Theta}_{k,h}} \int_{s_{h+1}^\dagger} \mathbb{T}_{\theta_h^2}(s_{h+1}^\dagger|s_{k,h}^\dagger, a_{k,h}) F_{h+1}^{\boldsymbol{\pi}^k,\widehat{\boldsymbol{\theta}}_k}(x_{k,h}-(y_{h+1}-y_{k,h})|s_{h+1}^\dagger) \\
= &\sum_{k=1}^{K}\sum_{h=1}^{H} \sup_{\theta_h^1 \in \widehat{\Theta}_{k,h}} [\mathbb{P}_{\theta_h^1} \widehat{F}_{k,h}](s_{k,h}, a_{k,h}) - \inf_{\theta_h^2 \in \widehat{\Theta}_{k,h}} [\mathbb{P}_{\theta_h^2} \widehat{F}_{k,h}](s_{k,h}, a_{k,h}) ,
\end{aligned}
$$

where the first inequality holds by $\widehat{\boldsymbol{\theta}}_k \in \widehat{\boldsymbol{\Theta}}_k$ and $\boldsymbol{\theta}^* \in \widehat{\boldsymbol{\Theta}}_k$ with high probability, the first equality is due to the definition of $x_{k,h}$, and the last equality holds by the definition of $\widehat{F}_{k,h}(s)$. By Lemma C.2 and Cauchy-Schwartz inequality, we have

$$I \leq \widetilde{\mathcal{O}}\left(H\sqrt{K\beta^{\mathrm{LSR}} \dim_E(\mathcal{W}_{\boldsymbol{\Theta}}, 1/\sqrt{K})}\right) .$$

Overall, we can conclude the result

$$\sum_{k=1}^{K} \left\| F_{Z_{\widehat{\boldsymbol{\theta}}_k}^{\boldsymbol{\pi}^k}} - F_{Z_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}^k}} \right\|_\infty \leq \widetilde{\mathcal{O}}\left(H\sqrt{K\beta^{\mathrm{LSR}} \dim_E(\mathcal{W}_{\boldsymbol{\Theta}}, 1/\sqrt{K})}\right) .$$

□

D. Model-Based Estimation by MLE Approach

In this section we propose our algorithm and analysis for model-based risk-sensitive RL via the MLE estimation approach.

Algorithm 5 M-Est-MLE$(\boldsymbol{\Theta}, \mathcal{H}_{k-1}, \beta)$

Input: History information $\mathcal{H}_{k-1}$, Model class $\boldsymbol{\Theta}$, and confidence radius β^{MLE}.
Estimate the transition model for every $h \in [H]$

$$\theta_{k,h}^{\mathrm{MLE}} = \operatorname*{argmax}_{\theta_h \in \boldsymbol{\Theta}_h} \sum_{i=1}^{k-1} \log \left[\mathbb{P}_{\boldsymbol{\theta}_h}(s_{i,h+1}|s_{i,h}, a_{i,h})\right] .$$

Construct Confidence set:

$$\widehat{\boldsymbol{\Theta}}_k = \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} : \sum_{i=1}^{k-1} \sum_{h=1}^{H} \log \mathbb{P}_{\theta_h}(s_{i,h+1}|s_{i,h}, a_{i,h}) \geq \sum_{i=1}^{k-1} \sum_{h=1}^{H} \log \mathbb{P}_{\theta_{k,h}^{\mathrm{MLE}}}(s_{i,h+1}|s_{i,h}, a_{i,h}) - \beta^{\mathrm{MLE}} \right\} .$$

Return $\widehat{\boldsymbol{\Theta}}_k^{\mathrm{MLE}}$.

D.1. Algorithm M-Est-MLE

Here we present our construction of confidence set via MLE in Algorithm 5, which is inspired by the OMLE algorithm of Liu et al. (2023). In this algorithm, we first calculate the maximal likelihood estimator $\theta_{k,h}^{\mathrm{MLE}}$ for each step h based on the history $\mathcal{H}_{k-1}$ before episode k by the following equation.

$$\theta_{k,h}^{\mathrm{MLE}} = \operatorname*{argmax}_{\theta_h \in \Theta_h} \sum_{i=1}^{k-1} \log \mathbb{P}_{\theta_h}(s_{i,h+1}|s_{i,h}, a_{i,h}) .$$

Then we can construct the confidence set centered at the maximal likelihood estimator with radius β:

$$\widehat{\boldsymbol{\Theta}}_k = \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} : \sum_{i=1}^{k-1} \sum_{h=1}^{H} \log \mathbb{P}_{\theta_h}(s_{i,h+1}|s_{i,h}, a_{i,h}) \geq \sum_{i=1}^{k-1} \sum_{h=1}^{H} \log \mathbb{P}_{\theta_{k,h}^{\mathrm{MLE}}}(s_{i,h+1}|s_{i,h}, a_{i,h}) - \beta^{\mathrm{MLE}} \right\} .$$

Where $\beta^{\mathrm{MLE}} = H \log \left(eHK\mathcal{N}_{[\cdot]}(\epsilon, \Theta, \|\cdot\|_1)/\delta\right)$.

D.2. Simulation Lemma in Augmented MDP

Next, we build the relationship between the supremum norm of the cumulative distribution function and the ℓ_1-norm of the trajectory probability kernel $\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}$. Throughout this section, we define $R(\boldsymbol{\tau})$ as the cumulative reward of the trajectory $\boldsymbol{\tau}$.

Lemma D.1 (Distribution difference)**.** *For any fixed model $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $Z_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}$ is the random variable of the cumulative reward collected by policy $\boldsymbol{\pi}$ in the MDP modeled by $\boldsymbol{\theta}$.Thus we have the following holds*

$$\left\| F_{Z_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}} - F_{Z_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}} \right\|_\infty \leq \|\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}} - \mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}\|_1 .$$

Proof. By the definition of the CDF, we have $F_{Z_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}}(x) = \int_{\boldsymbol{\tau}} \mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(\boldsymbol{\tau}) \mathbb{I}(R(\boldsymbol{\tau}) \leq x)$. Thus we have

$$\begin{aligned}
\left\| F_{Z_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}} - F_{Z_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}} \right\|_\infty &= \sup_{x \in [0,H]} \left| F_{Z_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}}(x) - F_{Z_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}}(x) \right| \\
&= \sup_{x \in [0,H]} \left| \int_{\boldsymbol{\tau}} \left(\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(\boldsymbol{\tau}) - \mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}(\boldsymbol{\tau})\right) \mathbb{I}(R(\boldsymbol{\tau}) \leq x) \right| \\
&\leq \int_{\boldsymbol{\tau}} |\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(\boldsymbol{\tau}) - \mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}(\boldsymbol{\tau})| \\
&= \|\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}} - \mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}\|_1 ,
\end{aligned}$$

where the first inequality holds by the triangle inequality, and the last equality is due to the definition of the ℓ_1-norm. □

We then establish the simulation lemma for augmented MDP, which connect the ℓ_1-norm difference of the probability measure $\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}$ on the augmented MDP with the ℓ_1-norm difference of transition probabilities. This is one of the key lemmas that bridge the gap between the analysis in augmented MDP and origin MDP.

Lemma D.2 (Augmented simulation lemma)**.**

$$\|\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}} - \mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}\|_1 \leq \sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim\nu_{\boldsymbol{\theta}^*,h}^{\boldsymbol{\pi}}} \left\|\mathbb{P}_{\theta_h}(s_h,a_h) - \mathbb{P}_{\theta_h^*}(s_h,a_h)\right\|_1 \leq 2H \|\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}} - \mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}\|_1 \ .$$

Proof. First, we decompose $\|\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}} - \mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}\|_1$. Following standard simulation lemma analysis techniques, we have for any trajectory $\boldsymbol{\tau}$,

$$\begin{aligned}
&|\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(\boldsymbol{\tau}) - \mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}(\boldsymbol{\tau})| \\
=&\left|\prod_{h=1}^{H} \mathbb{T}_{\theta_h^*}(s_{h+1}^\dagger|s_h^\dagger,a_h)\pi_h(a_h|s_h^\dagger) - \prod_{i=1}^{H} \mathbb{T}_{\theta_h}(s_{h+1}^\dagger|s_h^\dagger,a_h)\pi_h(a_h|s_h^\dagger)\right| \\
=&\left|\sum_{h=1}^{H}\prod_{i=1}^{h-1} \mathbb{T}_{\theta_i^*}(s_{i+1}^\dagger|s_i^\dagger,a_i)\pi_i(a_i|s_i^\dagger)\left(\mathbb{T}_{\theta_h^*}(s_{h+1}^\dagger|s_h^\dagger,a_h) - \mathbb{T}_{\theta_h}(s_{h+1}^\dagger|s_h^\dagger,a_h)\right)\pi_h(a_h|s_h^\dagger)\prod_{i=h+1}^{H} \mathbb{T}_{\theta_i}(s_{i+1}^\dagger|s_i^\dagger,a_i)\pi_i(a_i|s_i^\dagger)\right| \\
\leq&\sum_{h=1}^{H}\prod_{i=1}^{h-1} \mathbb{T}_{\theta_i^*}(s_{i+1}^\dagger|s_i^\dagger,a_i)\pi_i(a_i|s_i^\dagger)\left|\mathbb{T}_{\theta_h^*}(s_{h+1}^\dagger|s_h^\dagger,a_h) - \mathbb{T}_{\theta_h}(s_{h+1}^\dagger|s_h^\dagger,a_h)\right|\pi_h(a_h|s_h^\dagger)\,.
\end{aligned}$$

Integral the above inequality over the entire space Ω, we have

$$\begin{aligned}
&\int_\Omega |\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(\boldsymbol{\tau}) - \mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}(\boldsymbol{\tau})|\,d\boldsymbol{\tau} = \int_{\boldsymbol{\tau}} |\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}}(\boldsymbol{\tau}) - \mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}(\boldsymbol{\tau})| = \|\mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi}} - \mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}}\|_1 \\
\leq&\sum_{h=1}^{H}\int_\Omega \prod_{i=1}^{h-1} \mathbb{T}_{\theta_i^*}(s_{i+1}^\dagger|s_i^\dagger,a_i)\pi_i(a_i|s_i^\dagger)\pi_h(a_h|s_h^\dagger)\left|\mathbb{T}_{\theta_h^*}(s_{h+1}^\dagger|s_h^\dagger,a_h) - \mathbb{T}_{\theta_h}(s_{h+1}^\dagger|s_h^\dagger,a_h)\right| d\boldsymbol{\tau} \\
=&\sum_{h=1}^{H}\int_{(s_h^\dagger,a_h,s_{h+1}^\dagger)\in\mathcal{S}^\dagger\times\mathcal{A}\times\mathcal{S}^\dagger}\left|\mathbb{T}_{\theta_h^*}(s_{h+1}^\dagger|s_h^\dagger,a_h) - \mathbb{T}_{\theta_h}(s_{h+1}^\dagger|s_h^\dagger,a_h)\right| ds_{h+1}^\dagger d\mu_{\boldsymbol{\theta}^*,h}^{\boldsymbol{\pi}}(s_h^\dagger,a_h) \\
=&\sum_{h=1}^{H}\mathbb{E}_{(s_h^\dagger,a_h)\sim\mu_{\boldsymbol{\theta}^*,h}^{\boldsymbol{\pi}}}\left\|\mathbb{T}_{\theta_h}(s_h^\dagger,a_h) - \mathbb{T}_{\theta_h^*}(s_h^\dagger,a_h)\right\|_1 \ ,
\end{aligned}$$

where $\mu_{\boldsymbol{\theta},h}^{\boldsymbol{\pi}}(s_h^\dagger,a_h)$ represent the probability of arriving the augmented state-action pair $(s_h^\dagger,a_h)$ at step h with policy $\boldsymbol{\pi}$ in the MDP modeled by $\boldsymbol{\theta}$. Moreover, we define $\mu_{\boldsymbol{\theta},h}^{\boldsymbol{\pi}}(s_h^\dagger,a_h,s_{h+1}^\dagger)$ as the probability of visit $(s_h^\dagger,a_h,s_{h+1}^\dagger)$ at step h and $h+1$, i.e., $\mu_{\boldsymbol{\theta},h}^{\boldsymbol{\pi}}(s_h^\dagger,a_h,s_{h+1}^\dagger) = \mu_{\boldsymbol{\theta},h}^{\boldsymbol{\pi}}(s_h^\dagger,a_h)\mathbb{T}_h(s_{h+1}^\dagger|s_h^\dagger,a_h)$. With this fact, we can rewrite the summation of the ℓ_1-norm difference of the transition probabilities of the augmented MDP.

$$\begin{aligned}
&\mathbb{E}_{(s_h^\dagger,a_h)\sim\mu_{\boldsymbol{\theta}^*,h}^{\boldsymbol{\pi}}}\left\|\mathbb{T}_{\theta_h}(s_h^\dagger,a_h) - \mathbb{T}_{\theta_h^*}(s_h^\dagger,a_h)\right\|_1 \\
=&\int_{\mathcal{S}^\dagger\times\mathcal{A}} d\mu_{\boldsymbol{\theta}^*,h}^{\boldsymbol{\pi}}(s_h^\dagger,a_h)\int_{\mathcal{S}^\dagger}\left|\mathbb{T}_{\theta_h}(s_{h+1}^\dagger|s_h^\dagger,a_h) - \mathbb{T}_{\theta_h^*}(s_{h+1}^\dagger|s_h^\dagger,a_h)\right| ds_{h+1}^\dagger \\
\leq&\int_{\mathcal{S}^\dagger\times\mathcal{A}\times\mathcal{S}^\dagger}\left|\mu_{\boldsymbol{\theta}^*,h}^{\boldsymbol{\pi}}(s_h^\dagger,a_h,s_{h+1}^\dagger) - \mu_{\boldsymbol{\theta},h}^{\boldsymbol{\pi}}(s_h^\dagger,a_h,s_{h+1}^\dagger)\right| + \left|\mu_{\boldsymbol{\theta},h}^{\boldsymbol{\pi}}(s_h^\dagger,a_h) - \mu_{\boldsymbol{\theta}_h^*}^{\boldsymbol{\pi}}(s_h^\dagger,a_h)\right|\mathbb{T}_{\theta_h}(s_{h+1}^\dagger|s_h^\dagger,a_h)ds_h^\dagger da_h ds_{h+1}^\dagger\,.
\end{aligned}$$

By the fact that for any state action pair $(s_h^\dagger, a_h)$, we have

$$\begin{aligned}
&\int_{\mathcal{S}^\dagger\times\mathcal{A}} \left|\mu^{\boldsymbol{\pi}}_{\boldsymbol{\theta},h}(s_h^\dagger, a_h) - \mu^{\boldsymbol{\pi}}_{\boldsymbol{\theta}^*_h}(s_h^\dagger, a_h)\right| ds_h^\dagger da_h \\
=&\int_{s_h^\dagger, a_h} \int_{\{\boldsymbol{\tau}:\tau_h=(s_h^\dagger,a_h)\}} \left|\mu^{\boldsymbol{\pi}}_{\boldsymbol{\theta}}(\boldsymbol{\tau}) - \mu^{\boldsymbol{\pi}}_{\boldsymbol{\theta}^*}(\boldsymbol{\tau})\right| d\boldsymbol{\tau} \\
\leq&\int_{\boldsymbol{\tau}} \left|\mu^{\boldsymbol{\pi}}_{\boldsymbol{\theta}}(\boldsymbol{\tau}) - \mu^{\boldsymbol{\pi}}_{\boldsymbol{\theta}^*}(\boldsymbol{\tau})\right| \\
=&\left\|\mu^{\pi}_{\boldsymbol{\theta}^*} - \mu^{\pi}_{\boldsymbol{\theta}}\right\|_1 .
\end{aligned}$$

We can apply a similar method to give

$$\int_{\mathcal{S}^\dagger\times\mathcal{A}\times\mathcal{S}^\dagger} \left|\mu^{\boldsymbol{\pi}}_{\boldsymbol{\theta}^*,h}(s_h^\dagger, a_h, s_{h+1}^\dagger) - \mu^{\boldsymbol{\pi}}_{\boldsymbol{\theta},h}(s_h^\dagger, a_h, s_{h+1}^\dagger)\right| ds_h^\dagger da_h ds_{h+1}^\dagger \leq \left\|\mu^{\pi}_{\boldsymbol{\theta}^*} - \mu^{\pi}_{\boldsymbol{\theta}}\right\|_1 .$$

Thus we can get

$$\sum_{h=1}^{H} \mathbb{E}_{(s_h^\dagger, a_h)\sim\mu^{\boldsymbol{\pi}}_{\boldsymbol{\theta}^*,h}} \left\|\mathbb{T}_{\theta_h}(s_h^\dagger, a_h) - \mathbb{T}_{\theta_h^*}(s_h^\dagger, a_h)\right\|_1 \leq 2H \left\|\mu^{\pi}_{\boldsymbol{\theta}^*} - \mu^{\pi}_{\boldsymbol{\theta}}\right\|_1 .$$

At last, we need to prove the equivalence between augmented transition difference and original transition difference.

$$\begin{aligned}
&\mathbb{E}_{(s_h^\dagger, a_h)\sim\mu^{\pi}_{\boldsymbol{\theta}^*,h}} \left[\left\|\mathbb{T}_{\theta_h^*}(s_h^\dagger, a_h) - \mathbb{T}_{\theta_h}(s_h^\dagger, a_h)\right\|_1\right] \\
=&\int_{\mathcal{S}^\dagger\times\mathcal{A}} d\mu^{\pi}_{\theta^*,h}(s_h^\dagger, a_h) \int_{\mathcal{S}} \int_{[y_h,1]} \left|\mathbb{P}_{\theta_h^*}(s_{h+1}|s_h, a_h) - \mathbb{P}_{\theta_h}(s_{h+1}|s_h, a_h)\right| \mathbb{R}_h(y_{h+1} - y_h|s_h, a_h) dy_{h+1} ds_{h+1} \\
=&\int_{\mathcal{S}\times\mathcal{A}} \int_0^1 \left\|\mathbb{P}_{\theta_h^*}(s_h, a_h) - \mathbb{P}_{\theta_h}(s_h, a_h)\right\|_1 d\mu^{\boldsymbol{\pi}}_{\boldsymbol{\theta}^*,h}((s_h, y_h), a_h) \int_{y_h}^1 \mathbb{R}_h(y_{h+1} - y_h|s_h, a_h) dy_{h+1} \\
=&\int_{\mathcal{S}\times\mathcal{A}} \left\|\mathbb{P}_{\theta_h^*}(s_h, a_h) - \mathbb{P}_{\theta_h}(s_h, a_h)\right\|_1 d\nu^{\boldsymbol{\pi}}_{\boldsymbol{\theta}^*,h}(s_h, a_h) \\
=&\int_{\mathcal{S}\times\mathcal{A}} \left\|\mathbb{P}_{\theta_h^*}(s_h, a_h) - \mathbb{P}_{\theta_h'}(s_h, a_h)\right\|_1 d\nu^{\pi}_{\theta^*,h}(s_h, a_h) \\
=&\mathbb{E}_{(s_h, a_h)\sim\nu^{\boldsymbol{\pi}}_{\boldsymbol{\theta}^*,h}} \left[\left\|\mathbb{P}_{\theta_h^*}(s_h, a_h) - \mathbb{P}_{\theta_h'}(s_h, a_h)\right\|_1\right] .
\end{aligned}$$

where the first equality holds by definition, the second equality is due to the Fubini Theorem, and the third equality holds by the decomposition of the probability measures of augmented MDP. □

D.3. Concentration Condition of MLE Approach

In this section, we prove the concentration condition of the MLE approach. The key idea is utilizing the property of MLE which is also studied by Agarwal et al. (2020); Liu et al. (2022; 2023).

We define the ℓ_1-norm difference on the model set $\boldsymbol{\Theta}$. For any $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2 \in \boldsymbol{\Theta}$, we define $\|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2\|_1$ as:

$$\begin{aligned}
\|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2\|_1 :=& \sup_{h\in[H],(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} \|\mathbb{P}_{\theta_h^1}(s_h, a_h) - \mathbb{P}_{\theta_h^2}(s_h, a_h)\|_1 \\
=& \sup_{h\in[H],(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} \int_{\mathcal{S}} \left|\mathbb{P}_{\theta_h^1}(s_{h+1}|s_h, a_h) - \mathbb{P}_{\theta_h^2}(s_{h+1}|s_h, a_h)\right| ds_{h+1} .
\end{aligned}$$

Then we denote $\boldsymbol{\Theta}^\downarrow$ as the lower bracket function set of $\boldsymbol{\Theta}$ such that for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have $\boldsymbol{\theta}^\downarrow \in \boldsymbol{\Theta}^\downarrow$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}^\downarrow\|_1 \leq \epsilon$. We have $\left|\boldsymbol{\Theta}^\downarrow\right| \leq \mathcal{N}_{[\cdot]}(\boldsymbol{\Theta}, \epsilon, \|\cdot\|_1)$.

Lemma D.3 (Likelihood difference)**.** *Consider the probability constant $\delta \in (0, 1]$. For **all** $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ $h \in [H]$, and $k \in [K]$, we have:*

$$\sum_{i=1}^{k} \log\left[\frac{\mathbb{P}_{\theta_h}(s_{i,h+1}|s_{i,h}, a_{i,h})}{\mathbb{P}_{\theta_h^*}(s_{i,h+1}|s_{i,h}, a_{i,h})}\right] \leq \log\left(eK\mathcal{N}_{[\cdot]}(\boldsymbol{\Theta}, \epsilon, \|\cdot\|_1)/\delta\right)$$

holds for fixed h with probability at least $1 - \delta$.

Proof. This proof is standard (Geer, 2000; Liu et al., 2022; 2023). Consider the lower bracket set $\boldsymbol{\Theta}^{\downarrow}$. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we can find a $\boldsymbol{\theta}^{\downarrow} \in \boldsymbol{\Theta}^{\downarrow}$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\downarrow}\|_1 \leq \epsilon$. Moreover, for every $h \in [H]$, we denote $s_{i,h}, a_{i_h}$ as the corresponding state action in trajectory $\tau_{i,h}$, and we have:

$$
\begin{aligned}
&\mathbb{E}_{s_{i,h+1}, i\in[k]}\left[\exp\left(\sum_{i=1}^{k}\log\left[\frac{\mathbb{P}_{\theta_h^{\downarrow}}(s_{i,h+1}|s_{i,h},a_{i,h})}{\mathbb{P}_{\theta_h^{*}}(s_{i,h+1}|s_{i,h},a_{i,h})}\right]\right)\right]\\
=&\mathbb{E}_{s_{i,h+1}, i\in[k]}\left[\exp\left(\sum_{i=1}^{k-1}\log\left[\frac{\mathbb{P}_{\theta_h^{\downarrow}}(s_{i,h+1}|s_{i,h},a_{i,h})}{\mathbb{P}_{\theta_h^{*}}(s_{i,h+1}|s_{i,h},a_{i,h})}\right]\right)\cdot\mathbb{E}_{s_{k,h+1}}\left[\frac{\mathbb{P}_{\theta_h^{\downarrow}}(s_{k,h+1}|s_{k,h},a_{k,h})}{\mathbb{P}_{\theta_h^{*}}(s_{k,h+1}|s_{k,h},a_{k,h})}\right]\right]
\end{aligned}
$$

By the definition of the expectation operator, we have

$$
\begin{aligned}
&\mathbb{E}_{s_{i,h+1}, i\in[k]}\left[\exp\left(\sum_{i=1}^{k-1}\log\left[\frac{\mathbb{P}_{\theta_h^{\downarrow}}(s_{i,h+1}|s_{i,h},a_{i,h})}{\mathbb{P}_{\theta_h^{*}}(s_{i,h+1}|s_{i,h},a_{i,h})}\right]\right)\cdot\mathbb{E}_{s_{k,h+1}}\left[\frac{\mathbb{P}_{\theta_h^{\downarrow}}(s_{k,h+1}|s_{k,h},a_{k,h})}{\mathbb{P}_{\theta_h^{*}}(s_{k,h+1}|s_{k,h},a_{k,h})}\right]\right]\\
=&\mathbb{E}_{s_{i,h+1}, i\in[k]}\left[\exp\left(\sum_{i=1}^{k-1}\log\left[\frac{\mathbb{P}_{\theta_h^{\downarrow}}(s_{i,h+1}|s_{i,h},a_{i,h})}{\mathbb{P}_{\theta_h^{*}}(s_{i,h+1}|s_{i,h},a_{i,h})}\right]\right)\cdot\int_{\mathcal{S}}\left[\frac{\mathbb{P}_{\theta_h^{\downarrow}}(s_{k,h+1}|s_{k,h},a_{k,h})}{\mathbb{P}_{\theta_h^{*}}(s_{k,h+1}|s_{k,h},a_{k,h})}\right]d\mathbb{P}_{\theta_h^{*}}(s_{k,h+1}|s_{k,h},a_{k,h})\right]\\
=&\mathbb{E}_{s_{i,h+1}, i\in[k]}\left[\exp\left(\sum_{i=1}^{k-1}\log\left[\frac{\mathbb{P}_{\theta_h^{\downarrow}}(s_{i,h+1}|s_{i,h},a_{i,h})}{\mathbb{P}_{\theta_h^{*}}(s_{i,h+1}|s_{i,h},a_{i,h})}\right]\right)\cdot\int_{\mathcal{S}}\mathbb{P}_{\theta_h^{\downarrow}}(s_{k,h+1}|s_{k,h},a_{k,h})ds_{k,h+1}\right]\\
\leq&\mathbb{E}_{s_{i,h+1}, i\in[k]}\left[\exp\left(\sum_{i=1}^{k-1}\log\left[\frac{\mathbb{P}_{\theta_h^{\downarrow}}(s_{i,h+1}|s_{i,h},a_{i,h})}{\mathbb{P}_{\theta_h^{*}}(s_{i,h+1}|s_{i,h},a_{i,h})}\right]\right)\cdot(\|\mathbb{P}_{\theta_h}(s_{k,h},a_{k,h})\|_1+\|\mathbb{P}_{\theta_h}(s_{k,h},a_{k,h})-\mathbb{P}_{\theta_h^{\downarrow}}(s_{k,h},a_{k,h})\|_1)\right]\\
\leq&\mathbb{E}_{s_{i,h+1}, i\in[k]}\left[\exp\left(\sum_{i=1}^{k-1}\log\left[\frac{\mathbb{P}_{\theta_h^{\downarrow}}(s_{i,h+1}|s_{i,h},a_{i,h})}{\mathbb{P}_{\theta_h^{*}}(s_{i,h+1}|s_{i,h},a_{i,h})}\right]\right)\cdot(1+\epsilon)\right]\\
\leq&\cdots\\
\leq&(1+\epsilon)^k\,.
\end{aligned}
$$

where the first inequality holds by the triangle inequality, and the second inequality holds by the definition of $\theta_h^{\downarrow}$. Moreover, since $\epsilon < 1/K$, we have $(1+\epsilon)^k < e$. Therefore, by Markov's inequality, we have that

$$
\mathbb{P}\left[\sum_{i=1}^{k}\log\left[\frac{\mathbb{P}_{\theta_h^{\downarrow}}(s_{i,h+1}|s_{i,h},a_{i,h})}{\mathbb{P}_{\theta_h^{*}}(s_{i,h+1}|s_{i,h},a_{i,h})}\right]\geq\log(1/\delta')\right]\leq e\delta'\,.
$$

holds for $\delta' \in (0,1]$. Thus taking a union bound for all $\boldsymbol{\theta}^{\downarrow} \in \boldsymbol{\Theta}^{\downarrow}$, $h \in [H]$ and $k \in [K]$, we have with probability at least $1-\delta$,

$$
\sum_{i=1}^{k}\log\left[\frac{\mathbb{P}_{\theta_h^{\downarrow}}(s_{i,h+1}|s_{i,h},a_{i,h})}{\mathbb{P}_{\theta_h^{*}}(s_{i,h+1}|s_{i,h},a_{i,h})}\right]\leq\log(eHK|\boldsymbol{\Theta}^{\downarrow}|/\delta)\,.
$$

Since $|\boldsymbol{\Theta}^{\downarrow}| = \mathcal{N}_{[\cdot]}(\boldsymbol{\Theta}, \epsilon, \|\cdot\|_1)$ and $\mathbb{P}_{\boldsymbol{\theta}^{\downarrow}} \leq \mathbb{P}_{\boldsymbol{\theta}}$, we have

$$
\sum_{i=1}^{k}\log\left[\frac{\mathbb{P}_{\theta_h}(s_{i,h+1}|s_{i,h},a_{i,h})}{\mathbb{P}_{\theta_h^{*}}(s_{i,h+1}|s_{i,h},a_{i,h})}\right]\leq\log(eHK\mathcal{N}_{[\cdot]}(\boldsymbol{\Theta},\epsilon,\|\cdot\|_1)/\delta)
$$

holds for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $h \in [H]$ and $k \in [K]$. □

Lemma D.4 (MLE concentration)**.** *The estimation function* `M-Est-MLE` *satisfying the Condition 6.2, i.e., for $\delta \in (0,1]$, we have with probability at least $1-\delta$, $\boldsymbol{\theta}^* \in \widehat{\boldsymbol{\Theta}}_k$ for every $k \in [K]$.*

Proof. Apply Lemma D.3, we have for every $k \in [K], h \in [H]$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$
\sum_{i=1}^{k}\mathbb{P}_{\theta_h}(s_{i,h+1}|s_{i,h},a_{i,h})\leq\sum_{i=1}^{k}\mathbb{P}_{\theta_h^{*}}(s_{i,h+1}|s_{i,h},a_{i,h})+\log(eHK\mathcal{N}_{[\cdot]}(\boldsymbol{\Theta},\epsilon,\|\cdot\|_1)/\delta)
$$

with probability at least $1-\delta$. Recall that $\beta^{\mathrm{MLE}} = H\log(eHK\mathcal{N}_{[\cdot]}(\boldsymbol{\Theta},\epsilon,\|\cdot\|_1)/\delta)$. Summing the both sides of the inequality over $h\in[H]$ directly gives the result. □

D.4. Total Variance Distance of Transition Models

As a direct result of Lemma H.1, we have the following bound for model-based MLE estimation, which is also presented in Proposition 14 in Liu et al. (2022) and Proposition B.2 in Liu et al. (2023).

Lemma D.5 (TV distance and likelihood distance)**.**

$$\sum_{i=1}^{k-1}\mathbb{E}_{(s_h,a_h)\sim\mu^{\boldsymbol{\pi}^i}_{\boldsymbol{\theta}^*,h}}\left[\mathrm{TV}\left(\mathbb{P}_{\theta_h}(s_h,a_h)||\mathbb{P}_{\theta^*_h}(s_h,a_h)\right)^2\right]\le\mathcal{O}\left(\sum_{i=1}^{k-1}\log\left[\frac{\mathbb{P}_{\theta^*_h}(s_{i,h+1}|s_{i,h},a_{i,h})}{\mathbb{P}_{\theta_h}(s_{i,h+1}|s_{i,h},a_{i,h})}\right]+\beta\right).$$

Finally, we can bound the Total Variation (TV) distance of the transition models.

Lemma D.6 (TV distance)**.** *For $\delta\in(0,1]$ and any $\boldsymbol{\theta}_k\in\widehat{\boldsymbol{\Theta}}_k$, we have the following concentration on their TV-distance with :*

$$\sum_{i=1}^{k-1}\sum_{h=1}^{H}\mathbb{E}_{(s_h,a_h)\sim\mu^{\boldsymbol{\pi}^i}_{\boldsymbol{\theta}^*,h}}\left[\mathrm{TV}\left(\mathbb{P}_{\theta_{k,h}}(s_h,a_h)||\mathbb{P}_{\theta^*_h}(s_h,a_h)\right)^2\right]\le\mathcal{O}(\beta^{\mathrm{MLE}}).$$

Proof. For any $\boldsymbol{\theta}_k\in\widehat{\boldsymbol{\Theta}}^{\mathrm{MLE}}_k$, we have that:

$$\sum_{i=1}^{k-1}\sum_{h=1}^{H}\log\left[\mathbb{P}_{\theta_{k,h}}(s_{i,h+1}|s_{i,h},a_{i,h})\right]\ge\sum_{i=1}^{k-1}\sum_{h=1}^{H}\log\left[\mathbb{P}_{\theta^{\mathrm{MLE}}_{k,h}}(s_{i,h+1}|s_{i,h},a_{i,h})\right]-\beta^{\mathrm{MLE}}.$$

From the definition of $\theta^{\mathrm{MLE}}_{k,h}$, we have:

$$\sum_{i=1}^{k-1}\sum_{h=1}^{H}\log\left[\mathbb{P}_{\theta^*_h}(s_{i,h+1}|s_{i,h},a_{i,h})\right]\le\sum_{i=1}^{k-1}\sum_{h=1}^{H}\log\left[\mathbb{P}_{\theta^{\mathrm{MLE}}_{k,h}}(s_{i,h+1}|s_{i,h},a_{i,h})\right].$$

Thus, we have:

$$\sum_{i=1}^{k-1}\sum_{h=1}^{H}\log\left[\mathbb{P}_{\theta^*_h}(s_{i,h+1}|s_{i,h},a_{i,h})\right]\le\sum_{i=1}^{k-1}\sum_{h=1}^{H}\log\left[\mathbb{P}_{\theta_{k,h}}(s_{i,h+1}|s_{i,h},a_{i,h})\right]+\beta^{\mathrm{MLE}}.$$

Using Lemma D.5, we have:

$$\begin{aligned}&\sum_{i=1}^{k-1}\sum_{h=1}^{H}\mathbb{E}_{(s_h,a_h)\sim\mu^{\boldsymbol{\pi}^i}_{\boldsymbol{\theta}^*,h}}\left[\mathrm{TV}\left(\mathbb{P}_{\theta_{k,h}}(s_h,a_h)||\mathbb{P}_{\theta^*_h}(s_h,a_h)\right)^2\right]\\ \le&\mathcal{O}\left(\sum_{i=1}^{k-1}\sum_{h=1}^{H}\log\left[\frac{\mathbb{P}_{\theta^*_h}(s_{i,h+1}|s_{i,h},a_{i,h})}{\mathbb{P}_{\theta_{k,h}}(s_{i,h+1}|s_{i,h},a_{i,h})}\right]+\beta^{\mathrm{MLE}}\right)\le\mathcal{O}(\beta^{\mathrm{MLE}}).\end{aligned}$$

□

D.5. Elliptical Potential Condition of MLE Approach

First, we define the low witness rank, which is the structural complexity measure for this setting. We remark this is the Q-type witness rank and we provide a separate proof for V-type witness rank in Section D.6.

Definition D.7 (Q-type low witness rank (Definition 6.6 of Liu et al. (2023)))**.** We say the model class satisfies (d,α,B) witness rank condition, if there exists mappings: $\{f_h\}$ and $\{g_h\}$ from $\boldsymbol{\Theta}$ to $\mathbb{R}^d$, such that for any $h\in[H]$:

$$\begin{aligned}&\mathbb{E}_{(s_h,a_h)\sim\nu^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}_{\boldsymbol{\theta}^*,h}}\left[\mathrm{TV}\left(\mathbb{P}_{\boldsymbol{\theta}'_h}(s_h,a_h)||\mathbb{P}_{\boldsymbol{\theta}^*_h}(s_h,a_h)\right)\right]\le\langle f_h(\boldsymbol{\theta}),g_h(\boldsymbol{\theta}')\rangle\\ &\mathbb{E}_{(s_h,a_h)\sim\nu^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}_{\boldsymbol{\theta}^*,h}}\left[\mathrm{TV}\left(\mathbb{P}_{\boldsymbol{\theta}'_h}(s_h,a_h)||\mathbb{P}_{\boldsymbol{\theta}^*_h}(s_h,a_h)\right)\right]\ge\alpha^{-1}\langle f_h(\boldsymbol{\theta}),g_h(\boldsymbol{\theta}')\rangle\\ &\|f_h(\boldsymbol{\theta})\|_1\|g_h(\boldsymbol{\theta}')\|_\infty\le B.\end{aligned}$$

As a special case, we can show that the factored MDPs (Sun et al., 2019) have low witness rank.

Definition D.8 (Factored MDP)**.** In factored MDPs the states admit a factored structure. Each state contains m factors $(s[1], s[2], \cdots, s[m]) \in \mathcal{X}^m$. Each factor $i \in [m]$ has a parent set $pa_i \subset [m]$, with respect to which the transitions admit a factored form:

$$\mathbb{P}_h(s_{h+1}|s_h, a_h) = \prod_{i=1}^{m} \mathbb{P}_h^i(s_{h+1}[i]|s_h[pa_i], a_h)\,.$$

The following proposition establishes the low witness rank property for factored MDPs, which comes directly from Proposition 6.8 of Liu et al. (2023).

Proposition D.9. *Let $\boldsymbol{\Theta}$ denote all the factored MDPs with the same factorization structure, then $\boldsymbol{\Theta}$ satisfies low witness rank with $d = |\mathcal{A}| \sum_{i=1}^m |\mathcal{X}|^{|pa_i|}$, $\alpha = m$, $B = \sum_{i=1}^m |\mathcal{X}|^{|pa_i|}$.*

Proof of Condition B.2. From Lemma D.6 and using Cauchy inequality, we have:

$$\begin{aligned}
\sum_{h=1}^{H}\sum_{i=1}^{k-1} |\langle f_h(\boldsymbol{\theta}_i), g_h(\boldsymbol{\theta}_k)\rangle|^2 \leq& \alpha^2 \sum_{h=1}^{H}\sum_{i=1}^{k-1} \left(\mathbb{E}_{(s_h,a_h)\sim\mu^i_{\boldsymbol{\theta}^*,h}}\left[\mathrm{TV}\left(\mathbb{P}_{\theta_{k,h}}(s_h,a_h)||\mathbb{P}_{\theta^*_h}(s_h,a_h)\right)\right]\right)^2 \\
\leq& \alpha^2 \sum_{h=1}^{H}\sum_{i=1}^{k-1} \mathbb{E}_{(s_h,a_h)\sim\mu^i_{\boldsymbol{\theta}^*,h}}\left[\mathrm{TV}\left(\mathbb{P}_{\theta_{k,h}}(s_h,a_h)||\mathbb{P}_{\theta^*_h}(s_h,a_h)\right)^2\right] \\
\leq& \mathcal{O}\left(\alpha^2\beta^{\mathrm{MLE}}\right)\,.
\end{aligned}$$

Thus, we have by standard elliptical potential arguments (eg:Theorem 6.4 in Liu et al. (2023)):

$$\begin{aligned}
\sum_{h=1}^{H}\sum_{i=1}^{k} \mathbb{E}_{(s_h,a_h)\sim\mu^{\pi^i}_{\boldsymbol{\theta}^*,h}}\left[\mathrm{TV}\left(\mathbb{P}_{\theta_{k,h}}(s_h,a_h)||\mathbb{P}_{\theta^*_h}(s_h,a_h)\right)\right] \leq& \sum_{h=1}^{H}\sum_{i=1}^{k}\langle f_h(\boldsymbol{\theta}_i), g_h(\boldsymbol{\theta}_i)\rangle \\
\leq& \mathcal{O}\left(dB + \alpha\sqrt{d\beta^{\mathrm{MLE}}k}\right)\,.
\end{aligned}$$

Thus, from Lemma D.1 and Lemma D.2, we have:

$$\begin{aligned}
\sum_{k=1}^{K}\left\|F_{Z^{\boldsymbol{\pi}^k}_{\boldsymbol{\theta}_k}} - F_{Z^{\boldsymbol{\pi}_k}_{\boldsymbol{\theta}^*}}\right\|_\infty \leq& 2\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim\mu^k_{\boldsymbol{\theta}^*,h}}\left[\mathrm{TV}\left(\mathbb{P}_{\theta_{k,h}}(s_h,a_h)||\mathbb{P}_{\theta^*_h}(s_h,a_h)\right)\right] \\
\leq& \mathcal{O}\left(dB + \alpha\sqrt{d\beta^{\mathrm{MLE}}K}\right)\,.
\end{aligned}$$

Thus, using the L_∞ Lipschitz of ρ, we have:

$$\sum_{k=1}^{K} \rho(Z^{\boldsymbol{\pi}^k}_{\boldsymbol{\theta}_k}) - \rho(Z^{\boldsymbol{\pi}^k}_{\boldsymbol{\theta}^*}) \leq \mathcal{O}\left(L_\infty\left(dB + \alpha\sqrt{d\beta^{\mathrm{MLE}}K}\right)\right)\,.$$

□

D.6. Proof for Low Rank MDP

Extend our analysis above, we further provide an algorithm and regret analysis for low-rank MDPs, a special case of V-type witness rank, which also follows the MLE procedure above. The main difference is that we need to construct the mappings $f_h(\boldsymbol{\theta})$ and $g_h(\boldsymbol{\theta})$ in a novel manner to capture the low V-type witness rank of low rank MDPs in the augmented MDP.

First we define low rank MDPs as below (Agarwal et al., 2020; Uehara et al., 2021):

Definition D.10 (Low Rank MDP)**.** The transition kernel $\mathbb{P}_{\theta^*_h}(s_{h+1}|s_h, a_h)$ admits a low rank structure, i,e, there exists two sets of mappings $\boldsymbol{\phi}^* : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ and $\boldsymbol{\psi}^* : \mathcal{S} \to \mathbb{R}^d$, such that:

$$\mathbb{P}_{\theta^*_h}(s_{h+1}|s_h, a_h) = \langle \phi^*_h(s_h, a_h), \psi^*_h(s_{h+1})\rangle\,.$$

We have $\|\phi_h^*(s_h, a_h)\| \leq 1$ and $\left\|\int_{s_h} \psi_h^*(s_h) g(s_h)\right\| \leq \sqrt{d} \|g\|_\infty$ for all $h \in [H]$. Also, assume that we have access to two embedding classes Φ and Ψ such that $\boldsymbol{\phi}^* \in \Phi$ and $\boldsymbol{\psi}^* \in \Psi$.

The model class $\boldsymbol{\Theta}$ consists of all the transition kernels with the low rank structure defined by the inner-product of the embedding in Φ and Ψ, with $|\boldsymbol{\Theta}| = |\Psi||\Phi|$. Define the exploratory policy class for a policy $\boldsymbol{\pi_\theta}$ as $\Pi_{\exp}(\boldsymbol{\pi_\theta}) = \{\tilde{\pi}_{\theta,h} : \pi_\theta[1:h-1] \circ \mathcal{U}[h:H]\}_{h=1}^H$, where $\mathcal{U}$ is the uniform policy. $\tilde{\pi}_{\theta,h}$ is defined as following $\pi_{\boldsymbol{\theta}}$ for the first $h-1$ steps then taking uniform actions. We have $|\Pi_{\exp}| = H$. Define the two sets of mappings used to construct low witness rank for low rank MDPs as:

$$
\begin{aligned}
f_h(\boldsymbol{\theta}) &= \int_{s_{h-1}, a_{h-1}} \nu_{\boldsymbol{\theta}^*, h-1}^{\boldsymbol{\pi}_\theta}(s_{h-1}, a_{h-1}) \phi_{h-1}^*(s_{h-1}, a_{h-1}) \\
g_h(\boldsymbol{\theta}) &= \int_{s_h, a_h} \psi_h^*(s_h) \mathcal{U}(a_h) \left\|\mathbb{P}_{\theta_h^*}(s_h, a_h) - \mathbb{P}_{\theta_h}(s_h, a_h)\right\|_1 ,
\end{aligned}
$$

where $\boldsymbol{\pi_\theta}$ is the optimal risk-sensitive policy given model θ. We have that for any $\tilde{\pi}_{\theta,h} \in \Pi_{\exp}(\boldsymbol{\pi_\theta})$:

$$
\begin{aligned}
&2H \left\|\mu_{\boldsymbol{\theta}^*}^{\tilde{\pi}_{\theta,h}} - \mu_{\boldsymbol{\theta}'}^{\tilde{\pi}_{\theta,h}}\right\|_1 \\
\geq &\int_{s_h^\dagger, a_h} \mu_{\theta^*, h}^{\tilde{\pi}_{\theta,h}}(s_h^\dagger, a_h) \left\|\mathbb{T}_{\theta_h'}(s_h^\dagger, a_h) - \mathbb{T}_{\theta_h^*}(s_h^\dagger, a_h)\right\|_1 \\
= &\int_{s_h^\dagger, a_h} \mu_{\boldsymbol{\theta}^*, h}^{\tilde{\pi}_{\theta,h}}(s_h^\dagger, a_h) \left\|\mathbb{P}_{\theta_h^*}(s_h, a_h) - \mathbb{P}_{\theta_h'}(s_h, a_h)\right\|_1 \\
= &\int_{s_{h-1}^\dagger, a_{h-1}, s_h, a_h, y_h} \mu_{\boldsymbol{\theta}^*, h-1}^{\pi_\theta}(s_{h-1}^\dagger, a_{h-1}) \mathbb{R}(y_h - y_{h-1}|s_{h-1}, a_{h-1}) \mathbb{P}_{\theta_{h-1}^*}(s_h|s_{h-1}, a_{h-1}) \mathcal{U}(a_h) \qquad (3) \\
&\cdot \left\|\mathbb{P}_{\theta_h^*}(s_h, a_h) - \mathbb{P}_{\theta_h'}(s_h, a_h)\right\|_1 \\
= &\int_{s_{h-1}^\dagger, a_{h-1}} \mu_{\boldsymbol{\theta}^*, h-1}^{\pi_\theta}(s_{h-1}^\dagger, a_{h-1}) \phi_{h-1}^*(s_{h-1}, a_{h-1}) \int_{y_h} \mathbb{R}(y_h - y_{h-1}|s_{h-1}, a_{h-1}) \qquad (4) \\
&\cdot \int_{s_h, a_h} \psi_{h-1}^*(s_h) \mathcal{U}(a_h) \left\|\mathbb{P}_{\theta_h^*}(s_h, a_h) - \mathbb{P}_{\theta_h'}(s_h, a_h)\right\|_1 \\
= &\langle f_h(\boldsymbol{\theta}), g_h(\boldsymbol{\theta}')\rangle , \qquad (5)
\end{aligned}
$$

where the first inequality is from the right hand side of simulation lemma (Lemma D.2).

On the other hand, we have:

$$
\begin{aligned}
&\|\mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi_\theta}} - \mu_{\boldsymbol{\theta}}^{\boldsymbol{\pi_\theta}}\|_1 \\
\leq &\sum_{h=1}^H \int_{s_h^\dagger, a_h} \mu_{\boldsymbol{\theta}^*, h}^{\boldsymbol{\pi}_\theta}(s_h^\dagger, a_h) \left\|\mathbb{P}_{\theta_h^*}(s_h, a_h) - \mathbb{P}_{\theta_h}(s_h, a_h)\right\|_1 \\
= &\sum_{h=1}^H \int_{s_h^\dagger, a_h} \mu_{\boldsymbol{\theta}^*, h}^{\boldsymbol{\pi_\theta}}(s_h^\dagger) \pi_{\theta,h}(a_h|s_h^\dagger) \left\|\mathbb{P}_{\theta_h^*}(s_h, a_h) - \mathbb{P}_{\theta_h}(s_h, a_h)\right\|_1
\end{aligned} \qquad (6)
$$

Moreover,

$$
\begin{aligned}
\|\mu^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}_{\boldsymbol{\theta}^*} - \mu^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}_{\boldsymbol{\theta}}\|_1 \leq & A \sum_{h=1}^{H} \int_{s^\dagger_h, a_h} \mu^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}_{\boldsymbol{\theta}^*, h}(s^\dagger_h)\mathcal{U}(a_h) \left\|\mathbb{P}_{\theta^*_h}(s_h, a_h) - \mathbb{P}_{\theta_h}(s_h, a_h)\right\|_1 \\
\leq & A \sum_{h=1}^{H} \int_{s^\dagger_{h-1}, a_{h-1}} \mu^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}_{\boldsymbol{\theta}^*, h-1}(s^\dagger_{h-1}, a_h) \int_{y_h} \mathbb{R}(y_h - y_{h-1}|s_{h-1}, a_{h-1}) \qquad (7) \\
& \cdot \int_{s_h, a_h} \mathbb{P}_{\theta^*_{h-1}}(s_h|s_{h-1}, a_{h-1})\mathcal{U}(a_h) \left\|\mathbb{P}_{\theta^*_h}(s_h, a_h) - \mathbb{P}_{\theta_h}(s_h, a_h)\right\|_1 \\
= & A \sum_{h=1}^{H} \int_{s^\dagger_{h-1}, a_{h-1}} \mu^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}_{\boldsymbol{\theta}^*, h-1}(s^\dagger_{h-1}, a_h)\phi^*_{h-1}(s_{h-1}, a_{h-1}) \qquad (8) \\
& \cdot \int_{s_h, a_h} \psi^*_{h-1}(s_h)\mathcal{U}(a_h) \left\|\mathbb{P}_{\theta^*_h}(s_h, a_h) - \mathbb{P}_{\theta_h}(s_h, a_h)\right\|_1 \\
= & A \sum_{h=1}^{H} \langle f_h(\boldsymbol{\theta}), g_h(\boldsymbol{\theta})\rangle \,. \qquad (9)
\end{aligned}
$$

We present a modified version of the algorithm as Algorithm 6, where we modify the data collection process such that in

Algorithm 6 `RS-DisRL-Low-Rank-MDP`$(\boldsymbol{\Theta}, \beta)$

Input: Model class $\boldsymbol{\Theta}$, confidence radius $\beta^{\mathrm{MLE}} = \log(|\boldsymbol{\Theta}|/\delta)$.
Init: $\boldsymbol{\Theta}_1 \leftarrow \boldsymbol{\Theta}$
for $k = 1$ **to** K **do**
Optimistic Planning: $(\boldsymbol{\pi}_{\widehat{\theta}_k}, \widehat{\boldsymbol{\theta}}_k) = \mathrm{argmax}_{\boldsymbol{\pi}\in\Pi^\dagger, \boldsymbol{\theta}\in\widehat{\boldsymbol{\Theta}}_k} \rho(Z^{\boldsymbol{\pi}}_{\boldsymbol{\theta}})$.
Execute and collect information: For every policy $\tilde{\boldsymbol{\pi}} \in \Pi_{\exp}(\boldsymbol{\pi}_{\widehat{\theta}_k})$ Execute policy $\tilde{\boldsymbol{\pi}}$, add the collected data $\boldsymbol{\tau} = \left\{(s^\dagger_h, a_h, s^\dagger_{h+1})\right\}_{h=1}^{H}$ into history $\mathcal{H}_k = \mathcal{H}_{k-1} \cup \{\boldsymbol{\tau}\}$.
Estimate the MLE solution:

$$\theta^{\mathrm{MLE}}_{k+1,h} = (\phi^{\mathrm{MLE}}_{k+1,h}, \psi^{\mathrm{MLE}}_{k+1,h}) = \underset{\phi\in\Phi, \psi\in\Psi}{\mathrm{argmax}} \sum_{\boldsymbol{\tau}\in\mathcal{H}_k} \log\langle\phi(s^{\boldsymbol{\tau}}_h, a^{\boldsymbol{\tau}}_h), \psi(s^{\boldsymbol{\tau}}_{h+1})\rangle\,, \;\; \forall h \in [H]\,,$$

where $s^{\boldsymbol{\tau}}_h, a^{\boldsymbol{\tau}}_h$ denotes the h step state action pair in trajectory $\boldsymbol{\tau}$.
Construct confidence set:

$$\widehat{\boldsymbol{\Theta}}_{k+1} = \left\{\phi \in \Psi, \psi \in \Psi : \sum_{\boldsymbol{\tau}\in\mathcal{H}_k} \sum_{h=1}^{H} \log\left(\frac{\langle\phi^{\mathrm{MLE}}_{k+1,h}(s^{\boldsymbol{\tau}}_h, a^{\boldsymbol{\tau}}_h), \psi^{\mathrm{MLE}}_{k+1,h}(s^{\boldsymbol{\tau}}_{h+1})\rangle}{\langle\phi_h(s^{\boldsymbol{\tau}}_h, a^{\boldsymbol{\tau}}_h), \psi_h(s^{\boldsymbol{\tau}}_{h+1})\rangle}\right) \leq \beta\right\}.$$

end for

each episode, instead of executing policy $\boldsymbol{\pi}_{\widehat{\boldsymbol{\theta}}_k}$, we execute all the policies $\tilde{\boldsymbol{\pi}}_{\widehat{\theta}_k,h} \in \Pi_{\exp}(\boldsymbol{\pi}_{\widehat{\boldsymbol{\theta}}_k})$. Notice that in each step we collect H trajectories by taking the combination of the optimistic policy and the uniform exploratory policy. Similar to the proof in the Q-type witness rank MLE concentration, we have $\boldsymbol{\theta}^* \in \widehat{\boldsymbol{\Theta}}_k$ for all $k \in [K]$ by choosing $\beta^{\mathrm{MLE}} = \log(|\boldsymbol{\Theta}|/\delta)$. Also, we can bound the sum of the square distance similar as Lemma D.6:

$$\sum_{i=1}^{k} \sum_{\tilde{\boldsymbol{\pi}}\in\Pi_{\exp}(\boldsymbol{\pi}_{\widehat{\theta}_i})} \sum_{h=1}^{H} \mathbb{E}_{\mu^{\tilde{\boldsymbol{\pi}}}_{\boldsymbol{\theta}^*,h}} \left[\left\|\mathbb{P}_{\widehat{\theta}_{k,h}}(s_h, a_h) - \mathbb{P}_{\theta^*_h}(s_h, a_h)\right\|_1^2\right] \leq \mathcal{O}(\beta^{\mathrm{MLE}})\,.$$

Thus, using Lemma D.2 we have that for any $\widehat{\boldsymbol{\theta}}_k \in \widehat{\boldsymbol{\Theta}}_k^{\mathrm{MLE}}$:

$$\begin{aligned}
&\sum_{i=1}^{k} \sum_{\tilde{\boldsymbol{\pi}} \in \Pi_{\exp}(\boldsymbol{\pi}_{\widehat{\boldsymbol{\theta}}_i})} \left\| \mu_{\boldsymbol{\theta}^*}^{\tilde{\boldsymbol{\pi}}} - \mu_{\widehat{\boldsymbol{\theta}}_k}^{\tilde{\boldsymbol{\pi}}} \right\|_1^2 \\
\leq &\sum_{i=1}^{k} \sum_{\tilde{\boldsymbol{\pi}} \in \Pi_{\exp}(\boldsymbol{\pi}_{\widehat{\boldsymbol{\theta}}_i})} \left(\sum_{h=1}^{H} \mathbb{E}_{\mu_{\boldsymbol{\theta}^*,h}^{\tilde{\boldsymbol{\pi}}}} \left[\left\| \mathbb{P}_{\widehat{\theta}_{k,h}}(s_h, a_h) - \mathbb{P}_{\theta_h^*}(s_h, a_h) \right\|_1 \right] \right)^2 \\
\leq &\mathcal{O}(\mathrm{poly}(H)\beta^{\mathrm{MLE}}) \,.
\end{aligned}$$

Thus we have from Equation 3:

$$\sum_{i=1}^{k} \sum_{h=1}^{H} \left(\langle f_h(\widehat{\boldsymbol{\theta}}_i), g_h(\widehat{\boldsymbol{\theta}}_k) \rangle \right)^2 \leq \mathcal{O}\left(\mathrm{poly}(H)\beta^{\mathrm{MLE}}\right) \,.$$

Since we have that $\|f_h(\boldsymbol{\theta})\|_1 \leq \sqrt{d}$ and $\|g_h(\boldsymbol{\theta})\|_\infty \leq \sqrt{d}$, we have by standard elliptical arguments and Equation 6:

$$\begin{aligned}
\sum_{i=1}^{k} \left\| \mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}_{\widehat{\boldsymbol{\theta}}_i}} - \mu_{\widehat{\boldsymbol{\theta}}_i}^{\boldsymbol{\pi}_{\widehat{\boldsymbol{\theta}}_i}} \right\|_1 \leq &A \sum_{i=1}^{k} \sum_{h=1}^{H} \langle f_h(\widehat{\boldsymbol{\theta}}_i), g_h(\widehat{\boldsymbol{\theta}}_i) \rangle \\
\leq &\mathcal{O}(\mathrm{poly}(H) A \sqrt{d\beta^{\mathrm{MLE}} k}) \,,
\end{aligned}$$

where we ignore the constant and low order terms in k. Thus we have:

$$\begin{aligned}
\sum_{k=1}^{K} \rho(Z_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}^*}) - \rho(Z_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}^k}) \leq &\sum_{k=1}^{K} \rho(Z_{\widehat{\boldsymbol{\theta}}_k}^{\boldsymbol{\pi}_{\widehat{\boldsymbol{\theta}}_k}}) - \rho(Z_{\widehat{\boldsymbol{\theta}}_k}^{\boldsymbol{\pi}_{\widehat{\boldsymbol{\theta}}_k}}) \\
\leq &L_\infty \sum_{k=1}^{K} \left\| F_{Z_{\widehat{\boldsymbol{\theta}}_k}^{\boldsymbol{\pi}_{\widehat{\boldsymbol{\theta}}_k}}} - F_{Z_{\widehat{\boldsymbol{\theta}}_k}^{\boldsymbol{\pi}_{\widehat{\boldsymbol{\theta}}_k}}} \right\|_\infty \\
\leq &L_\infty \sum_{k=1}^{K} \left\| \mu_{\boldsymbol{\theta}^*}^{\boldsymbol{\pi}_{\widehat{\boldsymbol{\theta}}_k}} - \mu_{\widehat{\boldsymbol{\theta}}_k}^{\boldsymbol{\pi}_{\widehat{\boldsymbol{\theta}}_k}} \right\|_1 \\
\leq &\mathcal{O}(L_\infty \mathrm{poly}(H) A \sqrt{d\beta K}) \,.
\end{aligned}$$

E. General Model-free Framework: Algorithm `RS-DisRL-V`

In this section we review the general model-free framework algorithm.

Algorithm 7 `RS-DisRL-V`

Input: Function class $\boldsymbol{\mathcal{Z}} = \mathcal{Z}_1 \times \mathcal{Z}_2 \cdots \mathcal{Z}_H$, confidence radius γ.
Initialize: $\widehat{\mathcal{Z}}_{1,\boldsymbol{\pi}} \leftarrow \mathcal{Z}$.
for $k = 1$ **to** K **do**
  $(\boldsymbol{\pi}_k, \widehat{Z}^k) = \mathrm{argmax}_{\boldsymbol{\pi} \in \Pi^\dagger, Z \in \widehat{\mathcal{Z}}_{k,\boldsymbol{\pi}}} \rho(Z_1)$. //Optimistic planning
  Execute policy $\boldsymbol{\pi}^k$, add the collected data $\boldsymbol{\tau}_k = \{(s_{k,h}, a_{k,h}, r_{k,h})\}_{h=1}^H$ and $\boldsymbol{\pi}^k$, $\widehat{\boldsymbol{\theta}}_k$ into history $\mathcal{H}_k = \mathcal{H}_{k-1} \cup \{(\boldsymbol{\tau}_k, \boldsymbol{\pi}^k, \widehat{Z}^k)\}$. //Data collection
  $\widehat{\mathcal{Z}}_{k+1,\boldsymbol{\pi}} = \texttt{V-Est}(\mathcal{H}_k, \boldsymbol{\mathcal{Z}}, \boldsymbol{\pi}, \gamma)$. //Confidence set construction
end for

We restate Condition 7.2 and Condition 7.3 as below:

Condition E.1. For all policy $\boldsymbol{\pi} \in \Pi^\dagger$, we have that the random variable representing the true return is in the version space:

$$Z^{\boldsymbol{\pi}} \in \widehat{\mathcal{Z}}_{k+1,\boldsymbol{\pi}} \,.$$

established with probability at least $1 - \delta$, $\delta \in (0, 1]$.

Condition E.2. For $0 < \delta \leq 1$, we have

$$\sum_{k=1}^{K} \left\| F_{\widehat{Z}^k} - F_{Z^{\boldsymbol{\pi}^k}} \right\|_\infty \leq \zeta(K, H, \mathrm{d}, \boldsymbol{\mathcal{Z}}, \Pi^\dagger, \delta, \gamma)\,,$$

holds with probability at least $1 - \delta$. Here d is some structural complexity measure of the problem.

Given these two conditions, our regret bound can be stated as follows:

Theorem E.3 (Full version of Theorem 7.4)**.** *Under the general value function approximation (Assumption 7.1) If the given estimation function* `V-Est` *satisfies Condition E.1 and Condition E.2, we have:*

$$\sum_{k=1}^{K} \rho(Z^{\boldsymbol{\pi}^*}) - \rho(Z^{\boldsymbol{\pi}^k}) \leq L_\infty(\rho)\zeta(K, H, \mathrm{d}, \boldsymbol{\mathcal{Z}}, \Pi^\dagger, \delta, \gamma)\,.$$

holds with probability at least $1 - \delta$*,* $\delta \in (0, 1]$.

Proof. Since the concentration condition E.1 holds, we have for any $\boldsymbol{\pi} \in \Pi^\dagger$ and $k \in K$,

$$\rho(Z^{\boldsymbol{\pi}}) \leq \max_{Z \in \widehat{\mathcal{Z}}_{k,\boldsymbol{\pi}}} \rho(Z) \leq \rho(\widehat{Z}^k)\ .$$

Thus, we have:

$$\sum_{k=1}^{K} \rho(Z^{\boldsymbol{\pi}^*}) - \rho(Z^{\boldsymbol{\pi}^k}) \leq \sum_{k=1}^{K} \rho(\widehat{Z}^k) - \rho(Z^{\boldsymbol{\pi}^k}) \leq L_\infty \sum_{k=1}^{K} \left\| F_{\widehat{Z}^k} - F_{Z^{\boldsymbol{\pi}^k}} \right\|_\infty \leq L_\infty \zeta(K, H, \mathrm{d}, \mathcal{Z}, \Pi^\dagger, \delta, \gamma)\,.$$

which gives this result. □

E.1. Policy Cover

Notice that our regret is defined via the optimal policy in the policy set, which is adopted in many model free valued-based scenarios, such as Xie et al. (2021); Wang et al. (2023b). The main reason why our algorithm can only operate in a given policy set is that the optimal risk-sensitive policy can not be computed via dynamical programming. In contrast, in the risk neutral setting we can always select the greedy policy and ensure that it is the optimistic policy given our estimation. We remark, however, that when specified to specific risk measures such as OCE (Xu et al., 2023), CVaR (Wang et al., 2023a) and ERM (Fei et al., 2020), where the optimal policy have a similar greedy property, we can ensure global optimality without the policy set.

In this section, we discuss the policy covering given a policy class $\Pi^\dagger$. For any policy $\pi \in \Pi^\dagger : \mathcal{S}^\dagger \to \Delta(\mathcal{A})$, we define its lower ϵ-bracket $\pi^\downarrow$ as $\pi^\downarrow \leq \pi$ and $\left\|\pi^\downarrow(\cdot|s^\dagger) - \pi(\cdot|s^\dagger)\right\|_1 \leq \epsilon$ for all $s^\dagger$. Since $\pi^\downarrow$ may not be a valid distribution, we define its normalized version as: $\underline{\pi}(a|s^\dagger) = \pi^\downarrow(a|s^\dagger)/\int_{\mathcal{A}} \pi^\downarrow(a'|s^\dagger)$. Since $1 - \epsilon \leq \int_{\mathcal{A}} \pi^\downarrow(a'|s^\dagger) \leq 1$, we have that $\underline{\pi}(a|s^\dagger) \leq (1 + 2\epsilon)\pi^\downarrow(a|s^\dagger)$. Its bracketing number is denoted as $\mathcal{N}_{[\cdot]}(\Pi^\dagger, \epsilon, \|\cdot\|_1)$.

Instances Consider a softmax policy set $\Pi(U, \tau)$ with temperature τ and utility function $u \in U : \mathcal{S}^\dagger \times \mathcal{A} \to \mathbb{R}$: $\pi(a|s^\dagger) = \frac{e^{\tau u(s^\dagger, a)}}{\int_{\mathcal{A}} e^{\tau u(s^\dagger, a')}}$. We consider a $\epsilon' = \frac{\epsilon}{8\tau}$ covering of U, such that for any u, there exists $\underline{u}$ in the covering $\underline{U}$ with $\|\underline{u} - u\|_\infty \leq \epsilon'$. and we can construct the lower bracket as:$\pi^\downarrow = \frac{e^{\tau(\underline{u}(s^\dagger, a) - \epsilon')}}{\int_{\mathcal{A}} e^{\tau(\underline{u}(s^\dagger, a') + \epsilon')}}$. Its normalized version is $\underline{\pi} = \frac{e^{\tau \underline{u}(s^\dagger, a)}}{\int_{\mathcal{A}} e^{\tau \underline{u}(s^\dagger, a')}}$. We can verify that: $\pi^\downarrow \leq \pi(a|s^\dagger)$ and $\left\|\pi^\downarrow(\cdot|s^\dagger) - \pi(\cdot|s^\dagger)\right\|_1 = 1 - \int_{\mathcal{A}} \pi^\downarrow(a|s^\dagger) = 1 - e^{-2\tau\epsilon'} \leq \epsilon$. So the bracketing number is the same as the $\frac{\epsilon}{8\tau}$ covering number of the utility function. $\mathcal{N}_{[\cdot]}(\Pi(U, \tau), \epsilon, \|\cdot\|_1) = \mathcal{N}_C(U, \frac{\epsilon}{8\tau}, \|\cdot\|_\infty)$.

Moreover, we have for any $s^\dagger$, $\underline{\pi}(a|s^\dagger) = e^{2\tau\epsilon'} \frac{e^{\tau(\underline{u}(s^\dagger, a) - \epsilon')}}{\int_{\mathcal{A}} e^{\tau(\underline{u}(s^\dagger, a') + \epsilon')}} \leq (1 + \epsilon)\pi(a|s^\dagger)$ and $\underline{\pi}(a|s^\dagger) = e^{-2\tau\epsilon'} \frac{e^{\tau(\underline{u}(s^\dagger, a) + \epsilon')}}{\int_{\mathcal{A}} e^{\tau(\underline{u}(s^\dagger, a') - \epsilon')}} \geq (1 + \epsilon)\pi(a|s^\dagger)$. Thus, $\left\|\pi(\cdot|s^\dagger) - \underline{\pi}(\cdot|s^\dagger)\right\| \leq \int_a \left|\underline{\pi}(a|s^\dagger) - \pi(a|s^\dagger)\right| \leq \int_a \epsilon\pi(a|s^\dagger) = \epsilon$, we have that $\underline{\Pi}$ is also a ϵ-cover of Π under ℓ_1 norm.

F. Model-Free Estimation by LSR Approach

In a model-free environment, we assume that the random variable of the cumulative reward $Z^{\boldsymbol{\pi}}$ is determined by the cumulative distribution function $\boldsymbol{F}$. Since the algorithm is given a random variable function class $\boldsymbol{\mathcal{Z}}$, we assume for every $\boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}$, its CDF $\boldsymbol{F} \in \boldsymbol{\mathcal{F}}$. Investigate $\boldsymbol{\mathcal{Z}}$ is equal to explore the CDF class $\boldsymbol{\mathcal{F}}$. Throughout this section, we use CDF to characterize the random variable.

F.1. Estimation and Algorithms

In this section, we use the least squares regression to estimate the confidence set of CDF.

Here we need the covering for $\Pi^\dagger$ and $\mathcal{F}$, defined in Definition A.1. For the policy set $\Pi^\dagger$ and function set $\mathcal{F}$, we use the metric: $\rho(\pi^1, \pi^2) = \max_{s^\dagger} \left\|\pi_1\left(\cdot|s^\dagger\right) - \pi_2\left(\cdot|s^\dagger\right)\right\|_1$ and $\rho(F_1, F_2) = \max_{s^\dagger, a} \left\|F_1(\cdot|s^\dagger, a) - F_2(\cdot|s^\dagger, a)\right\|_\infty$. For any policy $\pi \in \Pi^\dagger$ and $F \in \mathcal{F}$, we denote its ϵ-approximation in the cover $\underline{\Pi}^\dagger$ and $\underline{\mathcal{F}}$ as $\underline{\pi}$ and $\underline{F}$ respectively. Since $\underline{\pi}$ and $\underline{F}$ are ϵ-approximations of π and F, we have $\|\pi - \underline{\pi}\| \le \epsilon$ and $\|F - \underline{F}\| \le \epsilon$. We denote the covering number as $\mathcal{N}_C\left(\Pi^\dagger, \epsilon, \|\cdot\|_1\right)$ and $\mathcal{N}_C(\mathcal{F}, \epsilon, \|\cdot\|_\infty)$ respectively.

Algorithm 8 `V-Est-LSR`$(\mathcal{H}_{k-1}, \boldsymbol{\mathcal{F}}, \boldsymbol{\pi}, \gamma^{\mathrm{LSR}})$

Input: History information $\mathcal{H}_k$, CDF Model class $\mathcal{F}$, and policy $\boldsymbol{\pi}$.

Return: $\widehat{\mathcal{F}}_{k,\boldsymbol{\pi}} \leftarrow \left\{\boldsymbol{F} \in \boldsymbol{\mathcal{F}} : F_h \in \widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi},F}, \forall h \in [H]\right\}$, where $\widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi},\widetilde{F}}$ is defined by

$$\widehat{F}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}} = \operatorname*{argmin}_{F_h \in \mathcal{F}_h} \sum_{i=1}^{k-1} \left(F_h(x_{i,h}^{\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}} \mid s_{i,h}^\dagger, a_{i,h}) - \int_{a_{h+1}} \pi_h(a_{h+1} \mid s_{i,h+1}^\dagger) \widetilde{F}_{h+1}(x_{i,h}^{\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}} - r_{i,h} \mid s_{i,h+1}^\dagger, a_{h+1}) \right)^2,$$

$$\widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi},\widetilde{F}} = \left\{ F_h \in \mathcal{F}_h : \sum_{i=1}^{k-1} \left(F_h(x_{i,h}^{\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}} \mid s_{i,h}^\dagger, a_{i,h}) - \widehat{F}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}}(x_{i,h}^{\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}} \mid s_{i,h}^\dagger, a_{i,h}) \right)^2 \le \gamma^{\mathrm{LSR}} \right\}.$$

We define the LSR-type distance function used here:

$$\mathrm{Dist}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi},\widetilde{F}}(F_1||F_2) = \sum_{i=1}^{k-1} \left(F_1(x_{i,h}^{\boldsymbol{\pi},\widetilde{F}} \mid s_{i,h}^\dagger, a_{i,h}) - F_2(x_{i,h}^{\boldsymbol{\pi},\widetilde{F}} \mid s_{i,h}^\dagger, a_{i,h}) \right)^2, \tag{10}$$

and we define $x_{i,h}^{\boldsymbol{\pi},\widetilde{F}}$ as below, which represents the direction with largest uncertainty.

$$x_{i,h}^{\boldsymbol{\pi},\widetilde{F}} = \operatorname*{argmax}_{x \in [0,H]} \left| \sup_{F_1 \in \widehat{F}_{i,h,\boldsymbol{\pi},\widetilde{F}}} F_1(x \mid s_{i,h}^\dagger, a_{i,h}) - \inf_{F_2 \in \widehat{F}_{i,h,\boldsymbol{\pi},\widetilde{F}}} F_2(x \mid s_{i,h}^\dagger, a_{i,h}) \right|. \tag{11}$$

We now describe our estimation procedure above. For a target random variable with CDF $\widetilde{F}_{h+1}$ and policy $\boldsymbol{\pi} \in \Pi^\dagger$, we estimate $\mathcal{T}^\dagger_{h,\boldsymbol{\pi}} \widetilde{F}_{h+1}$ via least squares: $\widehat{F}_{k,h,\boldsymbol{\pi},\widetilde{F}}$. Define the distance function

$$\mathrm{Dist}^{\mathrm{LSR}}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}} \left(F_h || \widehat{F}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}} \right) = \sum_{i=1}^{k-1} \left(F_h(x_{i,h}^{\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}} \mid s_{i,h}^\dagger, a_{i,h}) - \widehat{F}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}}(x_{i,h}^{\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}} \mid s_{i,h}^\dagger, a_{i,h}) \right)^2$$

Then we can rewrite our version space using the distance metric defined above:

$$\widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi},\widetilde{F}} = \left\{ F_h \in \mathcal{F}_h : \mathrm{Dist}^{\mathrm{LSR}}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}} \left(F_h || \widehat{F}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}} \right) \le \gamma^{\mathrm{LSR}} \right\},$$

with the confidence radius $\gamma^{\mathrm{LSR}} = 16\log(HK^2/\delta) + \log(\mathcal{N}_C(\Pi^\dagger, 1/K, \|\cdot\|_1)) + \log(\mathcal{N}_C(\boldsymbol{\mathcal{F}}, 1/K, \|\cdot\|_\infty))$.

The next lemma shows the one-step-back concentration guarantee, which will be used to prove Condition 7.2.

Lemma F.1. *For any $\pi \in \Pi^\dagger$, $F \in \mathcal{F}$, $h \in [H]$, we have with probability at least $1-\delta$, for all $k \in [K]$:*

$$\mathcal{T}^\dagger_{h,\boldsymbol{\pi}} \tilde{F}_{h+1} \in \widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}} = \widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi},\widetilde{F}}\,.$$

Proof. First we fix $h \in [H]$, $\tilde{F}$ and π. Since

$$\left\{\mathcal{T}^\dagger_{h,\boldsymbol{\pi}} \widetilde{F}_h(x^{\boldsymbol{\pi},\widetilde{F}}_{i,h} \mid s^\dagger_{i,h}, a_{i,h}) - \int_{a_{h+1}} \pi_{h+1}(a_{h+1} \mid s^\dagger_{i,h+1}) \widetilde{F}_{h+1}(x^{\boldsymbol{\pi},\widetilde{F}}_{i,h} - r_{i,h} \mid s^\dagger_{i,h+1}, a_{h+1})\right\}_{i=1}^k$$

is a 1-sub-Gaussian. Moreover, we have

$$\begin{aligned}
&\mathbb{E}\left[\int_{a_h+1} \pi_{h+1}(a_{h+1} \mid s^\dagger_{i,h+1}) \widetilde{F}_{h+1}(x^{\boldsymbol{\pi},\widetilde{F}}_{i,h} - r_{i,h} \mid s^\dagger_{i,h+1}, a_{h+1})\middle| \tau_{i,h}\right] \\
=&\int_{s_{h+1}} \mathbb{P}_h(s_{h+1}|s_{i,h}, a_{i,h}) \int_{r_{h+1}} \mathbb{R}_h(r_{h+1}|s_{i,h}, a_{i,h}) \int_{a_{h+1}} \pi_{h+1}(a_{h+1}|s^\dagger_{h+1}) \widetilde{F}_{h+1}(x^{\boldsymbol{\pi},\widetilde{F}}_{i,h}|s^\dagger_{h+1}, a_{h+1}) \\
=&\int_{s^\dagger_{h+1}} \mathbb{T}_h(s^\dagger_{h+1}|s^\dagger_{i,h}, a_{i,h}) \int_{a_{h+1}} \pi_{h+1}(a_{h+1}|s^\dagger_{h+1}) \widetilde{F}_{h+1}(x^{\boldsymbol{\pi},\widetilde{F}}_{i,h} - r_{h+1}|s^\dagger_{h+1}, a_{h+1}) \\
=&\mathcal{T}_{h,\boldsymbol{\pi}} \widetilde{F}_{h+1}(x^{\boldsymbol{\pi},\widetilde{F}}_{i,h}|s^\dagger_{i,h}, a_{i,h})\,,
\end{aligned}$$

where $\tau_{i,h}$ denotes history up to and include step h in episode i. Thus by Lemma H.2, we have with probability at least $1-\delta/H$, for all $k \in [K]$,

$$\mathrm{Dist}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi},\tilde{F}}\left(\mathcal{T}^\dagger_{h,\boldsymbol{\pi}} \tilde{F}_{h+1} || \widehat{F}_{k,h,\boldsymbol{\pi},\tilde{F}}\right) \le 8\log\left(\frac{2H}{\delta}\right) + 4\left(1 + \sqrt{\log\left(\frac{4HK^2}{\delta}\right)}\right)\,. \tag{12}$$

Applying a union bound for all $h \in [H]$, $\underline{\widetilde{F}} \in \underline{\mathcal{F}}$, and $\underline{\boldsymbol{\pi}} \in \underline{\Pi}$, we have:

$$\mathrm{Dist}^{\mathrm{LSR}}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}}\left(\mathcal{T}^\dagger_{h,\underline{\boldsymbol{\pi}}} \underline{\tilde{F}}_{h+1} || \widehat{F}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}}\right) \le \mathcal{O}(\gamma^{\mathrm{LSR}})\,.$$

Moreover, we have

$$\mathrm{Dist}^{\mathrm{LSR}}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}}(\mathcal{T}^\dagger_{h,\boldsymbol{\pi}} \widetilde{F}_{h+1} || \mathcal{T}^\dagger_{h,\underline{\boldsymbol{\pi}}} \underline{\widetilde{F}}_{h+1}) \le \mathrm{Dist}^{\mathrm{LSR}}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}}(\mathcal{T}^\dagger_{h,\boldsymbol{\pi}} \widetilde{F}_{h+1} || \mathcal{T}^\dagger_{h,\underline{\boldsymbol{\pi}}} \widetilde{F}_{h+1}) + \mathrm{Dist}^{\mathrm{LSR}}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}}(\mathcal{T}^\dagger_{h,\underline{\boldsymbol{\pi}}} \widetilde{F}_{h+1} || \mathcal{T}^\dagger_{h,\underline{\boldsymbol{\pi}}} \underline{\widetilde{F}}_{h+1})\,.$$

For the first term, we have:

$$\begin{aligned}
&\mathrm{Dist}^{\mathrm{LSR}}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}}(\mathcal{T}^\dagger_{h,\boldsymbol{\pi}} \widetilde{F}_{h+1} || \mathcal{T}^\dagger_{h,\underline{\boldsymbol{\pi}}} \widetilde{F}_{h+1}) \\
=&\sum_{i=1}^{k-1}\left(\int_{s^\dagger_{h+1}} \mathbb{T}_h(s^\dagger_{h+1}|s^\dagger_{i,h}, a_{i,h}) \int_{a_{h+1}} \left(\pi_{h+1}(a_{h+1}|s^\dagger_{h+1}) - \underline{\pi}_{h+1}(a_{h+1}|s^\dagger_{h+1})\right) \widetilde{F}_{h+1}(x^{\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}}_{i,h} - r_h|s^\dagger_{h+1}, a_{h+1})\right)^2 \\
\le&\sum_{i=1}^{k-1}\left(\max_{s^\dagger_{h+1}} \left\|\pi_{h+1}(\cdot|s^\dagger_{h+1}) - \underline{\pi}_{h+1}(\cdot|s^\dagger_{h+1})\right\|_1\right)^2 \\
\le&K\epsilon\,.
\end{aligned}$$

For the second term, we also have:

$$\begin{aligned}
&\mathrm{Dist}^{\mathrm{LSR}}_{k,h,\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}}(\mathcal{T}^\dagger_{h,\underline{\boldsymbol{\pi}}} \widetilde{F}_{h+1} || \mathcal{T}^\dagger_{h,\underline{\boldsymbol{\pi}}} \underline{\widetilde{F}}_{h+1}) \\
=&\sum_{i=1}^{k}\left(\int_{s^\dagger_{h+1}} \mathbb{T}_h(s^\dagger_{h+1}|s^\dagger_{i,h}, a_{i,h}) \int_{a_{h+1}} \underline{\pi}_{h+1}(a_{h+1}|s^\dagger_{h+1}) \left(\widetilde{F}_{h+1} - \underline{\widetilde{F}}_{h+1}\right)(x^{\underline{\boldsymbol{\pi}},\underline{\widetilde{F}}}_{i,h} - r_h|s^\dagger_{h+1}, a_{h+1})\right)^2 \\
\le&\sum_{i=1}^{k} \max_{a_{h+1}, s^\dagger_{h+1}} \left\|\underline{\widetilde{F}}_{h+1}(s^\dagger_{h+1}, a_{h+1}) - \widetilde{F}_{h+1}(s^\dagger_{h+1}, a_{h+1})\right\|_\infty^2 \\
\le&K\epsilon\,.
\end{aligned}$$

Thus we have

$$
\begin{aligned}
&\mathrm{Dist}^{\mathrm{LSR}}_{k,h,\underline{\pi},\widetilde{\underline{F}}}\left(\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}\widetilde{F}_{h+1}||\widehat{F}_{k,h,\underline{\pi},\widetilde{\underline{F}}}\right)\\
\leq&\mathrm{Dist}^{\mathrm{LSR}}_{k,h,\underline{\pi},\widetilde{\underline{F}}}\left(\mathcal{T}^{\dagger}_{h,\underline{\pi}}\widetilde{\underline{F}}_{h+1}||\widehat{F}_{k,h,\underline{\pi},\widetilde{\underline{F}}}\right)+\mathrm{Dist}^{\mathrm{LSR}}_{k,h,\underline{\pi},\widetilde{\underline{F}}}(\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}\widetilde{F}_{h+1}||\mathcal{T}^{\dagger}_{h,\underline{\pi}}\widetilde{\underline{F}}_{h+1})\\
\leq&\mathcal{O}(\gamma^{\mathrm{LSR}}+K\epsilon)\\
=&\mathcal{O}(\gamma^{\mathrm{LSR}})\,.
\end{aligned}
$$

From the definition of the confidence set, we have for any $\tilde{F}_{h+1}\in\mathcal{F}_{h+1}$ and $\boldsymbol{\pi}\in\Pi^{\dagger}$, $\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}\tilde{F}_{h+1}\in\widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\underline{\pi},\widetilde{\underline{F}}}=\widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi},\widetilde{F}}$ where the equality is because the ϵ-approximation of $\underline{\pi}$ and $\widetilde{\underline{F}}$ are themselves. □

Proof of Condition 7.2. We proof the Lemma via induction. If $F^{\boldsymbol{\pi}}_{h+1}\in\widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h+1,\boldsymbol{\pi},F}$, we have $F^{\boldsymbol{\pi}}_{h}=\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}F^{\boldsymbol{\pi}}_{h+1}\in\widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi},F}$. So for all $h\in[H]$ we have $F^{\boldsymbol{\pi}}_{h}=\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}F^{\boldsymbol{\pi}}_{h+1}\in\widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi},F}$. From the definition of $\mathcal{F}_{k,\pi}$, we have $\boldsymbol{F}^{\boldsymbol{\pi}}\in\widehat{\mathcal{F}}_{k,\boldsymbol{\pi}}$. □

The next lemma decomposes the supremum distance between the CDFs of the cumulative return via the bellman error, which is the distributional analogue of the performance difference lemma.

Lemma F.2 (Performance difference)**.** *For any random variable Z representing the estimated cumulative return, with CDF function $\boldsymbol{F}=F_1\times F_2\times\cdots F_H$, we can decompose the ℓ_∞ distance between the estimated return CDF F_Z and the real return CDF $F_{Z^{\boldsymbol{\pi}}}$ for policy $\boldsymbol{\pi}\in\Pi^{\dagger}$ by the bellman error as follows:*

$$
\|F_Z-F_{Z^{\boldsymbol{\pi}}}\|_\infty\leq\sum_{h=1}^{H}\mathbb{E}_{\mu^{\boldsymbol{\pi}}}\left\|F_h(\cdot|s^{\dagger}_h,a_h)-\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}F_{h+1}(s^{\dagger}_h,a_h)\right\|_\infty\,.
$$

Proof.

$$
\begin{aligned}
&\|F_Z-F_{Z^{\boldsymbol{\pi}}}\|_\infty\\
=&\sup_{x\in[0,H]}\left|\int_{a_1}\pi_1(a_1|s^{\dagger}_1)\left(F_1(x|s^{\dagger}_1,a_1)-F^{\boldsymbol{\pi}}_1(x|s^{\dagger}_1,a_1)\right)\right|\\
\leq&\mathbb{E}_{s^{\dagger}_1,a_1\sim\mu^{\boldsymbol{\pi}}}\left\|F_1(s^{\dagger}_1,a_1)-F^{\boldsymbol{\pi}}_1(s^{\dagger}_1,a_1)\right\|_\infty\\
=&\mathbb{E}_{s^{\dagger}_1,a_1\sim\mu^{\boldsymbol{\pi}}}\left\|F_1(s^{\dagger}_1,a_1)-\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}F_2(s^{\dagger}_1,a_1)\right\|_\infty\\
\leq&\mathbb{E}_{\mu^{\boldsymbol{\pi}}}\left\|F_1(s^{\dagger}_1,a_1)-\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}F_2(s^{\dagger}_1,a_1)\right\|_\infty+\mathbb{E}_{\mu^{\boldsymbol{\pi}}}\left\|\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}F_2(s^{\dagger}_1,a_1)-\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}F^{\boldsymbol{\pi}}_2(s^{\dagger}_1,a_1)\right\|_\infty\\
\leq&\mathbb{E}_{\mu^{\boldsymbol{\pi}}}\left\|F_1(s^{\dagger}_1,a_1)-\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}F_2(s^{\dagger}_1,a_1)\right\|_\infty+\mathbb{E}_{\mu^{\boldsymbol{\pi}}}\left\|F_2(s^{\dagger}_2,a_2)-F^{\boldsymbol{\pi}}_2(s^{\dagger}_2,a_2)\right\|_\infty\,,
\end{aligned}
$$

where the first and second inequalities holds by triangle inequality, the third inequality is because:

$$
\begin{aligned}
&\mathbb{E}_{\mu^{\boldsymbol{\pi}}}\left\|\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}F_2(s^{\dagger}_1,a_1)-\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}}F^{\boldsymbol{\pi}}_2(s^{\dagger}_1,a_1)\right\|_\infty\\
=&\int_{s^{\dagger}_1,a_1}\mu^{\boldsymbol{\pi}}(s^{\dagger}_1,a_1)\sup_{x\in[0,H]}\left|\int_{s^{\dagger}_1,a_2}\mathbb{T}(s^{\dagger}_2|s^{\dagger}_1,a_1)\pi(a_2|s^{\dagger}_2)\left(F_2(x|s^{\dagger}_2,a_2)-F^{\boldsymbol{\pi}}_2(x|s^{\dagger}_2,a_2)\right)\right|\\
\leq&\int_{s^{\dagger}_1,a_1}\mu^{\boldsymbol{\pi}}(s^{\dagger}_1,a_1)\int_{s^{\dagger}_1,a_2}\mathbb{T}(s^{\dagger}_2|s^{\dagger}_1,a_1)\pi(a_2|s^{\dagger}_2)\left\|F_2(s^{\dagger}_2,a_2)-F^{\boldsymbol{\pi}}_2(s^{\dagger}_2,a_2)\right\|_\infty\\
=&\mathbb{E}_{\mu^{\boldsymbol{\pi}}}\left\|F_2(s^{\dagger}_2,a_2)-F^{\boldsymbol{\pi}}_2(s^{\dagger}_2,a_2)\right\|_\infty\,.
\end{aligned}
$$

Repeat this analysis for every step $h \in [H]$, we have

$$\begin{aligned}
&\|F_Z - F_{Z^{\boldsymbol{\pi}}}\|_\infty \\
\leq &\mathbb{E}_{\mu^{\boldsymbol{\pi}}} \left\| F_1(s_1^\dagger, a_1) - \mathcal{T}_{h,\boldsymbol{\pi}}^\dagger F_2(s_1^\dagger, a_1) \right\|_\infty + \mathbb{E}_{\mu^{\boldsymbol{\pi}}} \left\| F_2(s_2^\dagger, a_2) - F_2^{\boldsymbol{\pi}}(s_2^\dagger, a_2) \right\|_\infty \\
\leq &\cdots \\
\leq &\sum_{h=1}^H \left\| F_h(\cdot|s_h^\dagger, a_h) - \mathcal{T}_{h,\boldsymbol{\pi}}^\dagger F_{h+1}(s_h^\dagger, a_h) \right\|_\infty
\end{aligned}$$

□

Equipped with the technical lemmas above, we are able to prove the Condition 7.3 for model-free LSR estimation function.

Lemma F.3 (Condition 7.3)**.** *For $0 < \delta \leq 1$, we have*

$$\sum_{k=1}^K \left\| F_{\widehat{Z}^k} - F_{Z^{\boldsymbol{\pi}^k}} \right\|_\infty \leq \mathcal{O}\left(\mathrm{poly}(H) \sqrt{K \gamma^{\mathrm{LSR}} \dim_E(\mathcal{F}, \sqrt{K})} \right) ,$$

holds with probability at least $1 - \delta$.

proof of Condition 7.3. Using Hoeffding inequality in Lemma F.2, we have:

$$\begin{aligned}
&\left\| F_{\widehat{Z}^k} - F_{Z^{\boldsymbol{\pi}^k}} \right\|_\infty \\
\leq &\sum_{h=1}^H \mathbb{E}_{\mu^{\boldsymbol{\pi}^k}} \left\| \widehat{F}_h^k(\cdot|s_h^\dagger, a_h) - \mathcal{T}_{h,\boldsymbol{\pi}}^\dagger \widehat{F}_{h+1}^k(s_h^\dagger, a_h) \right\|_\infty \\
\leq &\sum_{h=1}^H \left\| \widehat{F}_h^k(s_{k,h}^\dagger, a_{k,h}) - \mathcal{T}_{h,\boldsymbol{\pi}}^\dagger \widehat{F}_{h+1}^k(s_{k,h}^\dagger, a_{k,h}) \right\|_\infty + \mathcal{O}\left(\sqrt{KH \log \frac{1}{\delta}} \right) .
\end{aligned}$$

Since $\mathcal{T}_{h,\boldsymbol{\pi}}^\dagger \widehat{F}_{h+1}^k \in \widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\underline{\boldsymbol{\pi}}^k,\underline{\widehat{F}}^k}$ by Lemma F.1, we have:

$$\begin{aligned}
&\sum_{k=1}^K \sum_{h=1}^H \left\| \widehat{F}_h^k(s_{k,h}^\dagger, a_{k,h}) - \mathcal{T}_{h,\boldsymbol{\pi}}^\dagger \widehat{F}_{h+1}^k(s_{k,h}^\dagger, a_{k,h}) \right\|_\infty \\
\leq &\sum_{k=1}^K \sum_{h=1}^H \sup_{x \in [0,H]} \left| \sup_{F_1 \in \widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\underline{\boldsymbol{\pi}}^k,\underline{\widehat{F}}^k}} F_1(x|s_{k,h}^\dagger, a_{k,h}) - \inf_{F_2 \in \widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\underline{\boldsymbol{\pi}}^k,\underline{\widehat{F}}^k}} F_2(x|s_{k,h}^\dagger, a_{k,h}) \right| \\
= &\sum_{k=1}^K \sum_{h=1}^H \sup_{F_1 \in \widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi}^k,\widehat{F}^k}} F_1(x_{k,h}^{\underline{\boldsymbol{\pi}}^k,\underline{\widehat{F}}^k}|s_{k,h}^\dagger, a_{k,h}) - \inf_{F_2 \in \widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi}^k,\widehat{F}^k}} F_2(x_{k,h}^{\underline{\boldsymbol{\pi}}^k,\underline{\widehat{F}}^k}|s_{k,h}^\dagger, a_{k,h}) ,
\end{aligned}$$

which is by the definition of $x_{k,h}^{\underline{\boldsymbol{\pi}}^k,\underline{\widehat{F}}^k}$ in Equation 11. Denote

$$G_{k,h} = \sup_{F_1 \in \widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi}^k,\widehat{F}^k}} F_1(x_{k,h}^{\underline{\boldsymbol{\pi}}^k,\underline{\widehat{F}}^k}|s_{k,h}^\dagger, a_{k,h}) - \inf_{F_2 \in \widehat{\mathcal{F}}^{\mathrm{LSR}}_{k,h,\boldsymbol{\pi}^k,\widehat{F}^k}} F_2(x_{k,h}^{\underline{\boldsymbol{\pi}}^k,\underline{\widehat{F}}^k}|s_{k,h}^\dagger, a_{k,h}) .$$

Using similar techniques as Lemma 9 of Chen et al. (2023), we have:

$$\sum_{k=1}^K \sum_{h=1}^H G_{k,h}^2 \leq H + H \dim_E(\mathcal{F}, \sqrt{K}) + 4H\gamma^{\mathrm{LSR}} \dim_E(\mathcal{F}, \sqrt{K})(\log(K) + 1) .$$

Thus, using Cauchy inequality, we obtain:

$$\sum_{k=1}^K \left\| F_{\widehat{Z}^k} - F_{Z^{\boldsymbol{\pi}^k}} \right\|_\infty \leq \mathcal{O}\left(\mathrm{poly}(H) \sqrt{K \gamma^{\mathrm{LSR}} \dim_E(\mathcal{F}, \sqrt{K})} \right) .$$

□

G. Model-Free Estimation by MLE approach

G.1. Bellman Eluder Dimension

In this section, we define the bellman eluder dimension (Jin et al., 2021), which is a famous structural complexity. First, we define the ℓ_2 norm distributional eluder dimension for a function class (Definition 7 in Jin et al. (2021)).

Definition G.1 (ℓ_2 norm distributional eluder dimension)**.** We consider Φ be a function class on domain $\mathcal{X}$ where for $\phi \in \Phi$, $|\phi(x)| \leq 1$. $\mathcal{D}$ is a family of distributions on $\mathcal{X}$. Let L be the longest sequence that there exists $\epsilon' > \epsilon$ and $\mu_1 \cdots \mu_L \in \mathcal{D}$, for all $t \in [L]$, there exists $\phi \in \Phi$, $|\mathbb{E}_{\mu_t}[\phi(x)]| \geq \epsilon$ and $\sum_{i=1}^{t-1} \left(\mathbb{E}_{\mu_i}[\phi(x)]\right)^2 \leq \epsilon^2$. We denote L as the bellman eluder dimension $\mathrm{d}_{\mathrm{DE}}(\Phi, \mathcal{D}, \epsilon)$.

Given the function set have low eluder dimension, we have the standard elliptical potential lemma as below:

Lemma G.2 (Lemma 17 in Jin et al. (2021))**.** *Given a function class $\phi \in \Phi$ in domain $\mathcal{X}$ with $|\phi(x)| \leq 1$. Let $\mathcal{D}$ be a families of distributions on $\mathcal{X}$. Suppose $\{\phi_k\}_{[K]} \subset \Phi$ and $\{\mu_k\}_{[K]} \subset \mathcal{D}$ be two sequences. If for any $k \in [K]$, $\sum_{i=1}^{k} \left(\mathbb{E}_{\mu_i}[\phi_k]\right)^2 \leq \beta$, then for any $k \in [K]$, $\sum_{i=1}^{k} |\mathbb{E}_{\mu_i}[\phi_i]| \leq \tilde{\mathcal{O}}\left(\sqrt{\mathrm{d}_{\mathrm{DE}}(\Phi, \mathcal{D}, 1/K)\beta K}\right)$.*

In this section, we define our bellman eluder dimension as the distributional eluder dimension for the specific function class below:

Definition G.3 (Bellman Eluder Dimension)**.** Given a policy class $\Pi^\dagger$, and a PDF function class $\mathcal{F}$. For $h \in [H]$, we define the function class Φ_h as $\{\mathrm{TV}\left(f_h || \mathcal{T}_{h,\boldsymbol{\pi}} f_{h+1}\right)\}$, and the distribution family $\mathcal{D}_h$ as $\mu_h^{\boldsymbol{\pi}}$ with domain $\mathcal{X}_h$ as $s_h^\dagger, a_h$. We define the bellman eluder dimension of our problem as:

$$\mathrm{d}_{\mathrm{BE}} = \max_{h \in [H]} \mathrm{d}_{\mathrm{DE}}(\Phi_h, \mathcal{D}_h, 1/K)\,.$$

G.2. Setting

G.2.1. Notation

In this setting, we assume the density function of $Z_h^{\boldsymbol{\pi}} \in \mathcal{Z}_h$ belongs to a function class $\mathcal{F}_h$. We denote the density function of $Z_h^{\boldsymbol{\pi}}(s^\dagger, a) \in \mathcal{Z}_h$ at point $z \in \mathbb{R}$ as $f_h^{\boldsymbol{\pi}}(z|s^\dagger, a) \in \mathcal{F}_h$.

Consider an upper and lower ϵ-bracketing of $\mathcal{F}$ under $\|\cdot\|_1$, denoted as $\mathcal{F}^\uparrow$ and $\mathcal{F}^\downarrow$. We denote the corresponding lower bracket of g as $g^\downarrow$, and the upper bracket of f as $f^\uparrow$. Since $g^\downarrow$ may not be a valid distribution, we denote the normalized version as: $\underline{g}$, where $\underline{g} = g^\downarrow / \int_z g^\downarrow(z)$, and $1 - \epsilon \leq \int_z g^\downarrow(z) \leq 1$. Thus, we have: $\underline{g} \leq (1 + 2\epsilon) g^\downarrow$.

G.2.2. Bellman Completeness

Define the augmented bellman operator:

$$\mathcal{T}_{h,\boldsymbol{\pi}}^\dagger f_{h+1}(z|s_h^\dagger, a_h) = \int \mathbb{T}(s_{h+1}^\dagger | s_h^\dagger, a_h) \pi_{h+1}(a_{h+1}|s_{h+1}^\dagger) f_{h+1}(z - r_h | s_h^\dagger, a_h)\,.$$

Here for completeness we restate the distributional bellman completeness assumption, and give a corresponding example.

Assumption G.4 (augmented distributional bellman completeness)**.** For the density function class $\boldsymbol{\mathcal{F}} = \mathcal{F}_1 \times \cdots \mathcal{F}_H$ corresponding to the class of random variables $\boldsymbol{\mathcal{Z}} = \mathcal{Z}_1 \times \mathcal{Z}_2 \cdots \mathcal{Z}_H$, we have for any $h \in [H]$, such that for any $f_{h+1} \in \mathcal{F}_{h+1}$, we have for any $\pi \in \Pi^\dagger$, $\mathcal{T}_{h,\boldsymbol{\pi}}^\dagger f_{h+1} \in \mathcal{F}_h$.

Instances: for linear MDP with $\mathbb{P}_h(s_{h+1}|s_h, a_h) = \langle \phi(s_h, a_h), \mu_h(s_{h+1}) \rangle$ and suppose the reward is discretized into a uniform grid of M points $\{z_i\}_{i=1}^M$. Then we can write the reward distribution as $\mathbb{R}(z_i|s_h, a_h) = \langle \mathbf{1_M}, \theta(z_i) \rangle$, where $\mathbf{1_M}$ is a M dimensional vector with all the entries being 1, and $\theta(z_i)$ is a M dimensional vector with all 0 but the i th entry equal

to $\mathbb{R}(z_i|s_h, a_h)$. Then we have:

$$
\begin{aligned}
&\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1}(z|s^{\dagger}_h, a_h) \\
=&\sum_{s_{h+1}} \mathbb{P}(s_{h+1}|s_h, a_h) \sum_{z_i} \mathbb{R}(z_i|s_h, a_h) \sum_{s} \pi_{h+1}\left(a_{h+1}|(s_{h+1}, y_h + z_i)\right) f_{h+1}\left(z - z_i|(s_{h+1}, y_h + z_i), a_{h+1}\right) \\
=&\sum_{s_{h+1}} \sum_{z_i} \phi(s_h, a_h)^{\top} \mu_h(s_{h+1}) \theta_h(z_i)^{\top} \mathbf{1}_{\mathbf{M}} \sum_{s} \pi_{h+1}\left(a_{h+1}|(s_{h+1}, y_h + z_i)\right) f_{h+1}\left(z - z_i|(s_{h+1}, y_h + z_i), a_{h+1}\right) \\
=&\phi(s_h, a_h)^{\top} W^{\boldsymbol{\pi}}_h(z, y_h) \mathbf{1}_{\mathbf{M}}\,,
\end{aligned}
$$

where

$$
W^{\boldsymbol{\pi}}_h(z, y_h) = \sum_{s_{h+1}} \sum_{z_i} \mu_h(s_{h+1}) \theta_h(z_i)^{\top} \sum_{a} \pi_{h+1}\left(a_{h+1}|(s_{h+1}, y_h + z_i)\right) f_{h+1}\left(z - z_i|(s_{h+1}, y_h + z_i), a_{h+1}\right)
$$

depends only on y_h and z. We can also write the distribution function in a linear form, with $f^{\boldsymbol{\pi}}_h(z|s^{\dagger}_h, a_h) = \bar{\phi}(s_h, a_h)^{\top} w^{\boldsymbol{\pi}}_h(z, y_h)$, with $\bar{\phi}(s_h, a_h) = \phi(s_h, a_h) \otimes \mathbf{1}_{\mathbf{M}} \in \mathbb{R}^{d \times M}$, and $w^{\boldsymbol{\pi}}_h(z, y_h)$ is the flattened version of $W^{\boldsymbol{\pi}}_h(z, y_h)$ with $w^{\boldsymbol{\pi}}_h(z, y_h)[d \times i + m] = W^{\boldsymbol{\pi}}_h(z, y_h)[i, m]$. Thus the function class has a linear structure similar to the case of risk neutral setting in linear MDPs (Jin et al., 2020).

G.3. Estimation of confidence set

Here, we estimate the confidence set $\mathcal{Z}_{k,\boldsymbol{\pi}}$ via MLE using the density functions f.

Algorithm 9 `V-Est-MLE`$(\mathcal{H}_k, \mathcal{Z}, \boldsymbol{\pi}, \gamma^{\mathrm{MLE}})$

Input: History information $\mathcal{H}_k$, density function class $\boldsymbol{\mathcal{F}}$ of random variable class $\boldsymbol{\mathcal{Z}}$, and policy $\boldsymbol{\pi}$.
For all $i \in [k]$, and $(h, \boldsymbol{f}, \boldsymbol{\pi}) \in [H] \times \mathcal{F} \times \Pi^{\dagger}$, sample $z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h+1} \sim \underline{f}_{h+1}\left(s^{\dagger}_{i,h+1}, \underline{\pi}_{i,h+1}(s^{\dagger}_{i,h+1})\right)$, and let $z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h} = z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h+1} + r_{i,h}$. We define the version space as:

$$
\widehat{\mathcal{F}}^{\mathrm{MLE}}_{k+1,\boldsymbol{\pi}} = \left\{ \boldsymbol{f} \in \mathcal{F} : \sum_{i=1}^{k} \log f_h(z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h}|s^{\dagger}_{i,h}, a_{i,h}) \geq \max_{f' \in \mathcal{F}} \sum_{i=1}^{k} \log f'_h(z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h}|s^{\dagger}_{i,h}, a_{i,h}) - \gamma^{\mathrm{MLE}}, \forall h \in [H] \right\}.
$$

Return $\widehat{\mathcal{Z}}_{k+1,\boldsymbol{\pi}}$ as the set of random variables corresponding to $\widehat{\mathcal{F}}^{\mathrm{MLE}}_{k+1,\boldsymbol{\pi}}$.

We now describe the sampling procedure for our target function $g_{h+1} \in \mathcal{F}$ and $\boldsymbol{\pi} \in \Pi$. Define $\underline{g}_{h+1}$ and $\underline{\boldsymbol{\pi}}$ as the normalized lower bracket in Section G.2.1. For $1 \leq i \leq k$, we sample $z^{\underline{g},\underline{\boldsymbol{\pi}}}_{i,h+1} \sim \underline{g}_{h+1}(s^{\dagger}_{i,h+1}, \underline{\pi}_{h+1}(s^{\dagger}_{i,h+1}))$, then we construct a one-step-back sample as $z^{\underline{g},\underline{\boldsymbol{\pi}}}_{i,h} = z^{\underline{g},\underline{\boldsymbol{\pi}}}_{i,h+1} + r_{i,h}$ where $r_{i,h} = y_{i,h+1} - y_{i,h}$. We estimate the likelihood of f_h for $\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} g_{h+1}$ as $\log f_h(z^{\underline{g},\underline{\boldsymbol{\pi}}}_{i,h}|s^{\dagger}_{i,h}, a_{i,h})$. Then we can define the MLE confidence set as:

$$
\mathcal{F}^{\mathrm{MLE}}_{h,k,g,\boldsymbol{\pi}} = \left\{ f_h \in \mathcal{F}_h : \sum_{i=1}^{k} \log f_h(z^{\underline{g},\underline{\boldsymbol{\pi}}}_{i,h}|s^{\dagger}_{i,h}, a_{i,h}) \geq \max_{f' \in \mathcal{F}} \sum_{i=1}^{k} \log f'_h(z^{\underline{g},\underline{\boldsymbol{\pi}}}_{i,h}|s^{\dagger}_{i,h}, a_{i,h}) - \gamma^{\mathrm{MLE}} \right\},
$$

where $\gamma^{\mathrm{MLE}} = \log\left(\mathcal{N}_{[\cdot]}\left(\mathcal{F}, \epsilon, \|\|_{\infty}\right) \mathcal{N}_{[\cdot]}(\Pi^{\dagger}, \epsilon, \|\|_1)/\delta\right) + K\epsilon$. Then, we can show that w.h.p, we have

$$
\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} g_{h+1} \in \mathcal{F}^{\mathrm{MLE}}_{h,k,g,\boldsymbol{\pi}}
$$

We can define our version space as:

$$
\widehat{\mathcal{F}}^{\mathrm{MLE}}_{k,\boldsymbol{\pi}} = \left\{ f \in \mathcal{F} : f_h \in \mathcal{F}^{\mathrm{MLE}}_{h,k,f,\boldsymbol{\pi}}, h \in [H] \right\}.
$$

Thus, we can prove that $f^{\boldsymbol{\pi}} \in \widehat{\mathcal{F}}^{\mathrm{MLE}}_{k,\boldsymbol{\pi}}$, and we have that $Z^{\boldsymbol{\pi}} \in \widehat{\mathcal{Z}}_{k,\boldsymbol{\pi}}$

G.3.1. Proof of Condition E.1

In this section we prove the concentration Condition 7.2 (Condition E.1 in the Appendix)

Following standard MLE concentration analysis, we have:

Lemma G.5. *For any $f_h \in \mathcal{F}_h$, $h \in [H]$, there exists a constant c such that:*

$$\sum_{i=1}^{k} \log f_h(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h}) \leq \sum_{i=1}^{k} \log(\mathcal{T}_{h,\boldsymbol{\pi}}^{\dagger} g_{h+1})(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h}) + c\gamma^{\mathrm{MLE}}\,.$$

Proof. Consider an upper and lower ϵ-bracketing of $\mathcal{F}$ under $\|\cdot\|_1$, denoted as $\mathcal{F}^{\uparrow}$ and $\mathcal{F}^{\downarrow}$. We denote the corresponding lower bracket of g_{h+1} as $g^{\downarrow}$, and the upper bracket of f as $f^{\uparrow}$. Since $g^{\downarrow}$ may not be a valid distribution, we denote the normalized version as: $\underline{g}$, where $\underline{g} = g^{\downarrow}/\int_z g^{\downarrow}(z)$, and $1-\epsilon \leq \int_z g^{\downarrow}(z) \leq 1$

Then, we have:

$$\begin{aligned}
&\mathbb{E}_{\mu_h^{\boldsymbol{\pi}^i}}\left[\frac{f_h^{\uparrow}(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h})}{\mathcal{T}_{h,\boldsymbol{\pi}}^{\dagger} g_{h+1}^{\downarrow}(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h})}\right] \\
=&\int_{s_h^{\dagger},a_h,s_{h+1}^{\dagger},z} \mu_h^{\boldsymbol{\pi}^i}(s_h^{\dagger}, a_h, s_{h+1}^{\dagger}) \frac{\int_{\mathcal{A}} \underline{\pi}_{h+1}(a_{h+1}|s_{h+1}^{\dagger}) \underline{g}_{h+1}\left(z - r_h|s_{h+1}^{\dagger}, a_{h+1}\right) f_h^{\uparrow}(z|s_h^{\dagger}, a_h)}{\int \mathbb{T}(s_{h+1}^{\dagger}|s_h^{\dagger}, a_h) \int_{\mathcal{A}} \pi_{h+1}(a_{h+1}|s_{h+1}^{\dagger}) g_{h+1}^{\downarrow}\left(z - r_h|s_{h+1}^{\dagger}, a_{h+1}\right)} \\
=&\int_{s_h^{\dagger},a_h} \mu_h^{\boldsymbol{\pi}^i}(s_h^{\dagger}, a_h) \int_z f_h^{\uparrow}(z|s_h^{\dagger}, a_h) \int_{s_{h+1}^{\dagger}} \frac{\mathbb{T}(s_{h+1}^{\dagger}|s_h^{\dagger}, a_h) \int_{\mathcal{A}} \underline{\pi}_{h+1}(a_{h+1}|s_{h+1}^{\dagger}) \underline{g}_{h+1}\left(z - r_h|s_{h+1}^{\dagger}, a_{h+1}\right)}{\int_{s_{h+1}^{\dagger}} \mathbb{T}(s_{h+1}^{\dagger}|s_h^{\dagger}, a_h) \int_{\mathcal{A}} \pi_{h+1}(a_{h+1}|s_{h+1}^{\dagger}) g_{h+1}^{\downarrow}\left(z - r_h|s_{h+1}^{\dagger}, a_{h+1}\right)} \\
\leq&\int_{s_h^{\dagger},a_h} \mu_h^{\boldsymbol{\pi}^i}(s_h^{\dagger}, a_h) \int_z f_h^{\uparrow}(z|s_h^{\dagger}, a_h) \int_{s_{h+1}^{\dagger}} \frac{\mathbb{T}(s_{h+1}^{\dagger}|s_h^{\dagger}, a_h) \int_{\mathcal{A}} (1+2\epsilon)\pi_{h+1}^{\downarrow}(a_{h+1}|s_{h+1}^{\dagger}) \underline{g}_{h+1}\left(z - r_h|s_{h+1}^{\dagger}, a_{h+1}\right)}{\int_{s_{h+1}^{\dagger}} \mathbb{T}(s_{h+1}^{\dagger}|s_h^{\dagger}, a_h) \int_{\mathcal{A}} \pi_{h+1}(a_{h+1}|s_{h+1}^{\dagger}) g_{h+1}^{\downarrow}\left(z - r_h|s_{h+1}^{\dagger}, a_{h+1}\right)} \\
\leq&\int_{s_h^{\dagger},a_h} \mu_h^{\boldsymbol{\pi}^i}(s_h^{\dagger}, a_h) \int_z f_h^{\uparrow}(z|s_h^{\dagger}, a_h)(1+2\epsilon)^2 \\
\leq&1 + \frac{6}{K}\,.
\end{aligned}$$

Thus we obtain the result via Markov inequality:

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{k} f_h^{\uparrow}(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h}) - \log(\mathcal{T}_{h,\boldsymbol{\pi}}^{\dagger} g_{h+1}^{\downarrow})(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h}) \geq \log(1/\delta)\right) \\
\leq&\mathbb{E}\left[\exp\left(\sum_{i=1}^{k} \log \frac{f^{\uparrow}(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h})}{\mathcal{T}_{h,\boldsymbol{\pi}}^{\dagger} g^{\downarrow}(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h})}\right)\right] \exp(-\log(1/\delta)) \\
\leq&e^6\delta\,.
\end{aligned}$$

Applying a union bound, for all $f^{\uparrow}$ and $g^{\downarrow}$, we have w.p. $1-\delta$, there exists a constant c,

$$\sum_{i=1}^{k} \log f_h^{\uparrow}(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h}) - \log(\mathcal{T}_{h,\boldsymbol{\pi}}^{\dagger} g_{h+1}^{\downarrow})(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h}) \leq c\beta\,.$$

We conclude our result by the definition of upper and lower brackets:

$$\begin{aligned}
&\sum_{i=1}^{k} \log f_h(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h}) - \log(\mathcal{T}_{h,\boldsymbol{\pi}}^{\dagger} g_{h+1})(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h}) \\
\leq&\sum_{i=1}^{k} \log f_h^{\uparrow}(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h}) - \log(\mathcal{T}_{h,\boldsymbol{\pi}}^{\dagger} g_{h+1}^{\downarrow})(z_{i,h}^{\underline{g},\underline{\boldsymbol{\pi}}}|s_{i,h}^{\dagger}, a_{i,h}) \\
\leq&c\beta\,.
\end{aligned}$$

As a result, we have:

$$\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1} \in \mathcal{F}^{\mathrm{MLE}}_{h,k,f,\boldsymbol{\pi}}$$

by the definition of $\mathcal{F}^{\mathrm{MLE}}_{h,k,f,\boldsymbol{\pi}}$ □

We prove Condition E.1 via induction. If $f^{\boldsymbol{\pi}}_{h+1} \in \mathcal{F}^{\mathrm{MLE}}_{h+1,k,f^{\boldsymbol{\pi}},\boldsymbol{\pi}}$, From Lemma G.5, $f^{\boldsymbol{\pi}}_h = \mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f^{\boldsymbol{\pi}}_{h+1} \in \mathcal{F}^{\mathrm{MLE}}_{h,k,f^{\boldsymbol{\pi}},\boldsymbol{\pi}}$. Since $f^{\boldsymbol{\pi}}_H \in \mathcal{F}^{\mathrm{MLE}}_{H,k,f^{\boldsymbol{\pi}},\boldsymbol{\pi}}$ we have for every h, $f^{\boldsymbol{\pi}}_h \in \mathcal{F}_{h,k,f^{\boldsymbol{\pi}},\boldsymbol{\pi}}$. Thus $\boldsymbol{f}^{\boldsymbol{\pi}} \in \widehat{\mathcal{F}}^{\mathrm{MLE}}_{k,\boldsymbol{\pi}}$. As a result, $Z^{\boldsymbol{\pi}} \in \widehat{\mathcal{Z}}_{k,\boldsymbol{\pi}}$ for any $k \in [K]$ and $\boldsymbol{\pi} \in \Pi^{\dagger}$.

G.3.2. Proof of Condition E.2

In this section we prove the elliptical potential Condition 7.3 (Condition E.2 in the Appendix).

The following lemma is the standard result for MLE generalization bound.

Lemma G.6 (MLE concentration)**.** *We can bound the square TV distance of the bellman error for any $\boldsymbol{f} \in \mathcal{F}_{k,\boldsymbol{\pi}}$*

$$\sum_{i=1}^{k} \mathbb{E}_{s^{\dagger}_h, a_h \sim \mu^{\boldsymbol{\pi}^i}_h} \left[\mathrm{TV}^2 \left(f_h(s^{\dagger}_h, a_h) || \mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1}(s^{\dagger}_h, a_h) \right) \right] \leq \mathcal{O}(\gamma^{\mathrm{MLE}}) .$$

Proof. Since we have for any $\boldsymbol{f} \in \mathcal{F}_{k,\boldsymbol{\pi}}$

$$\begin{aligned}
&\sum_{i=1}^{k} \log \left(\mathcal{T}^{\dagger}_{h,\underline{\boldsymbol{\pi}}} \underline{f}_{h+1} \right) (z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h+1} | s^{\dagger}_{i,h}, a_{i,h}) - \log f_h(z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h+1} | s^{\dagger}_{i,h}, a_{i,h}) \\
\leq &\sum_{i=1}^{k} \log \left(\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}^{\downarrow}} f^{\downarrow}_{h+1} \right) (z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h+1} | s^{\dagger}_{i,h}, a_{i,h}) - 2\log(1-\epsilon) - \log f_h(z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h+1} | s^{\dagger}_{i,h}, a_{i,h})
\end{aligned}$$

which holds by the normalization constant of $\underline{f}$ and $\underline{\pi}$. By the definition of the lower bracket function $f^{\downarrow}_{h+1}$, we have

$$\begin{aligned}
&\sum_{i=1}^{k} \log \left(\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}^{\downarrow}} f^{\downarrow}_{h+1} \right) (z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h+1} | s^{\dagger}_{i,h}, a_{i,h}) - 2\log(1-\epsilon) - \log f_h(z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h+1} | s^{\dagger}_{i,h}, a_{i,h}) \\
\leq &\sum_{i=1}^{k} \log \left(\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1} \right) (z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h+1} | s^{\dagger}_{i,h}, a_{i,h}) - 2\log(1-\epsilon) - \log f_h(z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h+1} | s^{\dagger}_{i,h}, a_{i,h}) \\
\leq &\max_{f' \in \mathcal{F}} \sum_{i=1}^{k} \log f'_h(z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h} | s^{\dagger}_{i,h}, a_{i,h}) - \log f_h(z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h+1} | s^{\dagger}_{i,h}, a_{i,h}) + \mathcal{O}(k\epsilon) \\
\leq &\gamma^{\mathrm{MLE}} + \mathcal{O}(k\epsilon) \\
= &\mathcal{O}(\gamma^{\mathrm{MLE}}) ,
\end{aligned}$$

where the second inequality is due to the distributional bellman completeness, which ensures that $\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1} \in \mathcal{F}_h$. The last two inequalities are due to the construction of the confidence set and the choice of $\epsilon = 1/K$.

Since the conditional distribution of $z^{\underline{f},\underline{\boldsymbol{\pi}}}_{i,h+1}$ given $s^{\dagger}_{i,h}, a_{i,h}$ is the same as $\mathcal{T}^{\dagger}_{h,\underline{\boldsymbol{\pi}}} \underline{f}_{h+1}(s^{\dagger}_h, a_h)$, from Lemma H.1, we have:

$$\sum_{i=1}^{k} \mathbb{E}_{s^{\dagger}_h, a_h \sim \mu^{\boldsymbol{\pi}^i}_h} \left[\mathrm{TV}^2 \left(f_h(s^{\dagger}_h, a_h) || \mathcal{T}^{\dagger}_{h,\underline{\boldsymbol{\pi}}} \underline{f}_{h+1}(s^{\dagger}_h, a_h) \right) \right] \leq \mathcal{O}(\gamma^{\mathrm{MLE}}) .$$

We also have:

$$\begin{aligned}
&\mathrm{TV} \left(\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1}(s^{\dagger}_h, a_h) || \mathcal{T}^{\dagger}_{h,\underline{\boldsymbol{\pi}}} \underline{f}_{h+1}(s^{\dagger}_h, a_h) \right) \\
\leq &\mathrm{TV} \left(\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1}(s^{\dagger}_h, a_h) || \mathcal{T}^{\dagger}_{h,\underline{\boldsymbol{\pi}}} f_{h+1}(s^{\dagger}_h, a_h) \right) + \mathrm{TV} \left(\mathcal{T}^{\dagger}_{h,\underline{\boldsymbol{\pi}}} \underline{f}_{h+1}(s^{\dagger}_h, a_h) || \mathcal{T}^{\dagger}_{h,\underline{\boldsymbol{\pi}}} f_{h+1}(s^{\dagger}_h, a_h) \right) .
\end{aligned}$$

We can bound the first term as:

$$
\begin{aligned}
&\mathrm{TV}\left(\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1}(s^{\dagger}_h, a_h) || \mathcal{T}^{\dagger}_{h,\underline{\boldsymbol{\pi}}} f_{h+1}(s^{\dagger}_h, a_h)\right) \\
\leq &\int_{s^{\dagger}_{h+1}} \mathbb{T}(s^{\dagger}_{h+1}|s^{\dagger}_h, a_h) \int_{a_{h+1}} \left|\underline{\pi}_{h+1}(a_{h+1}|s^{\dagger}_{h+1}) - \pi_{h+1}(a_{h+1}|s^{\dagger}_{h+1})\right| \int_z (f_{h+1}(z - r_h|s^{\dagger}_{h+1}, a_{h+1}) \\
\leq &\max_{s^{\dagger}_{h+1}} \left\|\underline{\pi}_{h+1}(\cdot|s^{\dagger}_{h+1}) - \pi_{h+1}(\cdot|s^{\dagger}_{h+1})\right\|_1 \\
\leq &\epsilon\,.
\end{aligned}
$$

We can also bound the second term as:

$$
\begin{aligned}
&\mathrm{TV}\left(\mathcal{T}^{\dagger}_{h,\underline{\boldsymbol{\pi}}} \underline{f}_{h+1}(s^{\dagger}_h, a_h) || \mathcal{T}_{h,\underline{\boldsymbol{\pi}}^{\dagger}} f_{h+1}(s^{\dagger}_h, a_h)\right) \\
\leq &\int_{s^{\dagger}_{h+1}} \mathbb{T}(s^{\dagger}_{h+1}|s^{\dagger}_h, a_h) \int_{a_{h+1}} \underline{\pi}_{h+1}(a_{h+1}|s^{\dagger}_{h+1}) \int_z \left|\left(f_{h+1} - \underline{f}_{h+1}\right)(z - r_h|s^{\dagger}_{h+1}, a_{h+1})\right| \\
\leq &\max_{s^{\dagger}_{h+1}, a_{h+1}} \left\|f_{h+1}(s^{\dagger}_{h+1}, a_{h+1}) - \underline{f}_{h+1}(s^{\dagger}_{h+1}, a_{h+1})\right\|_1 \\
\leq &\mathcal{O}(\epsilon)\,,
\end{aligned}
$$

since $\underline{f} \leq (1+2\epsilon)f^{\downarrow} \leq (1+2\epsilon)f$ and $\underline{f} \geq f^{\downarrow}$ point wise. Then we can have $\left\|\underline{f} - f\right\|_1 = \int_z \left|f(z) - \underline{f}(z)\right| \leq \int_z \max\left\{2\epsilon f, f - f^{\downarrow}\right\} \leq \int_z 2\epsilon f + \int_z (f - f^{\downarrow}) \leq 3\epsilon$.

We can conclude that: $\sum_{i=1}^{k} \mathbb{E}_{s^{\dagger}_h, a_h \sim \mu_h^{\boldsymbol{\pi}^i}} \left[\mathrm{TV}^2\left(\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1}(s^{\dagger}_h, a_h) || \mathcal{T}^{\dagger}_{h,\underline{\boldsymbol{\pi}}} \underline{f}_{h+1}(s^{\dagger}_h, a_h)\right)\right] \leq \mathcal{O}(k\epsilon)$. Thus,

$$
\sum_{i=1}^{k} \mathbb{E}_{s^{\dagger}_h, a_h \sim \mu_h^{\boldsymbol{\pi}^i}} \left[\mathrm{TV}^2\left(f_h(s^{\dagger}_h, a_h) || \mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1}(s^{\dagger}_h, a_h)\right)\right] \leq \mathcal{O}(\gamma^{\mathrm{MLE}})\,.
$$

□

Next, we present the distribution difference lemma for our model free analysis.

Lemma G.7. *We can bound the distance between the CDFs of the return by the bellman error of each step as:*

$$
\|F_Z - F_{Z^{\boldsymbol{\pi}}}\|_{\infty} \leq \sum_{h=1}^{H} \mathbb{E}_{(s^{\dagger}_h, a_h) \sim \mu_h^{\boldsymbol{\pi}}} \left[\left\|f_h(s^{\dagger}_h, a_h) - \mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1}(s^{\dagger}_h, a_h)\right\|_1\right]\,.
$$

Proof. We begin by induction. By triangle inequality:

$$
\left\|f_h(s^{\dagger}_h, a_h) - f^{\boldsymbol{\pi}}_h(s^{\dagger}_h, a_h)\right\|_1 \leq \left\|f_h(s^{\dagger}_h, a_h) - \mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1}(s^{\dagger}_h, a_h)\right\|_1 + \left\|\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1}(s^{\dagger}_h, a_h) - \mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f^{\boldsymbol{\pi}}_{h+1}(s^{\dagger}_h, a_h)\right\|_1\,.
$$

We derive a recursion for the second term.

$$
\begin{aligned}
&\left\|\mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f_{h+1}(s^{\dagger}_h, a_h) - \mathcal{T}^{\dagger}_{h,\boldsymbol{\pi}} f^{\boldsymbol{\pi}}_{h+1}(s^{\dagger}_h, a_h)\right\|_1 \\
= &\int_z \left|\int_{s^{\dagger}_{h+1}, a_h} \mathbb{T}(s^{\dagger}_{h+1}|s^{\dagger}_h, a_h) \pi_{h+1}(a_{h+1}|s^{\dagger}_{h+1}) \left(f_{h+1}(z - r_h|s^{\dagger}_{h+1}, a_{h+1}) - f^{\boldsymbol{\pi}}_{h+1}(z - r_h|s^{\dagger}_{h+1}, a_{h+1})\right)\right| \\
\leq &\int_{s^{\dagger}_{h+1}, a_h} \mathbb{T}(s^{\dagger}_{h+1}|s^{\dagger}_h, a_h) \pi_{h+1}(a_{h+1}|s^{\dagger}_{h+1}) \int_z \left|f_{h+1}(z - r_h|s^{\dagger}_{h+1}, a_{h+1}) - f^{\boldsymbol{\pi}}_{h+1}(z - r_h|s^{\dagger}_{h+1}, a_{h+1})\right| \\
= &\mathbb{E}_{(s^{\dagger}_{h+1}, a_{h+1}) \sim (s^{\dagger}_h, a_h)} \left\|f_{h+1}(s^{\dagger}_{h+1}, a_{h+1}) - f^{\boldsymbol{\pi}}_{h+1}(s^{\dagger}_{h+1}, a_{h+1})\right\|_1\,.
\end{aligned}
$$

Then we have:

$$
\begin{aligned}
&\mathbb{E}_{(s_h^\dagger,a_h)\sim\mu_h^{\boldsymbol{\pi}}}\left[\left\|f_h(s_h^\dagger,a_h)-f_h^{\boldsymbol{\pi}}(s_h^\dagger,a_h)\right\|_1\right]\\
\leq&\mathbb{E}_{(s_h^\dagger,a_h)\sim\mu_h^{\boldsymbol{\pi}}}\left[\left\|f_h(s_h^\dagger,a_h)-\mathcal{T}_{h,\boldsymbol{\pi}}^\dagger f_{h+1}(s_h^\dagger,a_h)\right\|_1\right]\\
&+\int_{s_h^\dagger,a_h}\mu_h^{\boldsymbol{\pi}}(s_h^\dagger,a_h)\int_{s_{h+1}^\dagger,a_{h+1}}\mu^{\boldsymbol{\pi}}(s_{h+1}^\dagger,a_{h+1}|s_h^\dagger,a_h)\left\|f_{h+1}(s_{h+1}^\dagger,a_{h+1})-f_{h+1}^{\boldsymbol{\pi}}(s_{h+1}^\dagger,a_{h+1})\right\|_1\\
\leq&\mathbb{E}_{(s_h^\dagger,a_h)\sim\mu_h^{\boldsymbol{\pi}}}\left[\left\|f_h(s_h^\dagger,a_h)-\mathcal{T}_{h,\boldsymbol{\pi}}^\dagger f_{h+1}(s_h^\dagger,a_h)\right\|_1\right]+\mathbb{E}_{(s_{h+1}^\dagger,a_{h+1})\sim\mu_{h+1}^{\boldsymbol{\pi}}}\left[\left\|f_{h+1}(s_{h+1}^\dagger,a_{h+1})-f_{h+1}^{\boldsymbol{\pi}}(s_{h+1}^\dagger,a_{h+1})\right\|_1\right].
\end{aligned}
$$

Using the definition of $F_Z(x)=\int_{s_1^\dagger,a_1}\mu_1^{\boldsymbol{\pi}}(s_1^\dagger,a_1)\int_{z\leq x}f_1(z|s_1^\dagger,a_1)$, we have:

$$
\begin{aligned}
&\|F_Z-F_{Z^{\boldsymbol{\pi}}}\|_\infty\\
\leq&\mathbb{E}_{s_1^\dagger,a_1\sim\mu_1^{\boldsymbol{\pi}}}\left[\left\|f_1\left(s_1^\dagger,a_1\right)-f_1^{\boldsymbol{\pi}}(s_1^\dagger,a_1)\right\|_1\right]\\
\leq&\sum_{h=1}^H\mathbb{E}_{(s_h^\dagger,a_h)\sim\mu_h^{\boldsymbol{\pi}}}\left[\left\|f_h(s_h^\dagger,a_h)-\mathcal{T}_{h,\boldsymbol{\pi}}^\dagger f_{h+1}(s_h^\dagger,a_h)\right\|_1\right].
\end{aligned}
$$

□

Combining the elliptical potential condition for low bellman eluder dimension (Lemma G.2) and the concentration result Lemma G.6, we have for all $h\in[H]$ and any $\hat{\boldsymbol{f}}^k\in\widehat{\mathcal{F}}_{k,\boldsymbol{\pi}^k}^{\mathrm{MLE}}$:

$$
\sum_{k=1}^K\mathbb{E}_{s_h^\dagger,a_h\sim\mu_h^{\boldsymbol{\pi}^k}}\left[\mathrm{TV}\left(\hat{f}_h^k(s_h^\dagger,a_h)||\mathcal{T}_{h,\boldsymbol{\pi}^k}^\dagger\hat{f}_{h+1}^k(s_h^\dagger,a_h)\right)\right]\leq\mathcal{O}(\sqrt{\mathrm{d_{BE}}\,\gamma^{\mathrm{MLE}}K}).
$$

Here we invoke Lemma G.2 by setting $\mathcal{X}:\mathcal{S}^\dagger\times\mathcal{A}$, $\Phi:\mathrm{TV}\left(f_h(s^\dagger,a)||\mathcal{T}_{h,\boldsymbol{\pi}}^\dagger f_{h+1}(s^\dagger,a)\right)$ for all $(f,\boldsymbol{\pi})\in\mathcal{F}\times\Pi^\dagger$. $\mathcal{D}:\mu^{\boldsymbol{\pi}},\ \boldsymbol{\pi}\in\Pi^\dagger$ is the family of all the visitation measures defined on $(s^\dagger,a)$.

Thus, using Lemma G.7, we have:

$$
\sum_{k=1}^K\left\|F_{\widehat{Z}^k}-F_{Z^{\boldsymbol{\pi}^k}}\right\|_\infty\leq\mathcal{O}(H\sqrt{\mathrm{d_{BE}}\,\gamma^{\mathrm{MLE}}K}).
$$

H. Auxiliary Lemmas

Lemma H.1 (MLE generalization bound (Theorem 21 of Agarwal et al. (2020))). *Let $\mathcal{X}$ be a feature space and $\mathcal{Y}$ be the output space. Given a dataset $D=\{(x_i,y_i)\}_{i=1}^n$ which is collected from a martingale process: $x_i\sim\mathcal{D}_i(x_{1:i-1},y_{1:i-1})$, and $y_i\sim p(\cdot|x_i)$. Given a function set $\mathcal{F}:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$, we have the real conditional distribution $f^*(x,y)=p(y|x)\in\mathcal{F}$. Then, there exists a constant c, for any $\delta>0$, with probability at least $1-\delta$, we have:*

$$
\sum_{i=1}^n\mathbb{E}_{x\sim\mathcal{D}_i}\left[\mathrm{TV}\left(f(x,\cdot)||f^*(x,\cdot)\right)^2\right]\leq c\left(\sum_{i=1}^n\log\left(\frac{f^*(x_i,y_i)}{f(x_i,y_i)}\right)+\log\left(\mathcal{N}_{[\cdot]}(\epsilon,\mathcal{F},\|\cdot\|_1/)\delta\right)+n\epsilon\right)
$$

Lemma H.2 (Concentration Lemma(Theorem 5 in Ayoub et al. (2020))). *let $(X_p,Y_p)_{p=1,2\cdots}$ be a set of random variables, $X_p\in\mathcal{X}$ for some measurable set $\mathcal{X}$ and $Y_p\in\mathbb{R}$. Let $\mathcal{F}$ be a set of real valued measurable function with domain $\mathcal{X}$. Let $\mathbb{F}=(\mathbb{F}_p)_{p=0,1,2\cdots}$ be a filtration such that for all $p\geq1$, we have $(X_1,Y_1,\cdots,X_{p-1},Y_{p-1},X_p)$ is $\mathbb{F}_{p-1}$ measurable, and such that there exists some function $f_*\in\mathcal{F}$ such that $\mathbb{E}[Y_p|\mathbb{F}_{p-1}]=f_*(X_p)$ for all $p\geq1$. Let $\hat{f}_t=\mathrm{argmin}_{f\in\mathcal{F}}\sum_{p=1}^t\left(f(X_p)-Y_p\right)^2$. Let $\mathcal{N}(\mathcal{F},\alpha)$ be the α-covering number of set $\mathcal{F}$ under $\|\cdot\|_\infty$ metric at scale α. Define $\mathrm{Dist}_t(f||f_t)=\sum_{p=1}^t\left(f(X_p)-f_t(X_p)\right)^2$.*

If the functions in $\mathcal{F}$ are bounded by some constant $C>0$. Assume that for each $p\geq1$, $(Y_p-f_*(X_p))$ is conditionally σ-sub-gaussian given $\mathbb{F}_{p-1}$. Then, for any $\alpha>0$, with probability $1-\delta$ for all $t\geq1$, we have:

$$
\mathrm{Dist}\left(f_*||f_t\right)\leq8\sigma^2\log\left(\mathcal{N}(\mathcal{F},\alpha)/\delta\right)+4t\alpha\left(C+\sqrt{\sigma^2\log\left(t(t+1)/\delta\right)}\right)
$$

I. Linear CVaR

Similarly to Liu et al. (2023); Jin et al. (2021), the general algorithms provided for general version space are information theoretic, which means that they cannot be implemented efficiently in general. This is because we consider the general risk measure LRM and the general function approximation settings. However, when specified to the CVaR risk measure under the discretized linear MDP, a distributional extension of natural linear MDP (Jin et al., 2020), we can design and implement an efficient model-free algorithm that achieves sub-linear regret.

Definition I.1 (Discretized Linear MDP)**.** An augmented MDP $\mathcal{M}^\dagger$ is a discretized linear MDP with feature map $\phi : (s,a) \to \mathbb{R}^d$ and an uniform grid of M points $\{z_i\}_{i=1}^M$, if for any $h \in [H]$, $(r_h, y_h) \in \{z_i\}_{i=1}^M$, and there exists unknown measures $\mu_h : \mathcal{S} \to \mathbb{R}^d$ and $\theta_h : \{z_i\}_{i=1}^M \to \mathbb{R}^d$, such that:

$$\mathbb{P}_h(s_{h+1}|s_h, a_h) = \phi(s_h, a_h)^\top \mu_h(s_{h+1})$$
$$\mathbb{R}_h(z_i|s_h, a_h) = \phi(s_h, a_h)^\top \theta_h(z_i)\ ,$$

for all s_h, a_h, s_{h+1}, z_i.

This discretized linear MDP is the natural extension of the linear MDP assumption in Jin et al. (2020), where we consider the discretized distributional reward instead of determined reward. and we generalize the linear expected reward to its distributional counterpart. Another important ingredient in our definition is the discretized reward space, which is commonly used in practice. C51 and Rainbow (Bellemare et al., 2017; Hessel et al., 2018) both set $M = 51$ and achieved empirical success in Atari games. We need the discretized reward space mainly to bound the covering number of the value distribution, similar to Wang et al. (2023b).

In a discretized linear MDP, we have that the distribution function have a quadratic structure: $f_h^\pi(z_i|(s_h, y_h), a_h) = \phi(s_h, a_h) W_h^\pi(z_i, y_h)\phi(s_h, a_h)$. Thus, we can use linear regression in estimating statistical functionals of $Z_h^{\boldsymbol{\pi}}(s_h^\dagger, a_h)$. We present our regret bound as follows:

Theorem I.2. *If MDP $\mathcal{M}^\dagger$ is a discretized linear MDP satisfying Definition I.1, and the risk measure $\rho(Z^{\boldsymbol{\pi}}) = CVaR_\tau(Z^\pi) = \max_b \left\{b - \tau^{-1}\mathbb{E}[(b - Z^\pi)^+]\right\}$, we can bound the regret as:*

$$\mathrm{Regret}(K) \le \tilde{\mathcal{O}}(\tau^{-1} d^3 H^2 \sqrt{MK})$$

We highlight that this is the first efficient model free algorithm for linear MDPs, and the $\sqrt{M}$ dependency is due to the covering number of the value distribution class which also appears in other model free distributional RL algorithms, such as Wang et al. (2023b).

Define $\mathrm{CVaR}_\tau(Z^\pi) = \mathrm{argmax}_b \left\{b - \tau^{-1}\mathbb{E}[(b - Z^\pi)^+]\right\}$. We can define the statistical functionals Q and V as: $Q_h^\pi(s_h, y_h, a_h) = \sum_{z_i} f_h^\pi(z_i|s_h, y_h, a_h)[(-z_i - y_h)^+]$, and $V_h^\pi(s_h, y_h) = \sum_{z_i} f_h^\pi(z_i|s_h, y_h, \pi(s_h, y_h))[(-z_i - y_h)^+]$. Then we can write the CVaR objective as $\mathrm{CVaR}_\tau(Z^\pi) = \mathrm{argmax}_{b\in[0,H]} \left\{b - \tau^{-1} V_1^\pi(s_1, -b)\right\}$.

Define recursively $\pi_h^*(s_h, y_h) = \mathrm{argmin}_a Q_h^{\pi^*}(s_h, y_h, a_h)$, then by Wang et al. (2023a) we have for any $b \in [0, H]$,

$$V_1^*(s_1, -b) = V_1^{\pi^*}(s_1, -b) = \operatorname*{argmin}_{\pi\in\Pi^\dagger} V_1^\pi(s_1, -b)$$

We denote $\mathrm{CVaR}_\tau^* = \max_{\pi\in\Pi^\dagger} \mathrm{CVaR}_\tau^\pi = b^* - \tau^{-1} V_1^*(s_1, -b^*)$

I.1. Linear Augmented MDPs

When consider the linear function approximation (Definition I.1), we can also linearize the augmented MDPs. We ahve

$$\mathbb{T}_h(s_{h+1}, r_h|s_h, a_h) = \phi(s_h, a_h)^\top \theta(r_h)\mu(s_{h+1})^\top \phi(s_h, a_h) = \psi(s_h, a_h)\chi(r_h, s_{h+1})$$

where $\psi(s_h, a_h)$ and $\chi(r_h, s_{h+1})$ are the flattened versions of $\phi(s_h, a_h)\phi(s_h, a_h)^\top$ and $\theta(r_h)\mu\left(s_{h+1}\right)^\top$. Also, we assume the reward space is discretized into M points $z_1 \cdots z_M$ such that for all $h \in [H]$, $y_h \in \{z_i\}_{i=1}^M$. We have that $-H \le z_1 \le z_2 \cdots z_M \le H$. We highlight that this discretization is standard in practice as in C51 (Bellemare et al., 2017) and Rainbow (Hessel et al., 2018). Here we need this assumption to bound the complexity of the function class.

I.2. Linear Completeness

Because the density function of the rewards satisfy the distributional bellman equation:

$$f_h^\pi(z_i|s_h, y_h, a_h) = \sum_{s_{h+1}, r_h} \mathbb{T}_h(s_{h+1}, r_h|s_h, a_h) f_{h+1}^\pi(z_i - r_h|s_{h+1}, y_h + r_h, \pi(s_{h+1}, y_h + r_h))$$

We have the statistical functionals satisfy the augmented bellman equation:

$$\begin{aligned} Q_h^\pi(s_h, y_h, a_h) &= \sum_{z_i} \sum_{s_{h+1}, r_h} \mathbb{T}_h(s_{h+1}, r_h|s_h, a_h) f_{h+1}^\pi(z_i - r_h|s_{h+1}, y_h + r_h, \pi(s_{h+1}, y_h + r_h))(-z_i - y_h)^+ \\ &= \sum_{s_{h+1}, r_h} \mathbb{T}_h(s_{h+1}, r_h|s_h, a_h) \sum_{z_i} f_{h+1}^\pi(z_i|s_{h+1}, y_h + r_h, \pi(s_{h+1}, y_h + r_h))(-z_i - y_h - r_h)^+ \\ &= (\mathbb{T}_h V_{h+1})(s_h, y_h, a_h) \end{aligned}$$

Where we denote the augmented transition operating on a function V as:

$$\mathbb{T}_h V_{h+1}(s_h, y_h, a_h) = \sum_{s_{h+1}, r_h} \mathbb{T}_h(s_{h+1}, r_h|s_h, a_h) V_{h+1}(s_{h+1}, y_h + r_h)$$

Since $f_h^\pi(z_i|s_h, y_h, a_h) = \phi(s_h, a_h)^\top W_h^\pi(y_h, z_i)\phi(s_h, a_h)$, we have $Q_h^\pi(s_h, y_h, a_h) = \psi(s_h, a_h)^\top w_h^\pi(y_h)$. Where $w_h^\pi[i \cdot d + j] = \sum_{z_i} W_h^\pi(y_h, z_i)[i, j][(-z_i - y_h)^+]$

I.3. Algorithm

In this section, we present our computationally efficient algorithm `RSRL-Linear-CVaR`. Notice that we present Upper Confidence Bound Value Iteration (UCV-VI) in this algorithms instead of the general optimistic planning used in previous frameworks, due to its computation-tractable property.

I.4. Concentration and Covering

In this section we provide the concentration and covering arguments needed for our linear analysis.

Lemma I.3 (Concentration Inequality of Self-normalized Process (Jin et al., 2020)). *Let $\{x_\tau\}_{\tau=1}^\infty$ be a stochastic process on domain $\mathcal{X}$ with corresponding filtration $\{\mathcal{F}_\tau\}$. Let $\{\psi_\tau\}$ be an $\mathbb{R}^{d^2}$ valued stochastic process stochastic process such that $\psi_\tau \in \mathcal{F}_{\tau-1}$ and $\|\psi_\tau\| \leq 1$. Let $\Lambda_k = \sum_{\tau=1}^{k-1} \psi_\tau \psi_\tau^\top$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $k > 0$, and any $V \in \mathbb{V}$ such that $|\sup_x V(x) - \inf_x V(x)| \leq H$, we have:*

$$\left\| \sum_{\tau=1}^{k-1} \psi_\tau \left\{ V(x_\tau) - \mathbb{E}[V(x_\tau|\mathcal{F}_{\tau-1}]\right\} \right\|_{\Lambda_k^{-1}}^2 \leq H^2 \left[\frac{d}{2} \log\left(\frac{k+\lambda}{\lambda}\right) + \log\left(\mathcal{N}_C(\epsilon, \mathbb{V}, \|\cdot\|_\infty)/\delta\right) \right] + \frac{8k^2\epsilon^2}{\lambda}.$$

Lemma I.4 (Lemma D.6 in Jin et al. (2020)). *Let $\mathcal{V}$ denote a class of functions mapping from domain $\mathcal{X}$ to $\mathbb{R}$ with the following parametric form:*

$$V(\cdot) = w^\top \phi(\cdot) + \beta \|\phi(\cdot)\|_{\Lambda^{-1}}$$

where $\phi \in \mathbb{R}^d$ are features on domain $\mathcal{X}$. The parameters satisfy $\|w\| \leq L$, $\lambda_{\min}(\Lambda) \geq \lambda$, $\beta \in [0, B]$, and $\|\phi(\cdot)\| \leq 1$. Then, the log covering number can be bounded as:

$$\log\left((\mathcal{N}_C(\epsilon, \mathcal{V}, \|\cdot\|_\infty)\right) \leq d\log(1 + 4L/\epsilon) + d^2 \log(1 + 8d^{1/2}B^2/\lambda\epsilon^2)$$

Lemma I.5 (Lemma B.2 in Jin et al. (2020)). *For $h \in [H]$ and $k \in [K]$, if $V_{h+1}^k(s_{h+1}, y_{h+1}) \leq H$, then $\left\|w_h^k\right\|_2 \leq Hd\sqrt{k/\lambda}$*

Lemma I.6 (Lemma B.1 in Jin et al. (2020)). *For any $h \in [H]$, $\|w_h^\pi\| \leq 2Hd$*

Algorithm 10 `RSRL-Linear-CVaR`

Input: Features $\Psi : \psi(s,a)$ Bonus $b_h^k(s,a) = \beta \left\|\psi(s,a)\right\|_{\Lambda_{k,h}^{-1}}$ with $\beta = c_\beta d^2 H \sqrt{M \log(d^2 UMK/\delta)}$
for $k = 1$ **to** K **do**
  Set $V_{H+1}(s_{H+1}, y) = [(-y)^+]$ for all $y \in \{z_i\}_{i=1}^M$
  for $h = H$ **to** 1 **do**

$$\Lambda_{k,h} = \sum_{i=1}^{k-1} \psi(s_h^i, a_h^k)\psi(s_h^i, a_h^i)^\top + \lambda \mathbf{I}$$
$$w_h^k(y_h) = \Lambda_{k,h}^{-1} \sum_{i=1}^{k-1} \psi(s_h^i, a_h^i) V_{h+1}^k(s_{h+1}^i, y_h + r_h^i)$$
$$Q_h^k(s_h, y_h, a_h) = \psi(s_h, a_h) w_h^k(y_h) - b_h^k(s_h, a_h)$$
$$\pi^k(s_h, y_h) = \operatorname*{argmin}_a Q_h^k(s_h, y_h, a_h)$$
$$V_h^k(s_h, y_h) = \max\left\{Q_h^k(s_h, y_h, \pi^k(s_h, y_h)), 0\right\}$$

  end for

$$b_k = \operatorname*{argmax}_{b \in \{z_i\}_{i=1}^M} \left\{b - \tau^{-1} V_1^k(s_1, -b)\right\}$$

  Start at state $s_1^\dagger = (s_1, -\lambda^k)$ and execute policy π^k in the augmented MDP $\mathcal{M}^\dagger$, collect information $\left\{(s_h^k, a_h^k, r_h^k)\right\}_{h \in [H]}$
end for

Lemma I.7. *If $V_{h+1}^k \le H$ for any $k \in [K]$, then there exists a constant C, for any $y_h \in \{z_i\}_{i=1}^M$, $\delta > 0$, with probability $1 - \delta$, we can bound the self normalized martingale process as:*

$$\left\|\sum_{i=1}^{k-1} \psi_{i,h} \left(V_h^k(s_h^i, y_{h-1} + r_{h-1}^i) - \mathbb{T}_{h-1} V_h^k(s_{h-1}^i, y_{h-1}, a_{h-1}^i)\right)\right\|_{\Lambda_{k,h}^{-1}}^2 \le C d^4 H^2 M \log(c_\beta d^2 KM/\delta)$$

Proof. According to Lemma I.5, we have that $\left\|w_h^k\right\|_2 \le Hd\sqrt{k/\lambda}$. According to Lemma I.4, we have for a fixed z_i the covering number of the function class $V_h^k(s_h, z_i)$ can be bounded as (notice that $\psi \in \mathbb{R}^{d^2}$):

$$\log\left((\mathcal{N}_C(\epsilon, \mathcal{V}(\cdot, z_i), \|\cdot\|_\infty)\right) \le d^2 \log(1 + 4Hd\lambda/k\epsilon) + d^4 \log(1 + 8d\beta^2/\lambda\epsilon^2)$$

Then we can bound the entire function class by regarding each $V(\cdot, z_i)$ an individual function. Thus, the total covering number can be bounded as:

$$\log\left((\mathcal{N}_C(\epsilon, \mathcal{V}, \|\cdot\|_\infty)\right) \le Md^2 \log(1 + 4Hd\lambda/k\epsilon) + d^4 \log(1 + 8d\beta^2/\lambda\epsilon^2)$$

Thus, we can apply Lemma I.3, and we have the result by choosing $\epsilon = 1/k$, $\lambda = 1$ and $\beta = c_\beta d^4 HM \sqrt{\log(d^2 UMK/\delta)}$. □

I.5. Proof of Theorem I.2

Define the event Optimism at $h \in [H]$ and $k \in [K]$ as:

$$Q_h^k(s_h, y_h, a_h) \le Q_h^*(s_h, y_h, a_h) \tag{Optimism}$$

Also, define

$$\psi(s,a) w_h^k(y) - Q_h^\pi(s, y, a) - \mathbb{T}_h(V_{h+1}^k - V_{h+1}^\pi)(s, a, y) = \Delta_h^k(s,a)$$

Then we have the following concentration result:

Lemma I.8. *If Optimism holds at $h \in [H]$ and $k \in [K]$. We have that $\Delta_h^k(s,a) \leq \beta \left\|\psi(s,a)\right\|_{\Lambda_{k,h}^{-1}}$.*

Proof. Since we have:

$$\begin{aligned}\psi(s_h^k, a_h^k)\mathbb{T}_h V_{h+1}^\pi(s_h^k, a_h^k, y_h) =& \sum_{s_{h+1}, r_h} \psi(s_h^k, a_h^k)\psi(s_h^k, a_h^k)^\top \chi(s_{h+1}, r_h) V_{h+1}^\pi(s_{h+1}, y_h + r_h) \\ =& \psi(s_h^k, a_h^k)\psi(s_h^k, a_h^k)^\top w_h^\pi(y_h)\end{aligned}$$

We bound the bias as:

$$\begin{aligned}w_h^k(y_h) - w_h^\pi(y_h) =& \Lambda_{k,h}^{-1}\left\{-\lambda w_h^\pi + \left[\sum_{i=1}^{k-1} \psi_{i,h}\left(V_{h+1}^k(s_{h+1}^i, y_h + r_h^i) - \mathbb{T}_h V_{h+1}^\pi(s_h^i, a_h^i, y_h)\right)\right]\right\} \\ =& \underbrace{-\lambda \Lambda_{k,h}^{-1} w_h^\pi}_{q_1} + \underbrace{\Lambda_{k,h}^{-1}\sum_{i=1}^{k} \psi_{i,h}\left(V_{h+1}^k(s_{h+1}^i, y_h + r_h^i) - \mathbb{T}_h V_{h+1}^k(s_h^i, a_h^i, y_h)\right)}_{q_2} \\ &+ \underbrace{\Lambda_{k,h}^{-1}\sum_{i=1}^{k} \psi_{i,h}\mathbb{T}_h(V_{h+1}^k - V_{h+1}^\pi)(s_h^i, a_h^i, y_h)}_{q_3}\end{aligned}$$

We have according to Lemma I.6:

$$\left|\psi(s_h, a_h)^\top q_1\right| \leq \sqrt{\lambda}\left\|w_h^\pi\right\| \left\|\psi(s_h, a_h)\right\|_{\Lambda_{k,h}^{-1}} \leq 2\sqrt{\lambda} H d \left\|\psi(s_h, a_h)\right\|_{\Lambda_{k,h}^{-1}}$$

For the second term, on event Optimism, we have that $V_{h+1}^k \leq V_{h+1}^* \leq H$. Apply Lemma I.7 we have:

$$\left|\psi(s_h, a_h)^\top q_2\right| \leq C d^2 H \sqrt{M \log(c_\beta d^2 K M/\delta)} \left\|\psi(s_h, a_h)\right\|_{\Lambda_{k,h}^{-1}}$$

For the third term, we have:

$$\begin{aligned}\psi(s_h, a_h)^\top q_3 =& \psi(s_h, a_h)\Lambda_{k,h}^{-1}\sum_{i=1}^{k}\psi_{i,h}\psi_{i,h}^\top \sum_{s_{h+1}, r_h} \chi(s_{h+1}, r_h)(V_{h+1}^k - V_{h+1}^\pi)(s_{h+1}, y_h + r_h) \\ =& \underbrace{\psi(s_h, a_h)^\top \sum_{s_{h+1}, r_h} \chi(s_{h+1}, r_h)(V_{h+1}^k - V_{h+1}^\pi)(s_{h+1}, y_h + r_h)}_{p_1} \\ &- \underbrace{\lambda\, \psi(s_h, a_h) \sum_{s_{h+1}, r_h} \chi(s_{h+1}, r_h)(V_{h+1}^k - V_{h+1}^\pi)(s_{h+1}, y_h + r_h)}_{p_2}\end{aligned}$$

$$p_1 = \mathbb{T}_h(V_{h+1}^k - V_{h+1}^\pi)(s_h, a_h)$$

$$|p_2| \leq \sqrt{\lambda} d^2 U(H - h + 1)) \left\|\psi(s_h, a_h)\right\|_{\Lambda_{k,h}^{-1}}$$

Thus, by choosing the appropriate constants, we have:

$$\left|\psi(s_h, a_h)^\top w_h^k(y_h) - Q_h^\pi(s_h, y_h, a_h) - \mathbb{T}_h(V_{h+1}^k - V_{h+1}^\pi)(s_h, a_h, y_h)\right| \leq \beta \left\|\psi(s_h, a_h)\right\|_{\Lambda_{k,h}^{-1}}$$

□

Lemma I.9 (Optimism)**.** *For any $h \in [H]$ and $k \in [K]$, event Optimism holds*

Proof. We prove the Lemma via induction. For any $k \in [K]$, Optimism holds at step H. Suppose that Optimism holds at step $h+1$, then we have from Lemma I.8,

$$\begin{aligned} Q_h^k(s_h, y_h, a_h) =& \psi(s_h, a_h) w_h^k(y_h) - \beta \left\|\psi(s,a)\right\|_{\Lambda_{k,h}^{-1}} \\ =& Q_h^\pi(s_h, y_h, a_h) + \mathbb{T}_h(V_{h+1}^k - V_{h+1}^\pi)(s_h, a_h, y_h) + \Delta_h^k(s_h, a_h) \\ \leq& Q_h^\pi(s_h, y_h, a_h) \end{aligned}$$

So Optimism holds at step h. □

Proof of Thoerem I.2. On event Optimism for all $h \in [H]$ and $k \in [K]$, we have that:

$$b_k - \tau^{-1} V_1^k(s_1, -b_k) \geq b^* - \tau^{-1} V_1^k(s_1, -b^*) \geq b^* - \tau^{-1} V_1^*(s_1, -b^*) = \mathrm{CVaR}_\tau^*$$

Also, we have that $\mathrm{CVaR}_\tau^{\pi^k} = \mathrm{argmax}_b \left\{ b - \tau^{-1} V_1^{\pi^k}(s_1, -b) \right\} \geq b_k - \tau^{-1} V_1^{\pi^k}(s_1, -b_k)$ Thus, the total regret can be bounded as:

$$\sum_{k=1}^K \mathrm{CVaR}_\tau^* - \mathrm{CVaR}_\tau^{\pi^k} \leq \sum_{k=1}^K \tau^{-1} \left(V_1^{\pi^k}(s_1, -b_k) - V_1^k(s_1, -b_k) \right)$$

Define $\delta_h^k = V_h^{\pi^k}(s_h^k, y_h^k) - V_h^k(s_h^k, y_h^k)$, and $\zeta_{h+1}^k = \mathbb{E}[\delta_{h+1}^k | s_h^k, y_h^k, a_h^k] - \delta_{h+1}^k$. According to Lemma I.8, we have that:

$$\delta_h^k \leq \delta_{h+1}^k + \zeta_{h+1}^k + 2\beta \left\|\psi_{k,h}\right\|_{\Lambda_{k,h}^{-1}}$$

Then we have:

$$\sum_{k=1}^K V_1^{\pi^k}(s_1, -b_k) - V_1^k(s_1, -b_k) \leq \sum_{h=1}^H \sum_{k=1}^K \zeta_h^k + 2 \sum_{h=1}^H \sum_{k=1}^K \beta \left\|\psi_{k,h}\right\|_{\Lambda_{k,h}^{-1}}$$

Since we have event Optimism, we have $\left|\delta_h^k\right| \leq H$. Thus, using the Hoeffding inequality, we have:

$$\sum_{h=1}^H \sum_{k=1}^K \zeta_h^k \leq \mathcal{O}\left(H \sqrt{K \log(KH/\delta)} \right)$$

Using the standard linear elliptical potential lemma, we obtain:

$$\sum_{h=1}^H \sum_{k=1}^K \left\|\psi_{k,h}\right\|_{\Lambda_{k,h}^{-1}} \leq \mathcal{O}(Hd\sqrt{K \log(KH/\delta)})$$

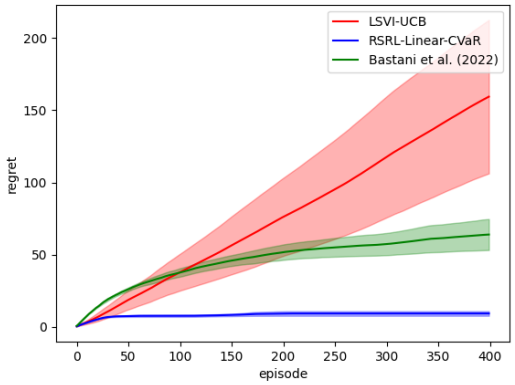
Then we have the regret bounded as:

$$\sum_{k=1}^K \mathrm{CVaR}_\tau^* - \mathrm{CVaR}_\tau^{\pi^k} \leq \tilde{\mathcal{O}}\left(\tau^{-1} H^2 d^3 \sqrt{MK} \right)$$

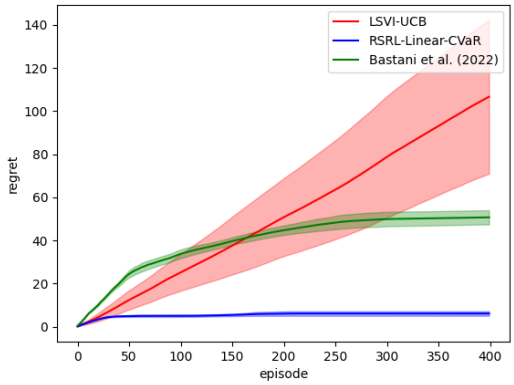
□

I.6. Numerical Experiment Results

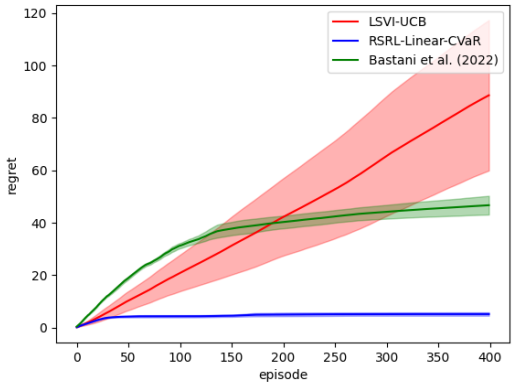
In this section we provide the details of our numerical experiments. We construct a zero-mean MDP where the expected return for all the state-action pairs are 0, thus risk-neutral algorithms such as LSVI-UCB of Jin et al. (2020) will learn nothing. We also compare our results with the optimistic MDP algorithm of Bastani et al. (2022). For simplicity we constructed a toy MDP with $S = 3$, $A = 2$, $d = 2$, $H = 6$, $M = 3$. The results are in Figure 1. From the figures we can see that the risk-neutral algorithm LSVI-UCB fails to learn anything, while the model-based algorithm of Bastani et al. (2022) converges much slower than ours. Also, with smaller τ we have larger regret, which is consistent with previous analysis.



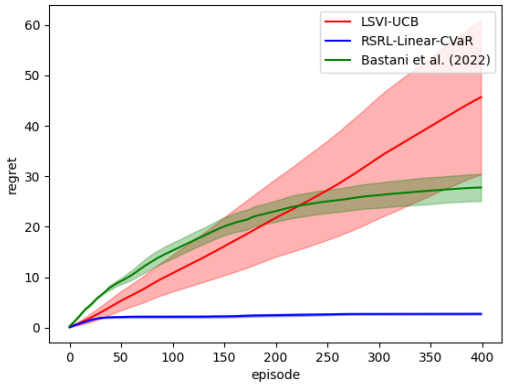
(a) $\mathrm{CVaR}_{0.2}$



(b) $\mathrm{CVaR}_{0.3}$.



(c) $\mathrm{CVaR}_{0.5}$



(d) $\mathrm{CVaR}_{0.7}$

Figure 1. Comparison for different algorithms for the CVaR objective CVaR_τ under different risk parameter τ.