Pessimism Meets Risk: Risk-Sensitive Offline Reinforcement Learning

Dake Zhang [1] **Boxiang Lyu** [1] **Shuang Qiu**[† 2] **Mladen Kolar** [3] **Tong Zhang** [4]

Abstract

We study risk-sensitive reinforcement learning (RL), a crucial field due to its ability to enhance decision-making in scenarios where it is essential to manage uncertainty and minimize potential adverse outcomes. Particularly, our work focuses on applying the entropic risk measure to RL problems. While existing literature primarily investigates the online setting, there remains a large gap in understanding how to efficiently derive a near-optimal policy based on this risk measure using only a pre-collected dataset. We center on the linear Markov Decision Process (MDP) setting, a well-regarded theoretical framework that has yet to be examined from a risk-sensitive standpoint. In response, we introduce two provably sample-efficient algorithms. We begin by presenting a risk-sensitive pessimistic value iteration algorithm, offering a tight analysis by leveraging the structure of the risk-sensitive performance measure. To further improve the obtained bounds, we propose another pessimistic algorithm that utilizes variance information and reference-advantage decomposition, effectively improving both the dependence on the space dimension d and the risk-sensitivity factor. To the best of our knowledge, we obtain the first provably efficient risk-sensitive offline RL algorithms.

1. Introduction

Reinforcement learning (RL) with risk-sensitivity is becoming increasingly popular in a variety of real-world risk-sensitive problems, such as finance (Föllmer & Schied, 2002; Hambly et al., 2021), optimal control (Nass et al., 2019), and neuroscience and psychology (Chateauneuf & Cohen, 1994; Nagengast et al., 2010; Braun et al., 2011; Niv et al., 2012). Whereas researchers have been focusing on analyzing and understanding how a risk-sensitive near-optimal policy can be learned in the online setting (Fei et al., 2021a;b; 2020; Fei & Xu, 2022; Liang & Luo, 2022; Du et al., 2023; Wang et al., 2023), i.e., when the learner is allowed to interact with and thereby explore the environment, in theory, little is known about how such policy can be learned with provable efficiency in the offline setting, where the learner has a pre-collected dataset but cannot interact with the environment (Urpí et al., 2021; Ma et al., 2021; Rigter et al., 2024).

Learning from interactions with the environment can be cost-prohibitive, thereby preventing us from actually learning and benefiting from these policies. Take the financial applications of RL for instance. Supposing we are training an RL agent to optimize the portfolio in the stock market (see Yu et al. (2019); Huang et al. (2021); Chaouki et al. (2020) for examples of such attempts), training an RL agent from scratch via online interactions in this setting may lead to significant financial losses, as these agents often require long periods of interacting with the environment before recovering a satisfactory policy. As such, many existing works focus on utilizing pre-collected datasets to learn an effective policy, falling into the offline RL setting. Moreover, many practical problems, such as financial applications, are highly risk-sensitive in nature. A rigorous theoretical analysis would help us better understand the possibilities and impossibilities of offline RL methods in such applications. Our work thus concentrates on answering the following critical question:

> *Can we design a provably efficient risk-sensitive RL algorithm with an offline dataset?*

Our work takes an initial step toward answering this question. Inspired by a long line of related works investigating risk-sensitivity in sequential decision-making (Howard & Matheson, 1972; Jaśkiewicz, 2007; Bäuerle & Rieder, 2014; Osogami, 2012; Tamar et al., 2012; Patek, 2001; Shapiro et al., 2021; Shen et al., 2014; Borkar, 2001; Mihatsch & Neuneier, 2002; Borkar, 2002; Borkar & Meyn, 2002; Di Masi & Stettner, 2007; Ma et al., 2020; Zhou et al., 2023; Moharrami et al., 2024), we investigate optimizing the follow-

---

[1]University of Chicago, IL, USA [2]Hong Kong University of Science and Technology, Hong Kong, China [3]University of Southern California, CA, USA [4]University of Illinois Urbana-Champaign, IL, USA [†]Corresponding Author. Correspondence to: Shuang Qiu <masqiu@ust.hk>.

ing entropic risk measure in Markov Decision Processes (MDPs), defined as

$$V_\beta := \frac{1}{\beta} \log \left\{ \mathbb{E}\left[e^{\beta R} \right] \right\}, \tag{1}$$

with $\beta \in \mathbb{R}$ being an adjustable parameter controlling the risk-sensitivity. Consider the second-order Taylor expansion of (1) for intuition, where $V_\beta = \mathbb{E}[R] + \frac{\beta}{2} \text{Var}[R] + O(\beta^2)$. When $\beta = 0$, we recover the risk-neutral objective. When $\beta < 0$, the objective is risk-averse, and for β's further away from 0, the objective further penalizes trajectories with larger reward variances. On the other hand, setting $\beta > 0$ leads to a risk-seeking objective, which can be used to model risk-seeking human behavior in psychology applications (Braun et al., 2011; Chateauneuf & Cohen, 1994). Moreover, to tackle the issue of large state space that can result in increased sample complexity in the tabular setting, our work considers the linear MDP, a popular theoretical framework for function approximation.

To the best of our knowledge, no existing work studies provably efficient offline RL with respect to the entropic risk measure in the tabular MDP setting, let alone the more general linear function approximation setting. Moreover, while it is common practice to design a pessimistic algorithm for risk-neutral offline RL, it remains elusive how pessimism could be implemented for risk-sensitive RL with the special entropic risk measure we focus on. Additionally, we also aim to design a variance-aware algorithm to sharpen our rates. However, we are unaware of any existing works incorporating variance estimation in risk-sensitive RL, even in those that focus on risk-sensitive online RL. Thus, it is technically challenging to provide the theoretical guarantees incorporating the entropic measure in offline RL for our proposed algorithms. In this work, by tackling the above challenges, we successfully propose two pessimistic algorithms to learn the optimal policy for the offline risk-sensitive RL with provable guarantees under the linear function approximation setting. We summarize our main contributions below.

Contributions. In this paper, we propose the first provably efficient risk-sensitive RL algorithms with linear function approximation for the offline setting. Specifically, our first algorithm is a pessimistic value iteration algorithm with a pessimistic bonus term devised using the structure of the entropic measure for eliminating spurious correlation. In addition, utilizing both variance information and reference-advantage decomposition, we develop a variance-aware pessimistic value iteration algorithm by devising a variance estimator for the entropic value function, aiming to improve our theoretical guarantee further. In our theoretical results, we show that the first algorithm is sufficiently efficient to learn an optimal policy with a guarantee depending on a risk-sensitivity factor and feature space dimension d, and then it can achieve a tighter guarantee using reference-advantage

decomposition with an improved feature space dimension dependence from d to $\sqrt{d}$ with an additional coverage assumption. Moreover, we prove that our variance-aware algorithm can effectively sharpen both the dependence on the feature space dimension d and the risk-sensitivity factor. When we take $\beta \to 0^+$, we can recover the best-known rate for risk-neutral offline RL. In our proofs, we provide a novel analysis of the covering number for risk-sensitive value function estimates in linear MDPs, which has not been studied in previous works, even under the online setting, making it of independent interest.

1.1. Related Work

Our work is related to a long line of works using linear function approximation in single-agent RL with an unknown transition and reward function. Particularly, whereas some works focus on the linear mixture MDPs (Zhou et al., 2021a; Hu et al., 2022; Chen et al., 2022; Zhou et al., 2021b; Cai et al., 2020; Zhang et al., 2021; Yang & Wang, 2019), where each state, action, next state tuple is mapped into a feature space, we focus on the linear MDP case where features depend on the state and action only (Jin et al., 2020; Zanette et al., 2021a; Jin et al., 2021; Wang et al., 2020a; Wagenmaker et al., 2022; Zhong & Zhang, 2024; Liu et al., 2024; Agarwal et al., 2020a). As the lines of research closely relevant to the linear MDP, the recent works further study low-rank MDPs (Agarwal et al., 2020b; Zhou et al., 2020; He et al., 2021; Uehara et al., 2021; Min et al., 2021; Qiu et al., 2022; Zhang et al., 2022a; Zheng et al., 2022; Mhammedi et al., 2024; Modi et al., 2024), where the state-action feature is learned instead of known a prior, and the kernel function approximation (Yang et al., 2020; Qiu et al., 2021), which covers the linear MDP as a special case. We leave the study of extending our method for linear MDPs to such settings as our future work.

We also draw inspiration from a line of work on risk-neutral offline RL (Levine et al., 2020) for the single-agent MDP, where the agent aims to recover a near-optimal policy from a pre-collected dataset under an unknown transition and reward function (Rashidinejad et al., 2021; Jin et al., 2021; Xie et al., 2021a; Cheng et al., 2022; Zanette et al., 2021b; Uehara & Sun, 2021; Xiong et al., 2022; Yin et al., 2022; Shi et al., 2022; Yin & Wang, 2021; Nguyen-Tang et al., 2023; Di et al., 2023; Li et al., 2024). Of these works, Xiong et al. (2022); Yin et al. (2022) achieve the tightest bounds for linear MDPs that we are aware of, and our work is able to recover their rates under similar assumptions when $\beta \to 0$ (i.e., risk-neutral). While the algorithm discussed in Section 4.2 is partially inspired by these works, it is important to emphasize that using variance estimation and reference-advantage decomposition in our risk-sensitive setting requires careful analysis of the specific problem structure defined by the entropic risk measure, which thus is not

straightforward given existing works.

Our work is closely related to a line of works studying the entropic risk measure in RL (Jaśkiewicz, 2007; Shen et al., 2013; Bäuerle & Rieder, 2014; Osogami, 2012; Patek, 2001; Nass et al., 2019; Shapiro et al., 2021; Shen et al., 2014; Borkar, 2001; Mihatsch & Neuneier, 2002; Borkar, 2002; Borkar & Meyn, 2002; Zhang et al., 2023; Di Masi & Stettner, 2007; Ma et al., 2020; Moharrami et al., 2024), dating back to Howard & Matheson (1972), which is consistent with the same risk measure's usage in financial mathematics (Detlefsen & Scandolo, 2005; Rudloff et al., 2008). Particularly, among recent works, Fei et al. (2021a;b; 2020); Fei & Xu (2022); Liang & Luo (2022) theoretically study online reinforcement learning under the entropic risk measure with theoretical guarantees, whereas we focus on the offline setting. Moreover, only Fei et al. (2021b) incorporates linear function approximation in the model, and the rest focuses only on tabular MDPs. Even for Fei et al. (2021b), the work only focuses on linear mixture MDPs, avoiding a challenging covering number analysis for the entropic risk value function estimates under linear MDPs. Orthogonal to our studies is another line of literature studying risk-sensitive RL with the CVaR risk measure (Prashanth, 2014; Lim & Malik, 2022; Bastani et al., 2022; Du et al., 2023; Wang et al., 2023; Zhao et al., 2023; Chen et al., 2023) or general risk measures (Wu & Xu, 2023) defined by different utility functions (Föllmer & Schied, 2002; Ben-Tal & Teboulle, 2007; Lee et al., 2020). It is interesting to further extend our analysis of the entropic risk measure to these various risk measures.

2. Preliminaries

Markov Decision Process. We study an episodic MDP characterized by a tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$, where $\mathcal{S}$ is a possibly infinite state space, $\mathcal{A}$ the agent's action space, H the episode length, $\mathbb{P} := \{\mathbb{P}_h\}_{h=1}^H$ the transition kernel with $\mathbb{P}_h(s'|s, a)$ being the probability density of transitioning from the current state s to the next state s' upon taking action $a \in \mathcal{A}$ at step h, and $r = \{r_h\}_{h=1}^H$ the reward function where $r_h : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$. We define $\pi = \{\pi_h\}_{h=1}^H$ with $\pi_h(\cdot|s) \in \Delta_{\mathcal{A}}$ for all $s \in \mathcal{S}$ as the policy for the agent where $\Delta_{\mathcal{A}}$ is the set of probability measures on $\mathcal{A}$. For interacting with the environment at step h, the agent at a state s_h takes an action $a_h \sim \pi_h(s_h)$ and transitions to the next state $s_{h+1} \sim \mathbb{P}_h(\cdot \mid s_h, a_h)$, receiving a reward $r_h(s_h, a_h)$ in the process. Without loss of generality, we assume that the interaction always starts from a fixed initial state s_1. We further assume that the transition kernel $\mathbb{P}$ and the reward function r are *unknown* to better reflect the challenges in real-world problems.

Linear MDP. When facing a large state space that may result in increased sample complexity in a tabular MDP, a common technique is making use of function approximation, such as the linear function approximation. In this paper, we consider the widely-studied linear MDP model, which admits a linear structure in both the reward function and the transition kernel, i.e.,

$$r_h(s, a) = \theta_h^\top \phi(s, a), \tag{2}$$
$$\mathbb{P}_h(\cdot|s, a) = \mu_h(\cdot)^\top \phi(s, a),$$

where $\max\{\int_{\mathcal{S}} \|\mu_h(s)\| \mathrm{d}s, \|\theta_h\|\} \leq \sqrt{d}$, and $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ is a feature map with $\|\phi(s, a)\| \leq 1$. In particular, with finite actions and states, by setting $d = |\mathcal{S}||\mathcal{A}|$ and $\phi(s, a) = e_{(s,a)}$ as the canonical basis, it reduces to the tabular MDP setting where the states and actions are discrete.

Value Function and Optimal Policy. Recall the entropic risk measure in (1). For risk-sensitive RL, we define the state-value function (Q-function) $Q_h^\pi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ as the expected cumulative rewards, measured by the entropic risk measure, under some policy π starting from a state-action tuple (s, a). More concretely, we let

$$Q_h^\pi(s, a) = \frac{1}{\beta} \log \left\{ \mathbb{E}_\pi \left[e^{\beta \sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})} \mid s_h = s, a_h = a \right] \right\},$$

where the expectation is taken over the trajectories induced by π. The value function $V_h^\pi : \mathcal{S} \mapsto \mathbb{R}$ is defined as the expected cumulative rewards under the policy π starting at a state s, which is

$$V_h^\pi(s) = \frac{1}{\beta} \log \left\{ \mathbb{E}_\pi \left[e^{\beta \sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})} \mid s_h = s \right] \right\}.$$

We introduce the following shorthand notations for the conditional expectation and variance of any function $f : \mathcal{S} \to \mathbb{R}$, taken over the randomness in the next step transition

$$(\mathbb{P}_h f)(s, a) = \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)}[f(s')|s_h = s, a_h = a],$$
$$(\mathrm{Var}_h f)(s, a) = (\mathbb{P}_h f^2)(s, a) - [(\mathbb{P}_h f)(s, a)]^2.$$

Letting $V_{H+1}^\pi(s) = Q_{H+1}^\pi(s, a) = 0$ for any (s, a), we have the following Bellman equation,

$$Q_h^\pi(s, a) = r_h(s, a) + \frac{1}{\beta} \log \left((\mathbb{P}_h e^{\beta V_{h+1}^\pi})(s, a) \right),$$
$$V_h^\pi(s) = \frac{1}{\beta} \log \left(\langle e^{\beta Q_h^\pi(s, \cdot)}, \pi(\cdot|s) \rangle_{\mathcal{A}} \right). \tag{3}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{A}}$ is the inner product over $\mathcal{A}$. The policy π can be stochastic in our setting. Based on the definition of the value function, we define the optimal policy π^* as $\pi^* := \operatorname{argmax}_\pi V_1^\pi(s_1)$. For brevity, we use Q_h^* and V_h^* to denote the optimal Q-function and optimal value function, respectively. Their relationship is characterized by the exponential Bellman optimality equation

$$e^{\beta Q_h^*(s,a)} = e^{\beta r_h(s,a)} \cdot (\mathbb{P}_h e^{\beta V_{h+1}^*})(s, a),$$
$$V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s, a).$$

Our goal is to learn a policy π that maximizes $V_1^\pi(s_1)$, and we measure the policy's performance by

$$\mathrm{SubOpt}(\pi) := V_1^*(s_1) - V_1^\pi(s_1),$$

that is, the suboptimality of the policy π given that the initial state is s_1. In addition, we say a policy π is ε-approximate optimal if $\mathrm{SubOpt}(\pi) \leq \varepsilon$.

Risk-Sensitive Offline RL. In offline RL, we assume that there exists a behavior policy that generates an offline dataset by interacting with the environment. With this dataset, we then employ offline RL algorithms to recover an estimate of π^* without interacting with the environment. Particularly, we make the following assumptions about the behavior policy and the offline dataset:

Assumption 2.1 (Offline Dataset). Suppose that there exists a behavior policy μ. After interacting with the environment for K rounds, we sample a dataset $\mathcal{D}$ consisting of K trajectories, which is defined as $\mathcal{D} := \{(s_h^\tau, a_h^\tau, r_h(s_h^\tau, a_h^\tau))\}_{h,\tau=1}^{H,K}$.

Assumption 2.1 does not enforce any coverage conditions on the dataset but precludes us from using techniques such as variance estimation and reference-advantage decomposition. As such, we make the following stronger but still common assumption on data coverage (Xiong et al., 2022; Wang et al., 2020b; Duan et al., 2020; Yin et al., 2022; Yin & Wang, 2021), with which we can obtain the improved dependence on d and tighter suboptimality guarantees.

Assumption 2.2 (Data Coverage). We assume that the smallest eigenvalue of the covariance matrix is bounded away from zero, that is

$$\kappa = \min_{h \in [H]} \lambda_{\min}(\mathbb{E}_{d_h^\mu}[\phi(s,a)\phi(s,a)^\top]) > 0,$$

where d_h^μ denotes the joint distribution of (s,a) over $\mathcal{S} \times \mathcal{A}$ at step h induced by the behavioral policy μ.

We note that Assumption 2.2 is not essential, and our algorithm can still efficiently learn a near-optimal policy without it, albeit at a slower rate. In the section on main results, we first prove a result for learning an ε-approximate optimal policy without this assumption in Theorem 5.1. Then, we target to obtain sharper guarantees based on Assumption 2.2. These sharper results are further provided in Theorem 5.2 using the same algorithm for Theorem 5.1, and in Theorem 5.3 with designing a different variance-aware algorithm.

3. Suboptimality Analysis of Risk-Sensitive RL

To motivate the algorithm design, we provide a high-level overview of the concentration terms we target during analysis and formally introduce the "shifting and scaling" technique. We begin by introducing the model evaluation error

under our setting, then show how the error relates to the suboptimality of a pessimistic value function estimate, and finally introduce the "shifting and scaling" technique, which ensures that the regression targets are on the same scale.

The suboptimality is closely related to the model evaluation error in offline RL (Jin et al., 2021). However, the standard analysis does not apply in our risk-sensitive setting because the Bellman equation has a different structure, as shown in (3), necessitating a new definition of model evaluation error. For any estimated optimal Q-function $\widehat{Q}_h$ and value function $\widehat{V}_h$, it seems natural to measure the estimation error of (3) by the difference between $\widehat{Q}_h(s,a)$ and $r_h(s,a) + \frac{1}{\beta}\log((\mathbb{P}_h e^{\beta\widehat{V}_{h+1}})(s,a))$. However, the direct approach has been shown to incur an extra factor $e^{|\beta|H^2}$ in the resulting upper bound, even in the tabular MDP setting considered by Fei et al. (2021a). Inspired by the work, we instead focus on the *exponential Bellman equation*, obtained by exponentiating both sides of (3),

$$e^{\beta Q_h^\pi(s,a)} = e^{\beta r_h(s,a)}(\mathbb{P}_h e^{\beta V_{h+1}^\pi})(s,a),$$
$$e^{\beta V_h^\pi(s)} = \langle e^{\beta Q_h^\pi(s,\cdot)}, \pi(\cdot|s) \rangle_{\mathcal{A}}. \tag{4}$$

The model evaluation error is then defined as

$$\iota_{\exp,h}(s,a) = e^{\beta r_h(s,a)}\mathbb{P}_h e^{\beta\widehat{V}_{h+1}}(s_h,a_h) - e^{\beta\widehat{Q}_h(s,a)},$$

where the subscript $\exp$ denotes that the evaluation error is defined with respect to the exponential Bellman equation.

Following the spirit of the exponential Bellman equation in (4), we leverage model evaluation error and pessimism to control $e^{\beta V_h^*(s)} - e^{\beta V_h^{\widehat{\pi}}(s)}$ directly, as opposed to controlling the term via bounding $V_h^*(s) - V_h^{\widehat{\pi}}(s)$.

Lemma 3.1. *Let $\widehat{Q}$ be a pessimistic estimate of Q, satisfying* $\mathrm{sign}(\beta)\iota_{\exp,h}(s,a) \geq 0$ *for any* $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. $\widehat{V}_h(s) = \max_{a \in \mathcal{A}} \widehat{Q}_h(s,a)$ *denote the value function induced by $\widehat{Q}$. Then, we have for all $\beta > 0$,*

$$\mathrm{SubOpt}(\widehat{\pi}) \leq \frac{1}{\beta}\left(e^{\beta V_1^*(s_1)} - e^{\beta V_1^{\widehat{\pi}}(s_1)}\right)$$

$$\leq \sum_{h=1}^H \frac{e^{\beta(h-1)}}{\beta}\mathbb{E}_{\pi^*}\left[\iota_{\exp,h}(s_h,a_h)|s_1\right],$$

and for all $\beta < 0$,

$$\mathrm{SubOpt}(\widehat{\pi}) \leq \frac{e^{-\beta H}}{\beta}\left(e^{\beta V_1^*(s_1)} - e^{\beta V_1^{\widehat{\pi}}(s_1)}\right)$$

$$\leq \frac{e^{-\beta H}}{\beta}\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\iota_{\exp,h}(s_h,a_h)|s_1\right],$$

where $\widehat{\pi}$ is the greedy policy taken with respect to $\widehat{Q}$, defined by $\widehat{\pi}_h(s) = \mathrm{argmax}_{a \in \mathcal{A}} \widehat{Q}_h(s,a)$.

We can observe that the model evaluation errors propagate backward differently. When $\beta < 0$ (risk-averse), the model valuation errors at each step scale the same in the bound, whereas errors at later steps are scaled upward when $\beta > 0$ (risk-seeking). The difference in scale, while inevitable, can be better accounted for via the following "shifting and scaling" technique.

Shifting and Scaling. We define the following shifting and scaling transformation:

$$\mathbb{S}_h f(\cdot) = \begin{cases} e^{\beta(h-1)}(e^{\beta f(\cdot)} - 1), & \beta > 0 \\ -e^{-\beta H}(e^{\beta f(\cdot)} - 1), & \beta < 0 \end{cases} \quad (5)$$

for any function f satisfying $f(s) \in [0, H+1-h]$. With the operator defined, at each step, rather than regressing directly on $\exp(V_h(s'))$, with s' denoting the observed state at the next step, we instead regress on $\mathbb{S}_h V_h(s')$ to rescale the regression targets at each step.

For the case $\beta > 0$, there is a scaling factor of $e^{\beta(h-1)}$ in (5), ensuring that the regression targets at each step are scaled according to the decomposition bounds in Lemma 3.1. For $\beta < 0$, the scaling factor $e^{-\beta H}$ is similarly constructed to ensure that the regression targets are on the same scale as the suboptimality bounds. Together with the shifting term, the range of the value function after the transformation satisfies $\mathbb{S}_h V_h(\cdot) \in [0, e^{|\beta|H} - 1]$ for any $\beta \neq 0$ and h. While shifting does not affect the scale of the model evaluation error, it ensures that the error's range starts at zero. For concentration analysis, we can then focus only on regression targets in the range $[0, e^{|\beta|H} - 1]$, regardless of the sign of β or the value of h. Our shifting and scaling operator integrates principles from shifted exponential V-functions in Fei et al. (2021b) and decaying bonus in Fei et al. (2021a) from a unified perspective. Additionally, our operator can rescale the target in advance and thus avoid adjusting the uncertainty bonus at each step as in Fei et al. (2021a).

This method provides multiple benefits. To understand the benefits of shifting, we examine our resulting upper bound when $\beta > 0$, which depends on β through a risk-sensitivity factor, $\frac{e^{\beta H} - 1}{\beta}$. If we overlook the property that $e^{\beta V_h} \geq 1$ and merely treat $e^{\beta V_h}$ as a function bounded by $e^{\beta H}$ without shifting it, the term $e^{\beta H} - 1$ in the upper bound will be replaced with $e^{\beta H}$. By shifting $e^{\beta V_h}$ by 1, the upper bound could be improved by a factor $(e^{\beta H} - 1)/e^{\beta H} < 1$, which could be much smaller, particularly when β approaches zero. Additionally, scaling is also beneficial. As shown in 3.1, the errors are scaled upward when $\beta > 0$. If errors at different steps are treated uniformly without incorporating the fact that V_hs have different ranges, the additional $e^{|\beta|H^2}$ term in the resulting upper bound in Fei et al. (2021b) cannot be entirely eliminated. Moreover, as the regression targets are scaled and shifted correctly, we avoid the need to derive suitable concentration bounds over the different

---

Algorithm 1 RSPVI Algorithm

1: **Input:** Dataset $\mathcal{D} := \{(s_h^\tau, a_h^\tau, r_h(s_h^\tau, a_h^\tau))\}_{h,\tau=1}^{H,K}$.

2: **Initialize:** $\widehat{V}_{H+1}(\cdot) = 0$.

3: **for** $h = H, \ldots, 1$ **do**

4: $\quad \Lambda_h \leftarrow \sum_{\tau=1}^K \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda I_d$

5: $\quad \widehat{\theta}_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^K \phi(s_h^\tau, a_h^\tau) r_h(s_h^\tau, a_h^\tau)$

6: $\quad \widehat{w}_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^K \phi(s_h^\tau, a_h^\tau) \mathbb{S}_{h+1}\widehat{V}_{h+1}(s_{h+1}^\tau)$

7: $\quad \widehat{r}_h(\cdot, \cdot) \leftarrow \{\phi(\cdot, \cdot)^\top \widehat{\theta}_h\}_{[0,1]}$

8: $\quad \Gamma_h(\cdot, \cdot) \leftarrow \gamma \|\phi(\cdot, \cdot)\|_{\Lambda_h^{-1}}$

9: $\quad \widehat{Q}_h(\cdot, \cdot) \leftarrow \{\frac{1}{\beta} \log(q_h(\cdot, \cdot))\}_{[0, H+1-h]}$ where q_h is

$\quad \bullet$ when $\beta > 0$,

$\quad e^{\beta(1-h)} \left[e^{\beta(\widehat{r}_h(\cdot, \cdot) - 1)}(\phi(\cdot, \cdot)^\top \widehat{w}_h + e^{\beta h}) - \Gamma_h(\cdot, \cdot) \right]$

$\quad \bullet$ when $\beta < 0$,

$\quad e^{\beta H} \left[e^{\beta \widehat{r}_h(\cdot, \cdot)}(e^{-\beta H} - \phi(\cdot, \cdot)^\top \widehat{w}_h) + \Gamma_h(\cdot, \cdot) \right]$

10: $\quad \widehat{\pi}_h(\cdot|\cdot) \leftarrow \operatorname{argmax}_{\pi_h} \langle \widehat{Q}_h(\cdot, \cdot), \pi_h(\cdot|\cdot) \rangle_{\mathcal{A}}$

11: $\quad \widehat{V}_h(\cdot) \leftarrow \frac{1}{\beta} \log\left(\langle \exp(\beta \widehat{Q}_h(\cdot, \cdot)), \widehat{\pi}_h(\cdot|\cdot) \rangle_{\mathcal{A}} \right)$

12: **end for**

13: **Output:** $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$

---

ranges of regression targets caused by the choices of β and h, streamlining the analysis.

4. Algorithm

In this section, we propose two offline algorithms for risk-sensitive RL under the linear MDP setting.

4.1. Risk-Sensitive Pessimistic Value Iteration

We begin by introducing the algorithm Risk-Sensitive Pessimistic Value Iteration (RSPVI) as summarized in Algorithm 1. The algorithm's performance relies on the shifting and scaling technique, and its construction draws inspiration from Jin et al. (2021).

Specifically, Lines 4-6 in Algorithm 1 perform two ridge regressions, and $\widehat{\theta}_h$ and $\widehat{w}_h$ are the solutions of the following minimization problems:

$$\min_{\theta \in \mathbb{R}^d} \sum_{\tau=1}^K \left[r_h(s_h^\tau, a_h^\tau) - \phi_h(s_h^\tau, a_h^\tau)^\top \theta \right]^2 + \lambda \|\theta\|_2^2,$$

and

$$\min_{w \in \mathbb{R}^d} \sum_{\tau=1}^K \left[\mathbb{S}_{h+1}\widehat{V}_{h+1}(s_{h+1}^\tau) - \phi_h(s_h^\tau, a_h^\tau)^\top w \right]^2 + \lambda \|w\|_2^2.$$

Combining such results, the estimated reward functions are constructed by $\widehat{r}_h(\cdot, \cdot) = \{\phi(\cdot, \cdot)^\top \widehat{\theta}_h\}_{[0,1]}$, where

---

Algorithm 2 VA-RSPVI Algorithm

1: **Input:** Dataset $\mathcal{D} := \{(s_h^\tau, a_h^\tau, r_h(s_h^\tau, a_h^\tau))\}_{h,\tau=1}^{H,K}$ and auxiliary dataset $\mathcal{D}^{\mathrm{aux}} := \{(\breve{s}_h^\tau, \breve{a}_h^\tau, r_h(\breve{s}_h^\tau, \breve{a}_h^\tau))\}_{h,\tau=1}^{H,K}$

2: **Initialize:** Set $\widehat{V}_{H+1}(\cdot) = 0$

3: Construct variance estimator $\widehat{\sigma}_h^2(\cdot, \cdot)$ with $\mathcal{D}^{\mathrm{aux}}$ via (6). Let $(\widehat{\sigma}_h^\tau)^2 := \widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)$.

4: **for** $h = H, \ldots, 1$ **do**

5:      $\Sigma_h \leftarrow \sum_{\tau=1}^K \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top / (\widehat{\sigma}_h^\tau)^2 + \lambda I_d$

6:      $\widehat{\theta}_h \leftarrow \Sigma_h^{-1} \sum_{\tau=1}^K \phi(s_h^\tau, a_h^\tau)r_h(s_h^\tau, a_h^\tau)/(\widehat{\sigma}_h^\tau)^2$

7:      $\widehat{w}_h \leftarrow \Sigma_h^{-1} \sum_{\tau=1}^K \phi(s_h^\tau, a_h^\tau)\mathbb{S}_{h+1}\widehat{V}_{h+1}(s_{h+1}^\tau)/(\widehat{\sigma}_h^\tau)^2$

8:      $\widehat{r}_h(\cdot, \cdot) \leftarrow \{\phi(\cdot, \cdot)^\top \widehat{\theta}_h\}_{[0,1]}$

9:      $\Gamma_h(\cdot, \cdot) \leftarrow \gamma \|\phi(\cdot, \cdot)\|_{\Sigma_h^{-1}}$

10:      $\widehat{Q}_h(\cdot, \cdot) \leftarrow \{\frac{1}{\beta}\log(q_h(\cdot, \cdot))\}_{[0,H+1-h]}$ where q_h is

       • when $\beta > 0$,

$$e^{\beta(1-h)}\left[e^{\beta(\widehat{r}_h(\cdot, \cdot)-1)}(\phi(\cdot, \cdot)^\top \widehat{w}_h + e^{\beta h}) - \Gamma_h(\cdot, \cdot)\right]$$

       • when $\beta < 0$,

$$e^{\beta H}\left[e^{\beta\widehat{r}_h(\cdot, \cdot)}(e^{-\beta H} - \phi(\cdot, \cdot)^\top \widehat{w}_h) + \Gamma_h(\cdot, \cdot)\right]$$

11:      $\widehat{\pi}_h(\cdot|\cdot) \leftarrow \mathrm{argmax}_{\pi_h}\langle\widehat{Q}_h(\cdot, \cdot), \pi_h(\cdot|\cdot)\rangle_\mathcal{A}$

12:      $\widehat{V}_h(\cdot) \leftarrow \frac{1}{\beta}\log\left(\langle\exp(\beta\widehat{Q}_h(\cdot, \cdot)), \widehat{\pi}_h(\cdot|\cdot)\rangle_\mathcal{A}\right)$

13: **end for**

14: **Output:** $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$

$\{z\}_{[x,y]}$ clips z into the range $[x,y]$, i.e., $\{z\}_{[x,y]} = \max\{x, \min\{z, y\}\}$. Our approach is justified in the linear MDP setting, as the transition operator $\mathbb{P}_h$ is linear in features. Therefore $\phi(s, a)^\top \widehat{w}_h$ can be viewed as the estimate of $\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s, a)$.

In Line 9, we undo the affine transformation $\mathbb{S}_h$ and combine the estimates of $r_h(s, a)$ and $\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s, a)$ into the estimate of $e^{\beta Q_h}$, or equivalently $e^{\beta r_h}\mathbb{P}_h(e^{\beta V_{h+1}})$. An uncertainty bonus of $\gamma\|\phi(s, a)\|_{\Lambda_h^{-1}}$ is subtracted for each (s, a) to enforce pessimism. The pessimistic estimates for Q, π, and V can then be obtained following Lines 9-11.

Algorithm 1 is attractive as it provably recovers a near-optimal policy even under insufficient coverage, achieving a performance guarantee in the absence of Assumption 2.2. Moreover, as we show in the sequel, under such a slightly stronger assumption on data coverage (Assumption 2.2), the algorithm can achieve a tighter dependence on d with changes to the theoretical analysis only.

4.2. Variance-Aware RSPVI

We further sharpen the suboptimality bound by incorporating variance information and propose Variance-Aware Risk-Sensitive Pessimistic Value Iteration (VA-RSPVI) in Algorithm 2. Particularly, by estimating the variances of the

value function estimates, we can reweigh the Bellman residuals to achieve a tighter bound. At a high level, the algorithm first uses Algorithm 1 as a subroutine to obtain variance estimates. A weighted ridge regression is then solved, and the Q-function, policy, and V-function estimates are obtained by solving the weighted regression problem rather than the unweighted ones used by Lines 4-6 of Algorithm 1. While the technique of utilizing the variance information has been studied in the risk-neural setting (Zhou et al., 2021a; Xiong et al., 2022; Yin et al., 2022), our work provides a non-trivial generalization of this technique to the risk-sensitive setting due to the exponential Bellman equation.

We first highlight a nuanced difference between VA-RSPVI and RSPVI. Note that we must ensure that the variance estimate is independent of the regression targets when estimating $\mathbb{S}_h\widehat{V}_h$. Otherwise, the variance estimates used to weigh the Bellman errors are correlated with the Bellman residual errors, complicating later convergence analysis. We assume the existence of auxiliary data as a "reference dataset", which is sampled from the same distribution as $\mathcal{D}$ yet independent of the dataset itself (Xie et al., 2021b; Xiong et al., 2022; Zhang et al., 2022b).

Assumption 4.1 (Auxiliary Offline Dataset). Suppose that we have another offline dataset sampled independently following the behavior policy μ as in Assumption 2.1. This auxiliary offline dataset $\mathcal{D}^{\mathrm{aux}}$ consisting of K trajectories is defined as $\mathcal{D}^{\mathrm{aux}} := \{(\breve{s}_h^\tau, \breve{a}_h^\tau, r_h(\breve{s}_h^\tau, \breve{a}_h^\tau))\}_{h,\tau=1}^{H,K}$.

We make the assumption only for ease of presentation. We note that when no such $\mathcal{D}^{\mathrm{aux}}$ is available, the learner can simply take the original dataset $\mathcal{D}$ and randomly split it by half over the trajectories, similar to the procedures used by Xie et al. (2021b); Zhang et al. (2022b). This will relax our suboptimality bound by only a negligible numerical constant factor and will not change the upper bound's dependence on dominating terms.

Variance Estimation. By the linear MDP assumption, we know that $[\mathbb{P}_h(\mathbb{S}_{h+1}V_{h+1}^*)](s, a) = \phi(s, a)^\top \eta_h^{(1)}$ and $[\mathbb{P}_h(\mathbb{S}_{h+1}V_{h+1}^*)^2](s, a) = \phi(s, a)^\top \eta_h^{(2)}$ for some $\eta_h^{(1)}$ and $\eta_h^{(2)}$. In other words, the second moment of the value function of the next step, with the expectation taken over the transition dynamics, remains linear in the feature vector $\phi(s, a)$. Using the auxiliary dataset $\mathcal{D}^{\mathrm{aux}} := \{(\breve{s}_h^\tau, \breve{a}_h^\tau, r_h(\breve{s}_h^\tau, \breve{a}_h^\tau))\}_{h,\tau=1}^{H,K}$, we can then calculate the estimate of the weights corresponding to the first moment, denoted as $\widehat{\eta}_h^{(1)}$, as the solution to the following minimization problem

$$\min_\eta \sum_{\tau=1}^K \left[\phi(\breve{s}_h^\tau, \breve{a}_h^\tau)^\top \eta - \mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}}(\breve{s}_{h+1}^\tau)\right]^2 + \lambda\|\eta\|_2^2,$$

and calculate the estimate of the weights corresponding to the second moment, denoted as $\widehat{\eta}_h^{(2)}$, as the solution to

another minimization problem

$$\min_{\eta} \sum_{\tau=1}^{K} \left[\phi(\breve{s}_h^\tau, \breve{a}_h^\tau)^\top \eta - (\mathbb{S}_{h+1} \widehat{V}_{h+1}^{\mathrm{aux}}(\breve{s}_{h+1}^\tau))^2 \right]^2 + \lambda \|\eta\|_2^2,$$

where $\widehat{V}_{h+1}^{\mathrm{aux}}$ is an estimator of V_{h+1}^*, obtained by calling Algorithm 1 as a subroutine, with $\mathcal{D}^{\mathrm{ref}}$ as the input. Doing so thus ensures that the weights $\widehat{\eta}_h^{(1)}$ and $\widehat{\eta}_h^{(2)}$ are independent of the Bellman residual errors calculated using $\mathcal{D}$, avoiding the potential dependence that may be introduced by variance estimation. Equipped with $\widehat{\eta}_h^{(1)}$ and $\widehat{\eta}_h^{(2)}$, the conditional variance of $\mathbb{S}_{h+1} V_{h+1}^*$ could then be estimated by

$$\widehat{\sigma}_h^2(\cdot, \cdot) := \max \left\{ \underline{\sigma}^2, \{\phi(\cdot, \cdot)^\top \widehat{\eta}_h^{(2)}\}_{[0,(e^{|\beta|H}-1)^2]} \right. \tag{6}$$
$$\left. - \left(\{\phi(\cdot, \cdot)^\top \widehat{\eta}_h^{(1)}\}_{[0,e^{|\beta|H}-1]} \right)^2 \right\}.$$

We clip the estimator from below by $\underline{\sigma}^2$ to avoid variance close to 0 and $\widehat{\sigma}_h^2$ is a consistent estimate of the clipped conditional variance $\sigma_h^2 := \max\{\underline{\sigma}^2, \mathrm{Var}_h(\mathbb{S}_{h+1} V_{h+1}^*)(s,a)\}$ under the conditions in Theorem 5.3.

With the variance estimators $\widehat{\sigma}_h^2(s,a)$ on hand, we develop Algorithm 2, namely Variance-Aware Risk-Sensitive Pessimistic Value Iteration, where the coefficients are constructed by weighted ridge regressions. The use of $\mathcal{D}^{\mathrm{aux}}$ in Line 3 of Algorithm 2 guarantees the independence between the variance estimate and the regression targets. Weighted by the variance estimators, $\widehat{w}_h$ and $\widehat{\theta}_h$ are the solutions of

$$\min_{\theta \in \mathbb{R}^d} \sum_{\tau=1}^{K} \frac{(r_h(s_h^\tau, a_h^\tau) - \phi_h(s_h^\tau, a_h^\tau)^\top \theta)^2}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)} + \lambda \|\theta\|_2^2,$$

and

$$\min_{w \in \mathbb{R}^d} \sum_{\tau=1}^{K} \frac{(\mathbb{S}_{h+1} \widehat{V}_{h+1}(s_{h+1}^\tau) - \phi_h(s_h^\tau, a_h^\tau)^\top w)^2}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)} + \lambda \|w\|_2^2.$$

Their closed forms are given by Lines 5-7 of Algorithm 2. The bonus function on Line 9 of Algorithm 2 changes correspondingly, and the Q-function, policy, and value function estimates are obtained accordingly on Lines 10-12.

Intuitively, by weighing each observation according to their residuals' variance, the procedure is now akin to generalized least squares with l_2 regularization (Amemiya, 1985). As we prove in the sequel, as long as the sample size is sufficiently large, the estimates' quality in Algorithm 2 will never be worse than that of Algorithm 1 under Assumption 2.2, thereby further improving our suboptimality bounds.

5. Main Results

In this section, we present our main theoretical results for both Algorithm 1 and Algorithm 2.

Suboptimality Bound for Algorithm 1. We begin with our baseline result characterizing the suboptimality of Algorithm 1. The result does not require any coverage assumptions and can be sharpened by the utilization of reference-advantage decomposition, as we will discuss later.

Theorem 5.1. *Under Assumption 2.1, if we set $\lambda = 1/e^{2|\beta|}$ and $\gamma = \widetilde{O}(d(e^{|\beta|H} - 1))$ in Algorithm 1, with probability at least $1 - \delta$, the suboptimality $\mathrm{SubOpt}(\widehat{\pi}) := V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1)$ admits an upper bound of*

$$\widetilde{O}(d) \frac{e^{|\beta|H} - 1}{|\beta|} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[\|\phi(s_h, a_h)\|_{\Lambda_h^{-1}} \Big| s_1 \right].$$

Here $\widetilde{O}(\cdot)$ omits terms that are logarithmic in d, H, K, and $1/\delta$, and the proof is deferred to the appendix. As the setting of risk-sensitive tends to the risk-neutral setting when $\beta \to 0^+$, one may achieve an upper bound $\widetilde{O}(dH) \sum_{h=1}^{H} [\|\phi(s,a)\|_{\Lambda_h^{-1}} |s_1]$, which coincides with the bound in Jin et al. (2021) up to logarithmic factors.

We stress that Theorem 5.1 does not require Assumption 2.2. We only need to assume the existence of an offline dataset and make no coverage assumption on the dataset. The coverage assumption, Assumption 2.2, can further improve our dependence on the feature dimension d by changing the strength of the uncertainty bonus γ, without any changes to the Algorithm 1. We present our result as follows:

Theorem 5.2. *Under Assumptions 2.1-2.2, with $K \geq \widetilde{\Omega}\left(d^2 H^2/\kappa + 1/\kappa^2\right)$, if we set $\lambda = 1/e^{2|\beta|}$ and $\gamma = \widetilde{O}(\sqrt{d}(e^{|\beta|H} - 1))$ in Algorithm 1, with probability at least $1 - \delta$, the suboptimality $\mathrm{SubOpt}(\widehat{\pi}) := V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1)$ admits an upper bound of*

$$\widetilde{O}(\sqrt{d}) \frac{e^{|\beta|H} - 1}{|\beta|} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[\|\phi(s_h, a_h)\|_{\Lambda_h^{-1}} \Big| s_1 \right].$$

Leveraging Assumptions 2.2, we show in Theorem 5.2 that we can improve the dependence on the feature dimension from $\widetilde{O}(d)$ to $\widetilde{O}(\sqrt{d})$ by only changing the hyperparameter choice and the analysis of the algorithm. As we detail in Section 6, this improvement is due to the reference-advantage decomposition, which, to the best of our knowledge, has not been applied in risk-sensitive RL by existing works.

Suboptimality Bound for Algorithm 2. Theorem 5.3 provides the performance bound for Algorithm 2. Our theoretical analysis is associated with a term $\xi(\underline{\sigma}^2) := \sup_{h,s,a,s' \sim \mathbb{P}_h(\cdot|s,a)} \frac{(\mathbb{S}_{h+1} V_{h+1}^*(s') - \mathbb{P}_h(\mathbb{S}_{h+1} V_{h+1}^*)(s,a))^2}{\max\{\underline{\sigma}^2, \mathrm{Var}_h(\mathbb{S}_{h+1} V_{h+1}^*)(s,a)\}}$ that characterizes the degree of deviation from the mean of $\mathbb{S}_{h+1} V_{h+1}^*(s')$ standardized by the truncated variance. In Theorem 5.3, we assume $\xi(\underline{\sigma}^2) = O(d)$. We note that Yin et al. (2022) imposed a similar condition with $\underline{\sigma}^2$ fixed at 1, while we allow the flexibility of adjusting $\underline{\sigma}^2$.

Theorem 5.3. *Under Assumption 2.1-2.2, if we have $K \geq \widetilde{\Omega}(d^2 H^2/\kappa + 1/\kappa^2) \cdot ((e^{|\beta|H} - 1)/\underline{\sigma})^4$ and $\xi(\underline{\sigma}^2) = O(d)$, setting $\lambda = 1/(e^{|\beta|(H+1)} - e^{|\beta|})^2$ and $\gamma = \widetilde{O}(\sqrt{d})$ in Algorithm 2, with probability at least $1 - \delta$, $\mathrm{SubOpt}(\widehat{\pi}) := V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1)$ admits an upper bound of*

$$\widetilde{O}(\sqrt{d}) \frac{1}{|\beta|} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[\|\phi(s_h, a_h)\|_{(\Sigma_h^*)^{-1}} \Big| s_1 \right],$$

where $\Sigma_h^* = \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top / \sigma_h^2(s_h^\tau, a_h^\tau) + \lambda I_d$ *and* $\sigma_h^2(s_h, a_h)$ *is the clipped conditional variance* $\max\{\underline{\sigma}^2, \mathrm{Var}_h(\mathbb{S}_{h+1}V_{h+1}^*)(s, a)\}$.

The condition $\xi(\underline{\sigma}^2) = O(d)$ can be ensured by using a sufficiently large $\underline{\sigma}^2$, which in turn guarantees that the magnitude of $\mathbb{S}_{h+1}V_{h+1}^*$ on the model evaluation error is negligible. A feasible choice is to set $\underline{\sigma}^2 \geq (e^{|\beta|H} - 1)^2/d$, which ensures $\xi(\underline{\sigma}^2) = O(d)$. One extreme case is $\underline{\sigma}^2 = (e^{|\beta|H} - 1)^2$. In that case, $\xi(\underline{\sigma}^2) = O(1)$ and $\widehat{\sigma}_h^2(s, a) = \underline{\sigma}^2$ for any h, s, a. In other words, in this extreme case, all targets in the regression are given the same weights and $\Sigma_h^* = \Lambda_h/\underline{\sigma}^2$. Consequently, Algorithm 2 is equivalent to Algorithm 1, and Theorem 5.3 provides the same upper bound as the one in Theorem 5.2 with appropriate choices of γ and λ.

Our bound implicitly depends on $e^{|\beta|H}$ through the scaled covariance matrix Σ_h^* and λ, as now $\lambda^{-1} = e^{2|\beta|}(e^{|\beta|H} - 1)^2$ as opposed to $e^{2|\beta|}$. Despite the implicit dependence on $e^{|\beta|H}$, as $\sigma_h^2(s_h, a_h) \leq (e^{|\beta|H} - 1)^2$ and λ is adjusted correspondingly, we have $(\Sigma_h^*)^{-1} \preceq (e^{|\beta|H} - 1)^2 \Lambda_h^{-1}$, ensuring that $\|\phi(s, a)\|_{(\Sigma_h^*)^{-1}} \leq (e^{|\beta|H} - 1) \|\phi(s, a)\|_{\Lambda_h^{-1}}$. In other words, Algorithm 2 is never worse than Algorithm 1 with appropriate parameters under Assumption 2.2.

Comparison with Existing Results. Finally, we discuss our results in the context of existing works on both risk-sensitive RL and offline RL. We begin by discussing our dependence on $\frac{e^{|\beta|H}-1}{|\beta|}$, which we dub the risk-sensitive factor. All of our bounds have a dependence on $\widetilde{O}(\frac{e^{|\beta|H}-1}{|\beta|})$. According to Fei et al. (2021a;b); Liang & Luo (2022), in the online setting, the lower bound also depends on such a factor. We can only conjecture such a dependence might exist inspired by the lower bound for the online setting. However, it remains an open question how to derive a lower bound for risk-sensitive offline RL under linear MDPs. Compared with Fei et al. (2021b), the only paper studying linear function approximation in risk-sensitive RL though for online setting, our bound can remove a potential factor of $e^{|\beta|H^2}$ by the shifting and scaling technique. Moreover, under a mild coverage assumption, our bound's dependence on d can be improved to $\widetilde{O}(\sqrt{d})$, improving over the result in Fei et al. (2021b) by a factor of $\sqrt{d}$. When compared to risk-neutral offline RL in linear MDPs, our dependence on d matches the performance guarantee in Xiong et al. (2022), the tightest

bounds for risk-neutral RL in linear MDPs that we are aware of, while also matching the lower bound for risk-neutral RL in the same paper.

6. Theoretical Analysis

In this section, we outline the analysis of our theorems. Formal proofs are deferred to the appendix. Specifically, Appendix B presents the proof of Theorem 5.1, Appendix C provides the proof of Theorem 5.2, and Appendix D presents the proof of Theorem 5.3.

6.1. Proof Sketch of Theorem 5.1

By Lemma 3.1, the bound of the suboptimality can be obtained by bounding the model evaluation error, which is controlled with the estimation error of Bellman operator as shown in Lemma B.3. Taking into account the new structure exponential Bellman equation in (4), in the risk-sensitive setting, the Bellman operators $\mathbb{B}_h$ and its estimate $\widehat{\mathbb{B}}_h$ are defined as

$$\mathbb{B}_h f = \begin{cases} e^{\beta(r_h - 1)}(\mathbb{P}_h f + e^{\beta h}), & \beta > 0 \\ e^{\beta r_h}(e^{-\beta H} - \mathbb{P}_h f), & \beta < 0 \end{cases}, \quad (7)$$

$$\widehat{\mathbb{B}}_h f = \begin{cases} e^{\beta(\widehat{r}_h - 1)}(\widehat{\mathbb{P}}_h f + e^{\beta h}), & \beta > 0 \\ e^{\beta \widehat{r}_h}(e^{-\beta H} - \widehat{\mathbb{P}}_h f), & \beta < 0 \end{cases}, \quad (8)$$

where $\widehat{\mathbb{P}}_h$ is an estimate of the transition operator. With the appropriate choice of λ, the dominating term in the estimation error $\widehat{\mathbb{B}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s, a) - \mathbb{B}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s, a)$ is $\left\| \sum_\tau \phi_h^\tau \epsilon_h^\tau (\mathbb{S}_{h+1}\widehat{V}_{h+1}) \right\|_{\Lambda_h^{-1}} \|\phi\|_{\Lambda_h^{-1}}$ where the term $\epsilon_h^\tau(\mathbb{S}_{h+1}\widehat{V}_{h+1})$ is the regression noise defined by $\mathbb{S}_{h+1}\widehat{V}_{h+1}(s_{h+1}^\tau) - \mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s_h^\tau, a_h^\tau)$, which can in turn be controlled by a uniform-concentration bound over the possible values of $\mathbb{S}_{h+1}\widehat{V}_{h+1}$. We remark that obtaining a uniform concentration of the term requires careful analysis of the covering number of the function class $\mathbb{S}_{h+1}\widehat{V}_{h+1}$ for risk-sensitive linear MDPs, which has not been examined prior to our work. Fortunately, the "shifting and scaling" operator $\mathbb{S}$ helps ensure that $\mathbb{S}_h\widehat{V}_h$ is scaled similarly at each h, streamlining our analysis. By a combination of "shifting and scaling" and uniform concentration analysis, we are able to relate the model evaluation error in risk-sensitive MDPs to the risk-neutral uncertainty quantifier $\gamma \|\phi(\cdot, \cdot)\|_{\Lambda_h^{-1}}$, which completes our proof.

6.2. Proof Sketch of Theorem 5.2

We sharpen Algorithm 1's dependence on d in Theorem 5.2 under Assumption 2.2 by better adjusting γ, yet without changing the algorithm's overall structure. An additional $\sqrt{d}$ amplification of the error upper bound is introduced in the proof of Theorem 5.1 due to the uniform concentration analysis. We further show that when the data has

sufficient coverage, the performance of offline risk-sensitive RL algorithms can benefit from reference-advantage decomposition by avoiding this amplification. The main idea is to set a fixed reference function V_{h+1}^{ref} and decompose the error of the transition $\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}) - \widehat{\mathbb{P}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})$ as $\mathbb{P}_h(\mathbb{S}_{h+1}V_{h+1}^{\text{ref}}) - \widehat{\mathbb{P}}_h(\mathbb{S}_{h+1}V_{h+1}^{\text{ref}})$ and $\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1} - \mathbb{S}_{h+1}V_{h+1}^{\text{ref}}) - \widehat{\mathbb{P}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1} - \mathbb{S}_{h+1}V_{h+1}^{\text{ref}})$.

As V_{h+1}^{ref} is fixed, we avoid the $\sqrt{d}$ amplification in the first term by eschewing uniform concentration. For the second term, if $\left\|\mathbb{S}_{h+1}\widehat{V}_{h+1} - \mathbb{S}_{h+1}V_{h+1}^{\text{ref}}\right\|_\infty \leq R_{h+1}$ for some constant R_{h+1}, the uniform concentration as in Theorem 5.1 leads to an upper bound $\widetilde{O}(dR_{h+1})\|\phi\|_{\Lambda_h^{-1}}$ which will be non-dominating as long as R_{h+1} is sufficiently small. Equipped with data coverage assumption, Assumption 2.2, the optimal value function V_h^* may serve as the reference function and $\left\|\mathbb{S}_h V_h^* - \mathbb{S}_h \widehat{V}_h\right\|_\infty$ is small enough for sufficiently large K, as the dataset now has sufficient coverage of the optimal risk-sensitive policy. Then we obtain a tighter bound on the error $\widehat{\mathbb{B}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a) - \mathbb{B}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a)$. Correspondingly, setting the parameter γ in Algorithm 1 as $\gamma = \widetilde{O}(\sqrt{d}(e^{|\beta|H} - 1))$ provides a tighter bound on the suboptimality.

6.3. Proof Sketch of Theorem 5.3

We can further sharpen the dependency on the risk-sensitivity factor $\frac{e^{|\beta|H}-1}{|\beta|}$ via variance information. Recall from the proof sketch of Theorem 5.1 that the key quantity is $\left\|\sum_\tau \frac{\phi_h^\tau}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)} \widetilde{\epsilon}_h^\tau(\mathbb{S}_{h+1}V_{h+1}^*)\right\|_{\Sigma_h^{-1}}$ with the weighted noise $\widetilde{\epsilon}_h^\tau(\mathbb{S}_{h+1}V_{h+1}^*) = \epsilon_h^\tau(\mathbb{S}_{h+1}V_{h+1}^*)/\widehat{\sigma}_h(s_h^\tau, a_h^\tau)$. Intuitively, since the variance of the noise term is not necessarily the same for all (s,a), weighing each observation accordingly improves the bound via the Gauss-Markov theorem (Amemiya, 1985). When data has sufficient coverage and the number of samples is sufficiently large, the variance can be accurately estimated, thereby improving the bounds.

We sketch our technique below. Let σ^2 and R denote the conditional variance and magnitude of $\widetilde{\epsilon}_h^\tau$. We use a Bernstein-type concentration inequality and improve the term's bound to $\widetilde{O}(\sqrt{d}\sigma + R)$, whereas the standard Hoeffding-type used to prove Theorem 5.1 yields $\widetilde{O}(\sqrt{d}R)$. Thus, as long as the estimator $\widehat{\sigma}_h^2$, defined in (6), is consistent, with an appropriate choice of $\underline{\sigma}^2$, we have $R = O(\sqrt{d})$ and $\sigma = O(1)$ leading to the upper bound in Theorem 5.3.

To show the consistency of $\widehat{\sigma}_h^2$, we first bound the estimation error between $\widehat{\sigma}_h^2$ and $\max\{\underline{\sigma}^2, \text{Var}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a)\}$ by similar technique as in Theorem 1. Then we can convert $\text{Var}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a)$ to $\text{Var}_h(\mathbb{S}_{h+1}V_{h+1}^*)(s,a)$ under the coverage assumption, since $\mathbb{S}_{h+1}\widehat{V}_{h+1}$ and $\mathbb{S}_{h+1}V_{h+1}^*$ are close enough for large K.

7. Conclusion

We study risk-sensitive offline reinforcement learning under the entropic risk measure, with a focus on the linear MDP. We begin by presenting a risk-sensitive pessimistic value iteration algorithm, offering a tight analysis by leveraging the structure of the risk-sensitive performance measure. To further improve the obtained bounds, we propose another pessimistic algorithm that utilizes variance information and reference-advantage decomposition, improving both the dependence on the space dimension d and the risk-sensitivity factor. To the best of our knowledge, we obtain the first provably efficient risk-sensitive offline RL algorithms.

Acknowledgements

Impact Statement

While motivated by practical applications, our work is theoretical in nature. Particularly, our results concern the theoretical guarantees for provably efficient offline RL algorithms and constitute generic algorithmic and theoretical contributions to reinforcement learning. In our research, we do not investigate specific application scenarios. The societal impact of our algorithms depends on how practitioners employ them and what scenarios they apply our algorithms to. Overall, this paper presents work whose goal is to advance the field of machine learning, especially from a theoretical perspective. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Agarwal, A., Henaff, M., Kakade, S., and Sun, W. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 33:13399–13412, 2020a.

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, 2020b.

Amemiya, T. *Advanced econometrics*. Harvard university press, 1985.

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, 2022.

Bäuerle, N. and Rieder, U. More risk-sensitive markov

decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

Ben-Tal, A. and Teboulle, M. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

Borkar, V. S. A sensitivity formula for risk-sensitive cost and the actor–critic algorithm. *Systems & Control Letters*, 44(5):339–346, 2001.

Borkar, V. S. Q-learning for risk-sensitive control. *Mathematics of operations research*, 27(2):294–311, 2002.

Borkar, V. S. and Meyn, S. P. Risk-sensitive optimal control for markov decision processes with monotone cost. *Mathematics of Operations Research*, 27(1):192–209, 2002.

Braun, D. A., Nagengast, A. J., and Wolpert, D. M. Risk-sensitivity in sensorimotor control. *Frontiers in human neuroscience*, 5:1, 2011.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Chaouki, A., Hardiman, S., Schmidt, C., Sérié, E., and De Lataillade, J. Deep deterministic portfolio optimization. *The Journal of Finance and Data Science*, 6:16–30, 2020.

Chateauneuf, A. and Cohen, M. Risk seeking with diminishing marginal utility in a non-expected utility model. *Journal of Risk and Uncertainty*, 9:77–91, 1994.

Chen, Y., He, J., and Gu, Q. On the sample complexity of learning infinite-horizon discounted linear kernel mdps. In *International Conference on Machine Learning*, pp. 3149–3183. PMLR, 2022.

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.

Cheng, C.-A., Xie, T., Jiang, N., and Agarwal, A. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pp. 3852–3878. PMLR, 2022.

Detlefsen, K. and Scandolo, G. Conditional and dynamic convex risk measures. *Finance and stochastics*, 9:539–561, 2005.

Di, Q., Zhao, H., He, J., and Gu, Q. Pessimistic nonlinear least-squares value iteration for offline reinforcement learning. *arXiv preprint arXiv:2310.01380*, 2023.

Di Masi, G. B. and Stettner, Ł. Infinite horizon risk sensitive control of discrete time markov processes under minorization property. *SIAM Journal on Control and Optimization*, 46(1):231–252, 2007.

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*, 2023.

Duan, Y., Jia, Z., and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pp. 2701–2709. PMLR, 2020.

Fei, Y. and Xu, R. Cascaded gaps: Towards logarithmic regret for risk-sensitive reinforcement learning. In *International Conference on Machine Learning*, pp. 6392–6417. PMLR, 2022.

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

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

Fei, Y., Yang, Z., and Wang, Z. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *International Conference on Machine Learning*, pp. 3198–3207. PMLR, 2021b.

Föllmer, H. and Schied, A. Convex measures of risk and trading constraints. *Finance and stochastics*, 6:429–447, 2002.

Hambly, B., Xu, R., and Yang, H. Recent advances in reinforcement learning in finance. *arXiv preprint arXiv:2112.04553*, 2021.

He, J., Zhou, D., and Gu, Q. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 4171–4180. PMLR, 2021.

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

Hu, P., Chen, Y., and Huang, L. Nearly minimax optimal reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 8971–9019. PMLR, 2022.

Huang, S.-H., Miao, Y.-H., and Hsiao, Y.-T. Novel deep reinforcement algorithm with adaptive sampling strategy for continuous portfolio optimization. *IEEE Access*, 9: 77371–77385, 2021.

Jaśkiewicz, A. Average optimality for risk-sensitive control with general state space. 2007.

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, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Lee, J., Park, S., and Shin, J. Learning bounds for risk-sensitive learning. *Advances in Neural Information Processing Systems*, 33:13867–13879, 2020.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260, 2024.

Liang, H. and Luo, Z.-Q. Bridging distributional and risk-sensitive reinforcement learning with provable regret bounds. *arXiv preprint arXiv:2210.14051*, 2022.

Lim, S. H. and Malik, I. Distributional reinforcement learning for risk-sensitive policies. *Advances in Neural Information Processing Systems*, 35:30977–30989, 2022.

Liu, Q., Weisz, G., György, A., Jin, C., and Szepesvári, C. Optimistic natural policy gradient: a simple efficient policy optimization framework for online rl. *Advances in Neural Information Processing Systems*, 36, 2024.

Ma, X., Xia, L., Zhou, Z., Yang, J., and Zhao, Q. Dsac: Distributional soft actor critic for risk-sensitive reinforcement learning. *arXiv preprint arXiv:2004.14547*, 2020.

Ma, Y., Jayaraman, D., and Bastani, O. Conservative offline distributional reinforcement learning. *Advances in neural information processing systems*, 34:19235–19247, 2021.

Mhammedi, Z., Block, A., Foster, D. J., and Rakhlin, A. Efficient model-free exploration in low-rank mdps. *Advances in Neural Information Processing Systems*, 36, 2024.

Mihatsch, O. and Neuneier, R. Risk-sensitive reinforcement learning. *Machine learning*, 49:267–290, 2002.

Min, Y., Wang, T., Zhou, D., and Gu, Q. Variance-aware off-policy evaluation with linear function approximation. *Advances in neural information processing systems*, 34: 7598–7610, 2021.

Modi, A., Chen, J., Krishnamurthy, A., Jiang, N., and Agarwal, A. Model-free representation learning and exploration in low-rank mdps. *Journal of Machine Learning Research*, 25(6):1–76, 2024.

Moharrami, M., Murthy, Y., Roy, A., and Srikant, R. A policy gradient algorithm for the risk-sensitive exponential cost mdp. *Mathematics of Operations Research*, 2024.

Nagengast, A. J., Braun, D. A., and Wolpert, D. M. Risk-sensitive optimal feedback control accounts for sensorimotor behavior under uncertainty. *PLoS computational biology*, 6(7):e1000857, 2010.

Nass, D., Belousov, B., and Peters, J. Entropic risk measure in policy search. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1101–1106. IEEE, 2019.

Nguyen-Tang, T., Yin, M., Gupta, S., Venkatesh, S., and Arora, R. On instance-dependent bounds for offline reinforcement learning with linear function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9310–9318, 2023.

Niv, Y., Edlund, J. A., Dayan, P., and O'Doherty, J. P. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, 32(2):551–562, 2012.

Osogami, T. Robustness and risk-sensitivity in markov decision processes. *Advances in neural information processing systems*, 25, 2012.

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

Prashanth, L. Policy gradients for cvar-constrained mdps. In *International Conference on Algorithmic Learning Theory*, pp. 155–169. Springer, 2014.

Qiu, S., Ye, J., Wang, Z., and Yang, Z. On reward-free rl with kernel and neural function approximations: Single-agent mdp and markov game. In *International Conference on Machine Learning*, pp. 8737–8747. PMLR, 2021.

Qiu, S., Wang, L., Bai, C., Yang, Z., and Wang, Z. Contrastive ucb: Provably efficient contrastive self-supervised learning in online reinforcement learning. In *International Conference on Machine Learning*, pp. 18168–18210. PMLR, 2022.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Rigter, M., Lacerda, B., and Hawes, N. One risk to rule them all: A risk-sensitive perspective on model-based offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Rudloff, B., Sass, J., and Wunderlich, R. Entropic risk constraints for utility maximization. *Festschrift in celebration of prof. Dr. Wilfried Grecksch's 60th birthday*, pp. 149–180, 2008.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Shen, Y., Stannat, W., and Obermayer, K. Risk-sensitive markov control processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672, 2013.

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

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pp. 19967–20025. PMLR, 2022.

Tamar, A., Di Castro, D., and Mannor, S. Policy gradients with variance related risk criteria. In *Proceedings of the twenty-ninth international conference on machine learning*, pp. 387–396, 2012.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR, 2016.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

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. *arXiv preprint arXiv:2102.05371*, 2021.

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., and Jamieson, K. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pp. 22384–22429. PMLR, 2022.

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

Wang, R., Du, S. S., Yang, L., and Salakhutdinov, R. R. On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 33:17816–17826, 2020a.

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020b.

Wu, Z. and Xu, R. Risk-sensitive markov decision process and learning under general utility functions. *arXiv preprint arXiv:2311.13589*, 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, 2021a.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021b.

Xiong, W., Zhong, H., Shi, C., Shen, C., Wang, L., and Zhang, T. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*, 2022.

Yang, L. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *International conference on machine learning*, pp. 6995–7004. PMLR, 2019.

Yang, Z., Jin, C., Wang, Z., Wang, M., and Jordan, M. Provably efficient reinforcement learning with kernel and neural function approximations. *Advances in Neural Information Processing Systems*, 33:13903–13916, 2020.

Yin, M. and Wang, Y.-X. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34:4065–4078, 2021.

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

Yu, P., Lee, J. S., Kulyatin, I., Shi, Z., and Dasgupta, S. Model-based deep reinforcement learning for dynamic portfolio optimization. *arXiv preprint arXiv:1901.08740*, 2019.

Zanette, A., Cheng, C.-A., and Agarwal, A. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pp. 4473–4525. PMLR, 2021a.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021b.

Zhang, R., Hu, Y., and Li, N. Regularized robust mdps and risk-sensitive mdps: Equivalence, policy gradient, and sample complexity. *arXiv preprint arXiv:2306.11626*, 2023.

Zhang, T., Ren, T., Yang, M., Gonzalez, J., Schuurmans, D., and Dai, B. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*, pp. 26447–26466. PMLR, 2022a.

Zhang, W., Zhou, D., and Gu, Q. Reward-free model-based reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34:1582–1593, 2021.

Zhang, X., Chen, Y., Zhu, X., and Sun, W. Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 5757–5773. PMLR, 2022b.

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.

Zheng, S., Wang, L., Qiu, S., Fu, Z., Yang, Z., Szepesvari, C., and Wang, Z. Optimistic exploration with learned features provably solves markov decision processes with neural dynamics. In *The Eleventh International Conference on Learning Representations*, 2022.

Zhong, H. and Zhang, T. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021a.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021b.

Zhou, H., Chen, J., Varshney, L. R., and Jagmohan, A. Nonstationary reinforcement learning with linear function approximation. *arXiv preprint arXiv:2010.04244*, 2020.

Zhou, R., Liu, M., Ren, K., Luo, X., Zhang, W., and Li, D. Is risk-sensitive reinforcement learning properly resolved? *arXiv preprint arXiv:2307.00547*, 2023.

Appendix

A. Proof of Lemma 3.1

Proof. As $\widehat{\pi}$ is the greedy policy, we have $\widehat{V}_h(s) = \frac{1}{\beta}\log(\langle e^{\beta \widehat{Q}_h(s,\cdot)}, \widehat{\pi}_h(\cdot|s)\rangle_{\mathcal{A}})$. Then, for any policy π', using $e^{\beta V_h^{\pi'}(s)} = \langle e^{\beta Q_h^{\pi'}(s,\cdot)}, \pi'_h(\cdot|s)\rangle_{\mathcal{A}} = \langle e^{\beta r_h(s,\cdot)}\mathbb{P}_h e^{\beta V_{h+1}^{\pi'}}(s,\cdot), \pi'_h(\cdot|s)\rangle_{\mathcal{A}}$ from (4), we have

$$
\begin{aligned}
e^{\beta V_h^{\pi'}(s)} - e^{\beta \widehat{V}_h(s)} =& \langle e^{\beta \widehat{Q}_h(s,\cdot)}, \pi'_h(\cdot|s) - \widehat{\pi}_h(\cdot|s)\rangle_{\mathcal{A}} + \langle \iota_{\exp,h}(s,\cdot), \pi'_h(\cdot|s)\rangle_{\mathcal{A}} \\
&+ \langle e^{\beta r_h(s,\cdot)}(\mathbb{P}_h e^{\beta V_{h+1}^{\pi'}}(s,\cdot) - \mathbb{P}_h e^{\beta \widehat{V}_{h+1}}(s,\cdot)), \pi'_h(\cdot|s)\rangle_{\mathcal{A}},
\end{aligned}
\tag{9}
$$

with $\iota_{\exp,h}(s,a) = e^{\beta r_h(s,a)}\mathbb{P}_h e^{\beta \widehat{V}_{h+1}}(s_h,a_h) - e^{\beta \widehat{Q}_h(s,a)}$. Then, we consider the case $\beta > 0$ and $\beta < 0$ separately.

Case 1: $\beta > 0$. As $\widehat{\pi}_h(\cdot|s) = \operatorname{argmax}_{\pi}\langle \widehat{Q}_h(s,\cdot), \pi(\cdot|s)\rangle_{\mathcal{A}}$, we have $\langle e^{\beta \widehat{Q}_h(s,\cdot)}, \pi'_h(\cdot|s) - \widehat{\pi}_h(\cdot|s)\rangle_{\mathcal{A}} \leq 0$. Given this property, if we set $\pi' = \pi$ in (9), we have

$$
\begin{aligned}
&e^{\beta V_h^*(s_h)} - e^{\beta \widehat{V}_h(s_h)} \\
&\leq \mathbb{E}_{\pi^*}[\iota_{\exp,h}(s_h,a_h)|s_h] + \mathbb{E}_{\pi^*}\left[e^{\beta r_h(s_h,a_h)}(\mathbb{P}_h e^{\beta V_{h+1}^*}(s_h,a_h) - \mathbb{P}_h e^{\beta \widehat{V}_{h+1}}(s_h,a_h))\Big|s_h\right].
\end{aligned}
\tag{10}
$$

If we set $\pi' = \widehat{\pi}$ in (9), we have

$$
\begin{aligned}
&e^{\beta V_h^{\widehat{\pi}}(s_h)} - e^{\beta \widehat{V}_h(s_h)} \\
&= \mathbb{E}_{\widehat{\pi}}[\iota_{\exp,h}(s_h,a_h)|s_h] + \mathbb{E}_{\widehat{\pi}}\left[e^{\beta r_h(s_h,a_h)}(\mathbb{P}_h e^{\beta V_{h+1}^{\widehat{\pi}}}(s_h,a_h) - \mathbb{P}_h e^{\beta \widehat{V}_{h+1}}(s_h,a_h))\Big|s_h\right].
\end{aligned}
\tag{11}
$$

Equipped with (10) and (11), we can show the following (12) by induction from $h = H$ to $h = 1$,

$$
\begin{aligned}
e^{\beta V_h^*(s_h)} - e^{\beta \widehat{V}_h(s_h)} &\leq \sum_{h'=h}^{H} e^{\beta(h'-h)}\mathbb{E}_{\pi^*}[\iota_{\exp,h'}(s_{h'},a_{h'})|s_h], \\
e^{\beta V_h^{\widehat{\pi}}(s_h)} - e^{\beta \widehat{V}_h(s_h)} &\geq 0.
\end{aligned}
\tag{12}
$$

To see this, we start with the base case $h = H$. As $\widehat{V}_{H+1} = V_{H+1}^{\widehat{\pi}} = V_{H+1}^* = 0$, (12) is implied by (10) and (11) directly with the assumption $\iota_{\exp,H} \geq 0$. Supposing that (10) holds for $h' \geq h+1$, we aim to show that it also holds for h. In fact, by (11), $\iota_{\exp,h} \geq 0$ and the induction assumption, we have

$$
e^{\beta V_h^{\widehat{\pi}}(s_h)} - e^{\beta \widehat{V}_h(s_h)} \geq \mathbb{E}_{\widehat{\pi}}[\iota_{\exp,h}(s_h,a_h)|s_h] \geq 0.
\tag{13}
$$

In addition, as $e^{\beta V_{h+1}^*(s_{h+1})} - e^{\beta \widehat{V}_{h+1}(s_{h+1})} \geq 0$, (10) leads to

$$
\begin{aligned}
&e^{\beta V_h^*(s_h)} - e^{\beta \widehat{V}_h(s_h)} \\
&\leq \mathbb{E}_{\pi^*}[\iota_{\exp,h}(s_h,a_h)|s_h] + e^{\beta}\mathbb{E}_{\pi^*}\left[\mathbb{P}_h e^{\beta V_{h+1}^*}(s_h,a_h) - \mathbb{P}_h e^{\beta \widehat{V}_{h+1}}(s_h,a_h)\Big|s_h\right] \\
&\leq \mathbb{E}_{\pi^*}[\iota_{\exp,h}(s_h,a_h)|s_h] + e^{\beta}\mathbb{E}_{\pi^*}\left[\sum_{h'=h+1}^{H} e^{\beta(h'-h-1)}\mathbb{E}_{\pi^*}[\iota_{\exp,h'}(s_{h'},a_{h'})|s_{h+1}]\Big|s_h\right] \\
&= \sum_{h'=h}^{H} e^{\beta(h'-h)}\mathbb{E}_{\pi^*}[\iota_{\exp,h'}(s_{h'},a_{h'})|s_h],
\end{aligned}
\tag{14}
$$

where the first inequality follows from $r_h \in [0,1]$ and the second inequality follows from the induction assumption that (12) holds for $h+1$. By induction, (13) and (14) imply that (12) holds for any $h \in [H]$. Moreover, setting $h = 1$ in (12), we have

$$
e^{\beta V_1^*(s_1)} - e^{\beta V_1^{\widehat{\pi}}(s_1)} \leq e^{\beta V_1^*(s_1)} - e^{\beta \widehat{V}_1(s_1)} \leq \sum_{h=1}^{H} e^{\beta(h-1)}\mathbb{E}_{\pi^*}[\iota_{\exp,h}(s_h,a_h)|s_1].
$$

Together with Lemma B.1, we have

$$\mathrm{SubOpt}(\widehat{\pi}) = V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1) \leq \frac{1}{\beta}\left(e^{\beta V_1^*(s_1)} - e^{\beta V_1^{\widehat{\pi}}(s_1)}\right) \leq \sum_{h=1}^{H} \frac{e^{\beta(h-1)}}{\beta} \mathbb{E}_{\pi^*}\left[\iota_{\exp,h}(s_h, a_h)|s_1\right].$$

Therefore, we conclude the proof of the case $\beta > 0$.

Case 2: $\beta < 0$. Similar to the positive β case, by setting $\pi' = \pi$ and $\pi' = \widehat{\pi}$ in (9) and the assumption $\iota_{\exp,h} \leq 0$, we have

$$e^{\beta V_h^*(s_h)} - e^{\beta \widehat{V}_h(s_h)}$$
$$\geq \mathbb{E}_{\pi^*}\left[\iota_{\exp,h}(s_h, a_h)|s_h\right] + \mathbb{E}_{\pi^*}\left[e^{\beta r_h(s_h, a_h)}(\mathbb{P}_h e^{\beta V_{h+1}^*}(s_h, a_h) - \mathbb{P}_h e^{\beta \widehat{V}_{h+1}}(s_h, a_h))\Big|s_h\right]$$

and

$$e^{\beta V_h^{\widehat{\pi}}(s_h)} - e^{\beta \widehat{V}_h(s_h)} \leq \mathbb{E}_{\widehat{\pi}}\left[e^{\beta r_h(s_h, a_h)}(\mathbb{P}_h e^{\beta V_{h+1}^{\widehat{\pi}}}(s_h, a_h) - \mathbb{P}_h e^{\beta \widehat{V}_{h+1}}(s_h, a_h))\Big|s_h\right].$$

Again, by induction from $h = H$ to $h = 1$, we can obtain a result similar to (12),

$$e^{\beta V_h^*(s_h)} - e^{\beta \widehat{V}_h(s_h)} \geq \sum_{h'=h}^{H} \mathbb{E}_{\pi^*}\left[\iota_{\exp,h'}(s_{h'}, a_{h'})|s_h\right],$$
$$e^{\beta V_h^{\widehat{\pi}}(s_h)} - e^{\beta \widehat{V}_h(s_h)} \leq 0, \tag{15}$$

and (15) leads to

$$e^{\beta V_1^*(s_1)} - e^{\beta V_1^{\widehat{\pi}}(s_1)} \geq e^{\beta V_1^*(s_1)} - e^{\beta \widehat{V}_1(s_1)} \geq \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\iota_{\exp,h}(s_h, a_h)|s_1\right].$$

Together with Lemma B.1, as $\beta < 0$, we have

$$\mathrm{SubOpt}(\widehat{\pi}) \leq \frac{e^{-\beta H}}{\beta}\left(e^{\beta V_1^*(s_1)} - e^{\beta V_1^{\widehat{\pi}}(s_1)}\right) \leq \frac{e^{-\beta H}}{\beta} \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\iota_{\exp,h}(s_h, a_h)|s_1\right].$$

The proof of the case $\beta < 0$ is concluded. $\qquad \square$

B. Proof of Theorem 5.1

In this section, we provide the detailed proofs for results attained by Algorithm 1 with $\gamma = \widetilde{O}(d(e^{|\beta|H} - 1))$. Furthermore, this section will serve as a warm-up for the subsequent section in which we sharpen the upper bound.

Proof of Theorem 5.1. We begin with the following result, which relates suboptimality to the transformed space induced by $\mathbb{S}_h$, the shifting and scaling operator.

Lemma B.1. *We have*

$$V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1) \leq \frac{1}{\beta}(e^{\beta V_1^*(s_1)} - e^{\beta V_1^{\widehat{\pi}}(s_1)}) = \frac{1}{|\beta|}(\mathbb{S}_1 V_1^*(s_1) - \mathbb{S}_1 V_1^{\widehat{\pi}}(s_1))$$

for all $\beta > 0$ and

$$V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1) \leq \frac{e^{-\beta H}}{\beta}(e^{\beta V_1^*(s_1)} - e^{\beta V_1^{\widehat{\pi}}(s_1)}) = \frac{1}{|\beta|}(\mathbb{S}_1 V_1^*(s_1) - \mathbb{S}_1 V_1^{\widehat{\pi}}(s_1))$$

for all $\beta < 0$.

The lemma then permits us to work almost entirely in the space induced by $\mathbb{S}_h$ for the rest of the proof, which is equipped with the property that $\mathbb{S}_h V_h$ is roughly of the same scale at all $h \in [H]$, thereby avoiding the need to adjust uniform concentration bounds at each step. We then define the model evaluation error after the transformation $\mathbb{S}_h$ as

$$\iota_h(s_h, a_h) = \begin{cases} e^{\beta(h-1)} \iota_{exp,h}(s_h, a_h), & \beta > 0 \\ -e^{-\beta H} \iota_{exp,h}(s_h, a_h), & \beta < 0 \end{cases}$$

The following lemma captures the relationship between suboptimality and the error. It can be viewed as a new version of Lemma 3.1 after introducing ι_h and $\mathbb{S}_h$, in the space induced by $\mathbb{S}_h$.

Lemma B.2. *If $\iota_h(s_h, a_h) \geq 0$ for all s_h, a_h, and h, then*

$$\mathbb{S}_h V_h^*(s_h) - \mathbb{S}_h \widehat{V}_h(s_h)$$
$$\leq \mathbb{E}_{\pi^*}[\iota_h(s_h, a_h)|s_h] + \mathbb{E}_{\pi^*}[\mathbb{S}_{h+1} V_{h+1}^*(s_{h+1}) - \mathbb{S}_{h+1} \widehat{V}_{h+1}(s_{h+1})|s_h], \tag{16}$$

and

$$\mathbb{S}_h V_h^{\widehat{\pi}}(s_h) - \mathbb{S}_h \widehat{V}_h(s_h) \geq 0 \tag{17}$$

Consequently, it holds that

$$\mathbb{S}_1 V_1^*(s_1) - \mathbb{S}_1 V_h^{\widehat{\pi}}(s_1) \leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\iota_h(s_h, a_h)|s_1].$$

The result then shows that if the estimate is pointwise pessimistic, in the sense that $\iota_h(s_h, a_h) \geq 0$ for all s_h, a_h, h, then the suboptimality upper bound defined in Lemma B.2 can be in turn translated to the expected model evaluation error, where expectation is taken over the state action visitation measure induced by the optimal policy. Naturally, the next step is to ensure that the condition $\iota_h(s_h, a_h) \geq 0$ holds and controlling the term from above via $\Gamma_h(s, a)$, the uncertainty bonus defined in Algorithm 1.

We first provide a more formal definition of the operators defined in (7) in the proof sketch. Particularly, let

$$\mathbb{B}_h f = \begin{cases} e^{\beta(r_h-1)}(\mathbb{P}_h f + e^{\beta h}), & \beta > 0 \\ e^{\beta r_h}(e^{-\beta H} - \mathbb{P}_h f), & \beta < 0 \end{cases} \quad \text{and} \quad \widehat{\mathbb{B}}_h f = \begin{cases} e^{\beta(\widehat{r}_h-1)}(\widehat{\mathbb{P}}_h f + e^{\beta h}), & \beta > 0 \\ e^{\beta \widehat{r}_h}(e^{-\beta H} - \widehat{\mathbb{P}}_h f), & \beta < 0 \end{cases}, \tag{18}$$

where $\widehat{\mathbb{P}}_h f(s, a) = \phi(s, a)^\top \widehat{w}(f)$ with

$$\widehat{w}(f) = \Sigma_h^{-1} \left(\sum_\tau \frac{\phi_h^\tau \cdot f(s_{h+1}^\tau)}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)} \right) \quad \text{and} \quad \Sigma_h = \sum_\tau \frac{\phi_h^\tau (\phi_h^\tau)^\top}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)} + \lambda \cdot I. \tag{19}$$

We use ϕ_h^τ to denote $\phi(s_h^\tau, a_h^\tau)$ for notation simplicity. With the definition in mind, we then show that conditioned on a so-called "good event", the term ι_h can be controlled simultaneously from above and below.

Lemma B.3. *On the event*

$$\mathcal{E}_h = \left\{ \left| \mathbb{B}_h(\mathbb{S}_{h+1} \widehat{V}_{h+1})(s, a) - \widehat{\mathbb{B}}_h(\mathbb{S}_{h+1} \widehat{V}_{h+1})(s, a) \right| \leq \Gamma_h(s, a) \right\},$$

we have $0 \leq \iota_h(s, a) \leq 2\Gamma_h(s, a)$.

Combining Lemma B.2 and B.3, we have

$$V_1^*(s_1) - \widehat{V}_1(s_1) \leq \frac{1}{|\beta|}(\mathbb{S}_h V_1^*(s_1) - \mathbb{S}_h \widehat{V}_1(s_1)) \leq \frac{2}{|\beta|} \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\Gamma_h(s_h, a_h)|s_1] \tag{20}$$

for some bonus function $\Gamma_h(s, a)$ if

$$\mathcal{E}_h = \left\{ \left| \mathbb{B}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s, a) - \widehat{\mathbb{B}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s, a) \right| \leq \Gamma_h(s, a) \right\} \tag{21}$$

holds for any h. In Algorithm 1, the bonus function is set to be $\Gamma_h(s, a) = \gamma \|\phi(s, a)\|_{\Lambda_h^{-1}}$. Then, to prove the theorem, it is sufficient to show that $\mathbb{P}(\cap_{h=1}^H \mathcal{E}_h) \geq 1 - \delta$ with $\gamma = \widetilde{O}(dR_\beta)$ where we use R_β to denote $e^{|\beta|H} - 1$. (21) suggests us to focus on the error between $\mathbb{B}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})$ and $\widehat{\mathbb{B}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})$. We first provide an upper bound on the error that is easier to control

Lemma B.4. *The following bound holds for all β, s, a, and h.*

$$\left| \mathbb{B}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s, a) - \widehat{\mathbb{B}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s, a) \right|$$
$$\leq e^{|\beta|H} |\beta| \underbrace{|r_h(s, a) - \widehat{r}_h(s, a)|}_{(i)} + \underbrace{\left| \mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s, a) - \widehat{\mathbb{P}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s, a) \right|}_{(ii)}. \tag{22}$$

Here, (i) is the error we incur when estimating the reward function r_h from data and (ii) from estimating the transition operator, $\mathbb{P}_h$, in the space induced by $\mathbb{S}_{h+1}$. Both can be controlled simultaneously by the following lemma:

Lemma B.5. *For a function $\|g_{h+1}\|_\infty \leq R$ with $\mathbb{P}_h g_{h+1}(s, a) = \phi(s, a)^\top w_h$, we have*

$$\left| \mathbb{P}_h g_{h+1}(s, a) - \widehat{\mathbb{P}}_h g_{h+1}(s, a) \right| \leq R\sqrt{d\lambda} \|\phi(s, a)\|_{\Sigma_h^{-1}}$$
$$+ \left\| \sum_\tau \frac{\phi_h^\tau}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)} \cdot \widetilde{\epsilon}_h^\tau(g_{h+1}) \right\|_{\Sigma_h^{-1}} \|\phi(s, a)\|_{\Sigma_h^{-1}},$$

where $\widetilde{\epsilon}_h^\tau(g) = \frac{g(s_{h+1}^\tau) - \mathbb{P}_h g(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)}$ and $\widehat{\mathbb{P}}_h$ is defined in (19). In addition, we have

$$|r_h(s, a) - \widehat{r}_h(s, a)| \leq \sqrt{d\lambda} \|\phi(s, a)\|_{\Sigma_h^{-1}}$$

for $\widehat{r}_h(s, a)$ given in Algorithms 1 and 2.

We note that the lemma is stated (and proven) in a more general fashion such that it can be applied to both Algorithm 1 and Algorithm 2. To specialize the result to Theorem 5.1, all we need is to set $\widehat{\sigma}_h^2(s, a) = 1$ for all s, a, and h, thereby obtaining

$$(i) = |r_h(s, a) - \widehat{r}_h(s, a)| \leq \sqrt{d\lambda} \|\phi(s, a)\|_{\Lambda_h^{-1}}. \tag{23}$$

In addition, as $\left\| \mathbb{S}_{h+1}\widehat{V}_{h+1} \right\|_\infty \leq R_\beta := e^{|\beta|H} - 1$, Lemma B.5 also leads to

$$(ii) \leq R_\beta \sqrt{d\lambda} \|\phi(s, a)\|_{\Lambda_h^{-1}} + \left\| \sum_\tau \phi_h^\tau \cdot \epsilon_h^\tau(\mathbb{S}_{h+1}\widehat{V}_{h+1}) \right\|_{\Lambda_h^{-1}} \|\phi(s, a)\|_{\Lambda_h^{-1}}, \tag{24}$$

where the first term is caused by the bias of ridge regression and $\epsilon_h^\tau(g) = g(s_{h+1}^\tau) - \mathbb{P}_h g(s_h^\tau, a_h^\tau)$. Plugging (23) and (24) into (22), we have

$$\left| \mathbb{B}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}) - \widehat{\mathbb{B}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}) \right|$$
$$\leq (R_\beta + e^{|\beta|H} |\beta|)\sqrt{d\lambda} \|\phi(s, a)\|_{\Lambda_h^{-1}} + \left\| \sum_\tau \phi_h^\tau \epsilon_h^\tau(\mathbb{S}_{h+1}\widehat{V}_{h+1}) \right\|_{\Lambda_h^{-1}} \|\phi(s, a)\|_{\Lambda_h^{-1}}$$
$$\leq 2R_\beta \sqrt{d} \|\phi(s, a)\|_{\Lambda_h^{-1}} + \left\| \sum_\tau \phi_h^\tau \epsilon_h^\tau(\mathbb{S}_{h+1}\widehat{V}_{h+1}) \right\|_{\Lambda_h^{-1}} \|\phi(s, a)\|_{\Lambda_h^{-1}}. \tag{25}$$

The last step in (25) follows the fact $\sqrt{\lambda} = e^{-|\beta|}$ and the inequality

$$e^{|\beta|H} |\beta| \leq e^{|\beta|H}(e^{|\beta|} - 1) \leq e^{|\beta|(H+1)} - e^{|\beta|} = \frac{R_\beta}{\sqrt{\lambda}}. \tag{26}$$

With (25), it is sufficient to bound $\left\| \sum_\tau \phi_h^\tau \epsilon_h^\tau (\mathbb{S}_{h+1} \widehat{V}_{h+1}) \right\|_{\Lambda_h^{-1}}$ as the first term is already $\widetilde{O}(dR_\beta) \|\phi(s,a)\|_{\Lambda_h^{-1}}$. As we do not have any coverage assumption on the offline dataset for Theorem 5.1, we do so via uniform concentration.

Uniform Concentration. In the backward iteration, $\widehat{V}_{h+1}$ depends on the data $(s_{h+1}^\tau, a_{h+1}^\tau)$ and thus a uniform concentration result for $\left\| \sum_\tau \phi_h^\tau \epsilon_h^\tau (\mathbb{S}_{h+1} \widehat{V}_{h+1}) \right\|_{\Lambda_h^{-1}}$ is needed. For brevity, in this section, we focus on the case $\beta > 0$ first and consider the following function class

$$\begin{aligned}
&\mathcal{U}_h(L_\theta, L_w, L_\gamma, \lambda) \\
&\quad = \{ U_h(s; \theta, w, \gamma, \Sigma) : \mathcal{S} \to [0, R_\beta] \text{ s.t. } \|\theta\| \leq L_\theta, \|w\| \leq L_w, \gamma \in [0, L_\gamma], \Sigma \gtrsim \lambda \cdot I \},
\end{aligned}$$

where

$$\begin{aligned}
&U_h(s; \theta, w, \gamma, \Sigma) \\
&\quad = \max_a \left\{ e^{\beta[\{\phi(s,a)^\top \theta\}_{[0,1]} - 1]}(\phi(s,a)^\top w + e^{\beta h}) - \gamma \sqrt{\phi(s,a)^\top \Sigma^{-1} \phi(s,a)} \right\}_{[e^{\beta(h-1)}, e^{\beta H}]}.
\end{aligned}$$

Given the construction of $\widehat{V}_h$ in Lines 6-9 of Algorithm 1, it is straightforward to see that $e^{\beta(\widehat{V}_h + h - 1)} \in \mathcal{U}_h(L_\theta, L_w, L_\gamma, \lambda)$. Controlling the size of $\mathcal{U}_h$ relies on controlling the norms of the estimates, which we provide in the following lemma.

Lemma B.6 (Upper bound of estimated coefficients). *$\widehat{w}_h$ and θ_h from Algorithm 1 satisfy*

$$\|\widehat{w}_h\| \leq R_\beta \sqrt{\frac{Kd}{\lambda}}, \quad \left\| \widehat{\theta}_h \right\| \leq \sqrt{\frac{Kd}{\lambda}},$$

where $R_\beta = e^{|\beta|H} - 1$. In addition, $\widehat{w}_h$ and θ_h from Algorithm 2 satisfy

$$\|\widehat{w}_h\| \leq R_\beta^2 \sqrt{\frac{Kd}{\lambda}}, \quad \left\| \widehat{\theta}_h \right\| \leq R_\beta \sqrt{\frac{Kd}{\lambda}}.$$

Using the upper bounds of $\|\widehat{w}_h\|$ and $\left\| \widehat{\theta}_h \right\|$ in Lemma B.6 and the fact that γ is set to be $\gamma = c \cdot dR_\beta \sqrt{\zeta}$ with $\zeta = \log(3dHKe^{|\beta|}/\delta)$ and some constant c, we have

$$L_\theta = \sqrt{Kd/\lambda}, \ L_w = R_\beta \sqrt{Kd/\lambda}, \ L_\gamma = c \cdot dR_\beta \sqrt{\zeta}, \ \lambda = e^{-2\beta}. \tag{27}$$

As $\mathbb{S}_h \widehat{V}_h = e^{\beta(\widehat{V}_h + h - 1)} - e^{\beta(h-1)}$, we have

$$\mathbb{S}_h \widehat{V}_h \in \widetilde{\mathcal{U}}_h := \left\{ U_h - e^{\beta(h-1)} : U_h \in \mathcal{U}_h(L_\theta, L_w, L_\gamma, \lambda) \right\}. \tag{28}$$

Let $\mathcal{N}_h(\varepsilon)$ be the ε-cover of $\widetilde{\mathcal{U}}_h$ with respect to $\|\cdot\|_\infty$ and $|\mathcal{N}_h(\varepsilon)|$ is the ϵ-covering number. Thanks to the shifting and scaling technique, the ϵ-covering number of $\mathcal{U}_h$ and $\widetilde{\mathcal{U}}_h$ should be the same, and we provide the result in the following lemma.

Lemma B.7 (Upper bound of the covering number). *For any $\varepsilon > 0$, let $|\mathcal{N}_h(\varepsilon)|$ and $|\mathcal{N}_h'(\varepsilon)|$ be the ε-covering number of the function space $\mathcal{U}_h(L_\theta, L_w, L_\gamma, \lambda)$ and $\mathcal{U}_h'(L_\theta, L_w, L_\gamma, \lambda)$ respectively, we have*

$$\begin{aligned}
&\max\{\log(|\mathcal{N}_h(\varepsilon)|), \log(|\mathcal{N}_h'(\varepsilon)|)\} \\
&\quad \leq d\log(1 + 8L_w/\varepsilon) + d\log(1 + 8|\beta|L_\theta(L_w + e^{|\beta|H})/\varepsilon) + d^2 \log(1 + 8d^{1/2}L_\gamma^2/(\lambda\varepsilon^2)).
\end{aligned}$$

Here $\mathcal{N}_h'(\varepsilon)$ refers to the covering number of the function class when $\beta < 0$. While we include its bound here for completeness, the term is defined formally in the proof of Lemma B.7 in Appendix E.7.

Setting $\varepsilon = \sqrt{\lambda} d R_\beta / K$, we have the following upper bound on the covering number.

$$
\begin{aligned}
\log|\mathcal{N}_h(\varepsilon)| &\le d\log(1 + 8L_w/\varepsilon) + d\log(1 + 8\beta L_\theta(L_w + e^{\beta H})/\varepsilon) + d^2\log(1 + 8d^{1/2}L_\gamma^2/(\lambda\varepsilon^2)) \\
&\le d\log(1 + 8e^\beta R_\beta\sqrt{Kd}/\varepsilon) \\
&\quad + d\log(1 + 16e^{3\beta}R_\beta Kd/\varepsilon) + d^2\log(1 + 8c^2e^{2\beta}R_\beta^2 d^{5/2}\zeta/\varepsilon^2) \\
&= d\log(1 + 8e^{2\beta}K^{3/2}d^{-1/2}) + d\log(1 + 16e^{4\beta}K^2) + d^2\log(1 + 8c^2e^{4\beta}d^{1/2}K^2\zeta) \\
&\le 3d^2\log(32c^2e^{4\beta}d^{1/2}K^2\zeta).
\end{aligned}
\tag{29}
$$

Here, the first step comes from Lemma B.7. The second uses the parameters in (27) and $e^{\beta H} - e^{\beta h} \le R_\beta$ and $e^\beta R_\beta \ge e^{\beta H}\beta$ from (26). The third step uses our choice of ε. The last step holds as long as $c \ge 1$.

Equipped with the upper bound of the covering number, we can obtain an upper bound of $\left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau(\mathbb{S}_{h+1}\widehat{V}_{h+1})\right\|_{\Lambda_h^{-1}}$.

Specifically, by the definition of the ε-cover, we can find $U_\varepsilon \in \mathcal{N}_{h+1}(\varepsilon)$ such that $\left\|U_\varepsilon - \mathbb{S}_{h+1}\widehat{V}_{h+1}\right\|_\infty \le \varepsilon$, then

$$
\begin{aligned}
&\left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau(\mathbb{S}_{h+1}\widehat{V}_{h+1})\right\|_{\Lambda_h^{-1}} \\
&\le \left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau(U_\varepsilon)\right\|_{\Lambda_h^{-1}} + \left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau(\mathbb{S}_{h+1}\widehat{V}_{h+1} - U_\varepsilon)\right\|_{\Lambda_h^{-1}} \\
&\le \underbrace{\sup_{U_\varepsilon' \in \mathcal{N}_{h+1}(\varepsilon)}\left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau(U_\varepsilon')\right\|_{\Lambda_h^{-1}}}_{(iii)} + \underbrace{\left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau(\mathbb{S}_{h+1}\widehat{V}_{h+1} - U_\varepsilon)\right\|_{\Lambda_h^{-1}}}_{(iv)},
\end{aligned}
\tag{30}
$$

where we use triangle inequality in the first step and $U_\varepsilon \in \mathcal{N}_{h+1}(\varepsilon)$ in the second step. We can thus control (iii) via the classic concentration bound on self-normalized processes in (Jin et al., 2021), which we write out in F.1 for completion, and have

$$
(iii) \le R_\beta\sqrt{2\log(|\mathcal{N}_{h+1}|\,H/\delta) + d\log(1 + K/\lambda)}
\tag{31}
$$

with probability $1 - \delta/H$. In addition, for (iv) we have

$$
(iv) \le \sum_{\tau=1}^K \left\|\Lambda_h^{-1/2}\phi_h^\tau\varepsilon\right\| \le K\left\|\Lambda_h^{-1/2}\right\|_2 \|\phi_h^\tau\| \cdot \varepsilon \le \frac{\varepsilon K}{\sqrt{\lambda}},
\tag{32}
$$

where $\|\cdot\|_2$ is the spectral norm of the matrix and $\|\phi_h^\tau\|$ comes from our assumption of linear MDP. Combing (29), (30), (31), (32) and $\varepsilon = \sqrt{\lambda} d R_\beta / K$, we have

$$
\begin{aligned}
&\left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau(\mathbb{S}_{h+1}\widehat{V}_{h+1})\right\|_{\Lambda_h^{-1}} \\
&\le R_\beta\sqrt{2\log(|\mathcal{N}_{h+1}|\,H/\delta) + d\log(1 + K/\lambda)} + \frac{\varepsilon K}{\sqrt{\lambda}} \\
&\le \sqrt{6}dR_\beta\sqrt{\log(32c^2e^{4\beta}d^{1/2}K^2\zeta) + \log(H/\delta) + \log(2e^{2\beta}K)} + dR_\beta \\
&\le 4dR_\beta\sqrt{7\zeta + 2\log(c)} \le \frac{c}{2}dR_\beta\sqrt{\zeta}
\end{aligned}
\tag{33}
$$

19

with probability $1 - \delta/H$. The second step comes from (29). The third step uses $\log(\zeta) \le \zeta$ and the definition of ζ. The last step holds as long as $c \ge 8\sqrt{7 + 2\log(c)/\log(2)}$. Consequently, plugging (33) into (25) shows that

$$
\begin{aligned}
\left| \mathbb{B}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}) - \widehat{\mathbb{B}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}) \right| &\le (2 + \tfrac{c}{2})dR_\beta\sqrt{\zeta}\,\|\phi(s,a)\|_{\Lambda_h^{-1}} \\
&\le cdR_\beta\sqrt{\zeta}\,\|\phi(s,a)\|_{\Lambda_h^{-1}} = \Gamma_h(s,a)
\end{aligned}
$$

holds with probability $1 - \delta/H$ if $c \ge 4$. Therefore, $\mathbb{P}(\cap_{h=1}^H \mathcal{E}_h) \ge 1 - \delta$ and (20) ensures that, on the event $\{\cap_{h=1}^H \mathcal{E}_h\}$, we have

$$
V_1^*(s_1) - \widehat{V}_1(s_1) \le \frac{2}{|\beta|}\mathbb{E}_{\pi^*}[\Gamma_h(s_h, a_h)|s_1] = \frac{e^{|\beta|H} - 1}{|\beta|} \cdot 2cd\sqrt{\zeta}\,\mathbb{E}_{\pi^*}[\|\phi(s_h, a_h)\|_{\Lambda_h^{-1}}|s_1].
$$

Thus completing the proof for Theorem 5.1 for when $\beta > 0$.

When $\beta < 0$, the only difference in the proof is that the function class $\mathcal{U}_h$ will be replaced by a slightly different function class $\mathcal{U}_h'$, whose definition we defer to (59). As shown in Lemma B.7, $\mathcal{U}_h'$ and $\mathcal{U}_h$ share the same upper bound for the covering number, and thus the above proof remains valid when $\beta < 0$ by replacing β with $|\beta|$. This completes the proof. □

C. Proof of Theorem 5.2

Proof. Note that the requisite lemmas in Appendix B remain valid for Theorem 5.2. Thus, to avoid redundancy, we focus on the reference-advantage technique under Assumption 2.2. In particular, here we show how the reference-advantage decomposition guides us to a tighter upper bound.

Without Assumption 2.2, the "good event" in Lemma B.3 requires a larger uncertainty bonus, Γ_h. As such, we instead consider the following two events, under which a factor of $\sqrt{d}$ can be shaved off of Γ_h,

$$
\begin{aligned}
\mathcal{E}_h &= \left\{ \left| \mathbb{B}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a) - \widehat{\mathbb{B}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a) \right| \le \Gamma_h(s,a) \right\}, \\
\widetilde{\mathcal{E}}_h &= \left\{ \left\| \mathbb{S}_h\widehat{V}_h - \mathbb{S}_h\widehat{V}_h \right\|_\infty \le R_h \right\}.
\end{aligned}
$$

The bonus function now is now $\Gamma_h(s,a) = \gamma\|\phi(s,a)\|_{\Lambda_h^{-1}}$ with $\gamma = 5\sqrt{d\zeta}R_\beta$ and $R_h = \frac{20\sqrt{d\zeta}R_\beta(H+1-h)}{\sqrt{K\kappa}}$ under Assumption 2.2. As suggested by (20) and (21) in the proof of Theorem 5.1, it is sufficient to show that $1 - \mathbb{P}(\cap_{h=1}^H \mathcal{E}_h) = \mathbb{P}(\cup_{h=1}^H \mathcal{E}_h^c) \le \delta$ with the new bonus function. In particular, we show a stronger result, proving that

$$
\mathbb{P}\left((\cup_{h'=h}^H \mathcal{E}_{h'}^c) \cup (\cup_{h'=h}^H \widetilde{\mathcal{E}}_{h'}^c) \right) \le \delta(H+1-h)/H \tag{34}
$$

for any $h \in [H]$ and the proof is established based on induction from $h = H$ to $h = 1$.

Base Case $h = H$. Our induction starts with the base case $h = H$. Recall that we have shown the upper bounded of the term $\left| \mathbb{B}_h(\mathbb{S}_{H+1}\widehat{V}_{H+1})(s,a) - \widehat{\mathbb{B}}_h(\mathbb{S}_{H+1}\widehat{V}_{H+1})(s,a) \right|$ in (25), by $\widehat{V}_{H+1} = V_{H+1}^* = 0$ and $\lambda = e^{-2|\beta|}$, (25) becomes

$$
\left| \mathbb{B}_H(\mathbb{S}_{H+1}\widehat{V}_{H+1})(s,a) - \widehat{\mathbb{B}}_H(\mathbb{S}_{H+1}\widehat{V}_{H+1})(s,a) \right| \le 2R_\beta\sqrt{d}\,\|\phi(s,a)\|_{\Lambda_H^{-1}} \le \Gamma_H(s,a).
$$

Therefore, $\mathbb{P}(\mathcal{E}_H) = 1$. To show that $\widetilde{\mathcal{E}}_H$ holds with high probability, we again use Lemma B.2, which gives

$$
\begin{aligned}
\mathbb{S}_h V_H^*(s_H) &- \mathbb{S}_H\widehat{V}_H(s_H) \\
&\le \mathbb{E}_{\pi^*}[\iota_H(s_H, a_H)|s_H] + \mathbb{E}_{\pi^*}[\mathbb{S}_{H+1}V_{H+1}^*(s_{H+1}) - \mathbb{S}_{H+1}\widehat{V}_{H+1}(s_{H+1})|s_H] \\
&= \mathbb{E}_{\pi^*}[\iota_H(s_H, a_H)|s_H].
\end{aligned}
$$

Conditioned on the event $\mathcal{E}_H$, we have

$$
\begin{aligned}
\mathbb{S}_H V_H^*(s_H) - \mathbb{S}_H\widehat{V}_H(s_H) &\le 2\mathbb{E}_{\pi^*}[\Gamma_H(s_H, a_H)|s_H] = 10\sqrt{d\zeta}R_\beta\mathbb{E}_{\pi^*}[\|\phi(s,a)\|_{\Lambda_H^{-1}}|s_H] \\
&\le \frac{20\sqrt{d\zeta}R_\beta}{\sqrt{K\kappa}} = R_H.
\end{aligned}
$$

with probability $1 - \delta/H$ given that $K \geq \max\{512 \log(2dH/\delta)/\kappa^2, 4\lambda/\kappa\}$. The first step comes from Lemma B.3, and the third step uses Assumption 2.2 and Lemma H.5 of (Min et al., 2021), which we restate Lemma F.3. Consequently, we have $\mathbb{P}(\widetilde{\mathcal{E}}_H^c \cup \mathcal{E}_H^c) = \mathbb{P}(\widetilde{\mathcal{E}}_H^c) \leq \delta/H$.

Induction from $h+1$ to h. In this part, we assume that (34) holds for $h+1$ and aim to show that it also holds for h. We will first show that $\mathcal{E}_h$ happens with high probability on the event $\widetilde{\mathcal{E}}_{h+1}$. Inspired by (25), it suffices to consider only $\left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau (\mathbb{S}_{h+1} \widehat{V}_{h+1})\right\|_{\Lambda_h^{-1}}$. We use the following reference-advantage decomposition to bound this term,

$$
\left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau (\mathbb{S}_{h+1} \widehat{V}_{h+1})\right\|_{\Lambda_h^{-1}} \leq \underbrace{\left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau (\mathbb{S}_{h+1} V_{h+1}^*)\right\|_{\Lambda_h^{-1}}}_{(v)}
$$
$$
+ \underbrace{\left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau (\mathbb{S}_{h+1} \widehat{V}_{h+1} - \mathbb{S}_{h+1} V_{h+1}^*)\right\|_{\Lambda_h^{-1}}}_{(vi)} . \tag{35}
$$

As V_h^* is independent of the dataset $\mathcal{D}$, the term (v) does not depend on the dataset and thus can be bounded Lemma F.1 directly. Specifically, we have, with probability $1 - \frac{\delta}{3H}$,

$$
(v) \leq R_\beta \sqrt{2\log(3H/\delta) + d\log(1 + K/\lambda)} \leq 2\sqrt{d\zeta} R_\beta. \tag{36}
$$

For the term (vi), as $\mathbb{S}_{h+1} \widehat{V}_{h+1} - \mathbb{S}_{h+1} V_{h+1}^*$ is correlated with $\mathcal{D}$, we require uniform concentration analysis similar to that in the proof of Theorem 5.1. We slightly abuse the notation and use f_{h+1} to denote $\mathbb{S}_{h+1} \widehat{V}_{h+1} - \mathbb{S}_{h+1} V_{h+1}^*$. For $\beta > 0$ and $\beta < 0$, we have different function classes for f_{h+1}. Similar to Theorem 5.1, it is sufficient to consider the positive β as the different function classes share the same upper bound of the covering number and will not affect the following proof. When $\beta > 0$, we can see that $f_{h+1} \in \widetilde{\mathcal{U}}_{h+1}^* := \{U_{h+1} - \mathbb{S}_{h+1} V_{h+1}^* : U_{h+1} \in \widetilde{\mathcal{U}}_{h+1}\}$ for $\widetilde{\mathcal{U}}_{h+1}$ defined in (28).

As $\mathbb{S}_{h+1} V_{h+1}^*$ is a fixed function of s, the ε-covering number of $\widetilde{\mathcal{U}}_{h+1}^*$ and $\widetilde{\mathcal{U}}_{h+1}$ are the same, which is $|\mathcal{N}_{h+1}(\varepsilon)|$ in (29). For notation simplicity, we thus still use $\mathcal{N}_{h+1}(\varepsilon)$ to denote the ε-cover of $\widetilde{\mathcal{U}}_{h+1}^*$. By definition, we can find $f_\varepsilon \in \mathcal{N}_{h+1}(\varepsilon)$ such that $\|f_{h+1} - f_\varepsilon\|_\infty \leq \varepsilon$, which implies that $\|f_\varepsilon\|_\infty \leq R_{h+1} + \varepsilon$ on the event $\widetilde{\mathcal{E}}_{h+1}$. Consequently, we have

$$
\left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau (f_{h+1})\right\|_{\Lambda_h^{-1}} \mathbb{1}\{\|f_{h+1}\|_\infty \leq R_{h+1}\}
$$
$$
\leq \left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau (f_\varepsilon)\right\|_{\Lambda_h^{-1}} \mathbb{1}\{\|f_\varepsilon\|_\infty \leq R_{h+1} + \varepsilon\} + \left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau (f_{h+1} - f_\varepsilon)\right\|_{\Lambda_h^{-1}} .
$$

Similar to (31) and (32), with $\varepsilon = \sqrt{\lambda d} R_{h+1}/K$, we have

$$
\left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau (f_{h+1})\right\|_{\Lambda_h^{-1}} \mathbb{1}\{\|f_{h+1}\|_\infty \leq R_{h+1}\}
$$
$$
\leq (R_{h+1} + \varepsilon) \sqrt{2\log(3H|\mathcal{N}_{h+1}(\varepsilon)|/\delta) + d\log(1 + K/\lambda)} + \frac{\varepsilon K}{\sqrt{\lambda}} \tag{37}
$$
$$
\leq 2R_{h+1} \sqrt{2\log(|\mathcal{N}_{h+1}(\varepsilon)|) + 2\log(3H/\delta) + d\log(2e^{2\beta} K)} + \sqrt{d} R_{h+1}
$$

with probability $1 - \frac{\delta}{3H}$. As $R_{h+1}/R_\beta = \frac{20\sqrt{d\zeta}(H-h)}{\sqrt{K\kappa}} \leq \frac{20\sqrt{d\zeta} H}{\sqrt{K\kappa}}$ and $\varepsilon = \sqrt{\lambda d} R_{h+1}/K$, given that $K \geq \widetilde{\Omega}(d^2 H^2/\kappa)$,

21

the upper bound in (29) becomes

$$
\begin{aligned}
\log|\mathcal{N}_h(\varepsilon)| &\le d\log(1 + 8e^{\beta}R_{\beta}\sqrt{Kd}/\varepsilon) + d\log(1 + 16e^{3\beta}R_{\beta}Kd/\varepsilon) \\
&\quad + d^2\log(1 + 8c^2e^{2\beta}R_{\beta}^2 d^{5/2}\zeta/\varepsilon^2) \\
&\le d\log\Big(\frac{320e^{2\beta}K^{3/2}H\sqrt{d\zeta}}{\sqrt{K\kappa}}\Big) + d\log\Big(\frac{640e^{4\beta}R_{\beta}K^2 Hd\sqrt{\zeta}}{\sqrt{K\kappa}}\Big) \\
&\quad + d^2\log\Big(\frac{6400c^2e^{4\beta}K^{3/2}Hd^{5/2}\zeta^2}{\sqrt{K\kappa}}\Big) \\
&\le 3d^2\log\Big(\frac{6400c^2e^{4\beta}K^2 Hd^{5/2}\zeta^2}{\sqrt{K\kappa}}\Big) \\
&\le O\Big(3d^2\log(6400c^2e^{4\beta}K^2 d^{3/2}\zeta^2)\Big) \\
&\le O\Big(3d^2(10\zeta + 2\log(c))\Big).
\end{aligned}
\tag{38}
$$

Combining (37) and (38), we have

$$
\Big\|\sum_{\tau}\phi_h^{\tau}\epsilon_h^{\tau}(f_{h+1})\Big\|_{\Lambda_h^{-1}}\mathbb{1}\{\|f_{h+1}\|_{\infty}\le R_{h+1}\} = O\Big(d\sqrt{\zeta}R_{h+1}\Big) \le O\Big(\frac{20d^{3/2}\zeta H}{\sqrt{K\kappa}}R_{\beta}\Big).
$$

When $K \ge \widetilde{\Omega}(d^2 H^2/\kappa)$, the term is smaller than $\sqrt{d}R_{\beta}$ and thus non-dominating. In other words, on the event $\widetilde{\mathcal{E}}_{h+1} = \{\|f_{h+1}\|_{\infty}\le R_{h+1}\}$, (vi) in (35) is smaller than $\sqrt{d}R_{\beta}$. Together with (36) and (25), we have

$$
\begin{aligned}
\Big|\mathbb{B}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a) - \widehat{\mathbb{B}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a)\Big| &\le (2\sqrt{d} + 2\sqrt{d\zeta} + \sqrt{d})R_{\beta}\|\phi(s,a)\|_{\Lambda_h^{-1}} \\
&\le 5\sqrt{d\zeta}R_{\beta}\|\phi(s,a)\|_{\Lambda_h^{-1}} = \Gamma_h(s,a)
\end{aligned}
$$

with probability at least $1 - \frac{2\delta}{3H}$ on the event $\widetilde{\mathcal{E}}_{h+1}$, which is equivalent to

$$
\mathbb{P}(\mathcal{E}_h^c \cap \widetilde{\mathcal{E}}_{h+1}) \le \frac{2\delta}{3H}.
\tag{39}
$$

In addition, on event $\mathcal{E}_h \cap \widetilde{\mathcal{E}}_{h+1}$, Lemma B.2, Lemma B.3, and Lemma F.3 (which we recall is a restatement of Lemma H.5 of (Min et al., 2021)) jointly imply that

$$
\begin{aligned}
\mathbb{S}_h V_h^*(s_h) - \mathbb{S}_h\widehat{V}_h(s_h) &\le 2\mathbb{E}_{\pi^*}[\Gamma_h(s_h, a_h)|s_h] + \mathbb{E}_{\pi^*}[\mathbb{S}_{h+1}V_{h+1}^*(s_{h+1}) - \mathbb{S}_{h+1}\widehat{V}_{h+1}(s_{h+1})|s_h] \\
&\le 10\sqrt{d\zeta}R_{\beta}\mathbb{E}_{\pi^*}[\|\phi(s_h, a_h)\|_{\Lambda_h^{-1}}|s_h] + R_{h+1} \\
&\le \frac{20\sqrt{d\zeta}R_{\beta}}{\sqrt{K\kappa}} + R_{h+1} = R_h
\end{aligned}
$$

with probability $1 - \frac{\delta}{3H}$ for that $K \ge \max\{512\log(6dH/\delta)/\kappa^2, 4\lambda/\kappa\}$. Therefore, we have

$$
\mathbb{P}(\widetilde{\mathcal{E}}_h^c \cap \mathcal{E}_h \cap \widetilde{\mathcal{E}}_{h+1}) \le \frac{\delta}{3H}.
\tag{40}
$$

Using (39) and (40), we have

$$
\mathbb{P}\big((\widetilde{\mathcal{E}}_h^c \cup \mathcal{E}_h^c) \cap \widetilde{\mathcal{E}}_{h+1}\big) = \mathbb{P}(\widetilde{\mathcal{E}}_h^c \cap \mathcal{E}_h \cap \widetilde{\mathcal{E}}_{h+1}) + \mathbb{P}(\mathcal{E}_h^c \cap \widetilde{\mathcal{E}}_{h+1}) \le \frac{2\delta}{3H} + \frac{\delta}{3H} = \frac{\delta}{H}.
\tag{41}
$$

Consequently, by direct calculation, we have

$$
\begin{aligned}
&\mathbb{P}\Big((\cup_{h'=h}^H \mathcal{E}_{h'}^c) \cup (\cup_{h'=h}^H \widetilde{\mathcal{E}}_{h'}^c)\Big) \\
&= \mathbb{P}\Big((\cup_{h'=h+1}^H \mathcal{E}_{h'}^c) \cup (\cup_{h'=h+1}^H \widetilde{\mathcal{E}}_{h'}^c)\Big) + \mathbb{P}\Big((\widetilde{\mathcal{E}}_h^c \cup \mathcal{E}_h^c) \cap \big((\cup_{h'=h+1}^H \mathcal{E}_{h'}^c) \cup (\cup_{h'=h+1}^H \widetilde{\mathcal{E}}_{h'}^c)\big)\Big) \\
&\le \mathbb{P}\Big((\cup_{h'=h+1}^H \mathcal{E}_{h'}^c) \cup (\cup_{h'=h+1}^H \widetilde{\mathcal{E}}_{h'}^c)\Big) + \mathbb{P}\big((\widetilde{\mathcal{E}}_h^c \cup \mathcal{E}_h^c) \cap \widetilde{\mathcal{E}}_{h+1}\big) \\
&\le \frac{\delta(H-h)}{H} + \frac{\delta}{H} = \frac{\delta(H+1-h)}{H}.
\end{aligned}
\tag{42}
$$

In the last step of (42), we use (41) and the induction assumption that (34) holds for $h+1$.

By induction, (42) shows that (34) holds for any $h \in [H]$, which implies that $1 - \mathbb{P}(\cap_{h=1}^{H} \mathcal{E}_h) = \mathbb{P}(\cup_{h=1}^{H} \mathcal{E}_h^c) \leq \delta$ with the bonus parameter $\gamma = 5\sqrt{d\zeta}R_\beta$. The rest of the proof of Theorem 5.2 then follows that of Theorem 5.1, completing the proof. $\qquad\square$

D. Proof of Theorem 5.3

Proof. A key technique used by Algorithm 2 is incorporating variance information in estimation. In this section, we highlight how the technique remains viable and beneficial in the risk-sensitive setting.

Let us recall the ingredients from Appendix B and Appendix C that we reuse. In Appendix B, we have provided the outline for proving the performance of value-iteration style algorithms in risk-averse offline RL in linear MDPs. In Appendix C, we show how Assumption 2.2 and reference-advantage decomposition sharpens the uncertainty bonus Γ_h. Finally, in this section, we show how incorporating the variance estimator ensures Algorithm 2's performance is never worse than that of Algorithm 1 under Assumption 2.2.

We start with the consistency of the variance estimator $\widehat{\sigma}_h^2(s,a)$ and then use the Bernstein-type inequality for the self-normalized process to achieve the tighter upper bound in Theorem 5.3.

Consistency of Variance Estimator. In this section, we will show that, with probability $1 - \frac{\delta}{2H}$,

$$\left| \sigma_h^2(s_h, a_h) - \widehat{\sigma}_h^2(s_h, a_h) \right| = \widetilde{O}\left(\frac{dH^2 R_\beta^2}{\sqrt{K\kappa}} \right). \tag{43}$$

As the clipping at $\underline{\sigma}^2$ is non-expansive, it is sufficient to show that

$$\left| \underbrace{\{\phi(s,a)^\top \widehat{\eta}_h^{(2)}\}_{[0,(e^{|\beta|H}-1)^2]} - \left(\{\phi(s,a)^\top \widehat{\eta}_h^{(1)}\}_{[0,e^{|\beta|H}-1]} \right)^2}_{\text{(vii)}} - \mathrm{Var}_h(\mathbb{S}_{h+1}V_{h+1}^*)(s,a) \right| = \widetilde{O}\left(\frac{dH^2 R_\beta^2}{\sqrt{K\kappa}} \right) \tag{44}$$

holds with high probability. To prove (44), we first show that $\{\phi^\top \widehat{\eta}_h^{(2)}\}_{[0,R_\beta^2]} - \left(\{\phi^\top \widehat{\eta}_h^{(1)}\}_{[0,R_\beta]} \right)^2$ is close to $\mathrm{Var}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}}) = \mathbb{P}_h((\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}})^2) - (\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}}))^2$ recalling that $R_\beta := e^{|\beta|H} - 1$. For the estimation of the first moment, that is the term $\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}})$, the proof is the same as that of Theorem 5.1 for the term (ii) in (22), which gives us

$$\left| \phi(s,a)^\top \widehat{\eta}_h^{(1)} - \mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}})(s,a) \right| = \widetilde{O}(dR_\beta)\|\phi(s,a)\|_{\Lambda_h^{-1}} = \widetilde{O}\left(\frac{dR_\beta}{\sqrt{K\kappa}} \right) \tag{45}$$

with probability $1 - \frac{\delta}{8H}$ given that $K \geq \max\{512 \log(16dH/\delta)/\kappa^2, 4\lambda/\kappa\}$, where we use Lemma F.3 in the last step. For the second moment, it is sufficient to bound $\left\| \sum_\tau \phi_h^\tau \epsilon_h^\tau ((\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}})^2) \right\|_{\Lambda_h^{-1}}$, which we control by a uniform concentration similar to that in Appendix B.

Consider the same ε-cover $\mathcal{N}_{h+1}(\varepsilon)$ as in the proof of Theorem 5.1, similar to (30) and (33), with $\varepsilon = \sqrt{\lambda}dR_\beta/K$, we have, with probability $1 - \frac{\delta}{8H}$,

$$\left\| \sum_{\tau \in \mathcal{D}^{\mathrm{aux}}} \phi_h^\tau \epsilon_h^\tau ((\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}})^2) \right\|_{\Lambda_h^{-1}} = \widetilde{O}(dR_\beta^2) + \frac{2\varepsilon K R_\beta}{\sqrt{\lambda}} = \widetilde{O}(dR_\beta^2), \tag{46}$$

where the second term in the first equation shares the same spirit of (32). Specifically, for $U_\varepsilon \in \mathcal{N}_{h+1}(\varepsilon)$ such that $\left\| U_\varepsilon - \mathbb{S}_{h+1}\widehat{V}_{h+1} \right\|_\infty \leq \varepsilon$, we have

$$\left\| \sum_{\tau \in \mathcal{D}^{\mathrm{aux}}} \phi_h^\tau \epsilon_h^\tau ((\mathbb{S}_{h+1}\widehat{V}_{h+1})^2 - U_\varepsilon^2) \right\|_{\Lambda_h^{-1}} \leq 2 \sum_{\tau=1}^{K} \|\phi_h^\tau\| \left\| \Lambda_h^{-1/2} \right\|_2 R_\beta \varepsilon \leq \frac{2\varepsilon K R_\beta}{\sqrt{\lambda}},$$

where the first step follows the fact that $\left|a^2 - b^2\right| \leq 2\max\{|a|, |b|\} \cdot |a - b|$. With (46) and Lemma B.5, a standard analysis of ridge regression leads to

$$\left|\phi(s,a)^\top \widehat{\eta}_h^{(2)} - \mathbb{P}_h((\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}})^2)(s,a)\right| = \widetilde{O}(dR_\beta^2)\left\|\phi(s,a)\right\|_{\Lambda_h^{-1}} = \widetilde{O}\left(\frac{dR_\beta^2}{\sqrt{K\kappa}}\right). \tag{47}$$

Combing (45) and (47), we have, with probability $1 - \frac{\delta}{4H}$

$$\left\|\{\phi^\top\widehat{\eta}_h^{(2)}\}_{[0,(e^{|\beta|H}-1)^2]} - \left(\{\phi^\top\widehat{\eta}_h^{(1)}\}_{[0,e^{|\beta|H}-1]}\right)^2 - \mathrm{Var}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}})\right\|_\infty$$
$$\leq \left\|\left(\{\phi^\top\widehat{\eta}_h^{(1)}\}_{[0,e^{|\beta|H}-1]}\right)^2 - (\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}}))^2\right\|_\infty + \left\|\phi^\top\widehat{\eta}_h^{(2)} - \mathbb{P}_h((\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}})^2)\right\|_\infty \tag{48}$$
$$\leq 2R_\beta\left\|\phi^\top\widehat{\eta}_h^{(1)} - \mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}})\right\|_\infty + \left\|\phi^\top\widehat{\eta}_h^{(2)} - \mathbb{P}_h((\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}})^2)\right\|_\infty = \widetilde{O}\left(\frac{dR_\beta^2}{\sqrt{K\kappa}}\right).$$

Recalling (34) in the proof of Theorem 5.2, we know $\left\|\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}} - \mathbb{S}_{h+1}V_{h+1}^*\right\|_\infty \leq R_{h+1}$ with probability $1-\delta$. Replacing δ by $\frac{\delta}{4H}$, it still holds that, with probability $1 - \frac{\delta}{4H}$,

$$\left\|\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}} - \mathbb{S}_{h+1}V_{h+1}^*\right\|_\infty = \widetilde{O}(R_{h+1}) = \widetilde{O}\left(\frac{\sqrt{d}HR_\beta}{\sqrt{K\kappa}}\right).$$

Consequently, we have

$$\left\|\mathrm{Var}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}}) - \mathrm{Var}_h(\mathbb{S}_{h+1}V_{h+1}^*)\right\|_\infty$$
$$\leq \left\|\mathbb{P}_h\left((\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}})^2 - (\mathbb{S}_{h+1}V_{h+1}^*)^2\right)\right\|_\infty + \left\|\left(\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}})\right)^2 - \left(\mathbb{P}_h(\mathbb{S}_{h+1}V_{h+1}^*)\right)^2\right\|_\infty$$
$$\leq 2R_\beta\left\|\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}} - \mathbb{S}_{h+1}V_{h+1}^*\right\|_\infty + 2R_\beta\left\|\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}^{\mathrm{aux}}) - \mathbb{P}_h(\mathbb{S}_{h+1}V_{h+1}^*)\right\|_\infty \tag{49}$$
$$=\widetilde{O}\left(\frac{\sqrt{d}HR_\beta^2}{\sqrt{K\kappa}}\right),$$

with probability $1 - \frac{\delta}{4H}$. Combing (48) and (49) completes the proof of (44).

Bernstein-Type Inequality for Self-Normalized Process. A well-known result in RL is that using a Bernstein-type concentration analysis, as opposed to a Hoeffding-type analysis, sharpens the bounds, and we show that the same technique benefits offline risk-sensitive RL in linear MDPs. Similar to the proof of Theorem 5.1 and 5.2, suggested by Lemma B.5, it is sufficient to show that

$$\underbrace{\left\|\sum_{\tau\in\mathcal{D}}\frac{\phi_h^\tau}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)}\widetilde{\epsilon}_h^\tau(\mathbb{S}_{h+1}\widehat{V}_{h+1})\right\|_{\Sigma_h^{-1}}}_{(\mathrm{viii})} = \widetilde{O}(\sqrt{d}). \tag{50}$$

where $\widetilde{\epsilon}_h^\tau(g) = \frac{g(s_{h+1}^\tau) - \mathbb{P}_h g(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)}$. Similar to (35), we use the reference-advantage decomposition an obtain

$$(\mathrm{viii}) \leq \underbrace{\left\|\sum_{\tau\in\mathcal{D}}\frac{\phi_h^\tau}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)}\widetilde{\epsilon}_h^\tau(\mathbb{S}_{h+1}V_{h+1}^*)\right\|_{\Sigma_h^{-1}}}_{(\mathrm{ix})} + \underbrace{\left\|\sum_{\tau\in\mathcal{D}}\frac{\phi_h^\tau}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)}\widetilde{\epsilon}_h^\tau(\mathbb{S}_{h+1}\widehat{V}_{h+1} - \mathbb{S}_{h+1}V_{h+1}^*)\right\|_{\Sigma_h^{-1}}}_{(\mathrm{x})}.$$

Following the proof of Theorem 5.2, we focus on the reference term (ix) first. As V_{h+1}^* is independent of the data, we can apply a Bernstein-style bound, which we detail in Lemma F.2, to (ix) directly.

To utilize Lemma F.2, we need to show that the conditional mean of $\widetilde{\epsilon}_h^\tau(\mathbb{S}_{h+1}V_{h+1}^*)$ is 0 and bound the magnitude and conditional variance of $\widetilde{\epsilon}_h^\tau(\mathbb{S}_{h+1}V_{h+1}^*)$. Specifically, we define the σ-algebra

$$\mathcal{F}_{h,\tau-1} = \sigma\big(\{(s_h^j, a_h^j)\}_{j=1, j\in\mathcal{D}}^\tau \cup \{s_{h+1}^j\}_{j=1, j\in\mathcal{D}}^{\tau-1}\big).$$

As $\widehat{\sigma}_h^2$ is calculated from the dataset $\mathcal{D}^{\mathrm{aux}}$ and therefore independent of $\mathcal{D}$, $\widetilde{\epsilon}_h^\tau(\mathbb{S}_{h+1}V_{h+1}^*)$ is $\mathcal{F}_{h,\tau}$-measurable and

$$\mathbb{E}[\widetilde{\epsilon}_h^\tau(\mathbb{S}_{h+1}V_{h+1}^*)|\mathcal{F}_{h,\tau-1}] = \frac{\mathbb{E}[\mathbb{S}_{h+1}V_{h+1}^*(s_{h+1}^\tau)|s_h^\tau, a_h^\tau] - \mathbb{P}_h(\mathbb{S}_{h+1}V_{h+1}^*)(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(s_h^\tau, a_h^\tau)} = 0. \tag{51}$$

To bound the magnitude of $\widetilde{\epsilon}_h^\tau(\mathbb{S}_{h+1}V_{h+1}^*)$, we use (43) and the assumption on K that $K \geq \widetilde{\Omega}(d^2 H^2 R_\beta^4/(\kappa\underline{\sigma}^4))$, which guarantees that the bias in (43) is smaller than $\underline{\sigma}^2/2$ for sufficiently large K and thus

$$\frac{1}{2}\sigma_h^2(s_h, a_h) \leq \sigma_h^2(s_h, a_h) - \frac{1}{2}\underline{\sigma}^2 \leq \widehat{\sigma}_h^2(s_h, a_h) \leq \sigma_h^2(s_h, a_h) + \frac{1}{2}\underline{\sigma}^2 \leq \frac{3}{2}\sigma_h^2(s_h, a_h) \leq \frac{3}{2}R_\beta^2. \tag{52}$$

Therefore, we have

$$\left|\widetilde{\epsilon}_h^\tau(\mathbb{S}_{h+1}V_{h+1}^*)\right| \leq \frac{\left|\mathbb{S}_{h+1}V_{h+1}^*(s') - \mathbb{P}_h(\mathbb{S}_{h+1}V_{h+1}^*)(s, a)\right|}{\sqrt{\frac{1}{2}\max\{\underline{\sigma}^2, \mathrm{Var}_h(\mathbb{S}_{h+1}V_{h+1}^*)(s, a)\}}} \leq \sqrt{2\xi(\underline{\sigma}^2)}, \tag{53}$$

where we let

$$\xi(\underline{\sigma}^2) = \sup_{h,s,a,s'\sim\mathbb{P}_h(\cdot|s,a)} \frac{(\mathbb{S}_{h+1}V_{h+1}^*(s') - \mathbb{P}_h(\mathbb{S}_{h+1}V_{h+1}^*)(s, a))^2}{\max\{\underline{\sigma}^2, \mathrm{Var}_h(\mathbb{S}_{h+1}V_{h+1}^*)(s, a)\}}.$$

Using (52), the conditional variance of $\widetilde{\epsilon}_h^\tau(\mathbb{S}_{h+1}V_{h+1}^*)$ can also be bounded by

$$\mathrm{Var}[\widetilde{\epsilon}_h^\tau(\mathbb{S}_{h+1}V_{h+1}^*)|\mathcal{F}_{h,\tau-1}] = \frac{\mathrm{Var}[\mathbb{S}_{h+1}V_{h+1}^*(s_{h+1}^\tau)|s_h^\tau, a_h^\tau]}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)} \leq \frac{\sigma_h^2(s_h^\tau, a_h^\tau)}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)} \leq 2. \tag{54}$$

Plugging (51), (53) and (54) into Lemma F.2, with the assumption that $\xi(\underline{\sigma}^2) = O(d)$, we have

$$(\mathrm{ix}) = \widetilde{O}(\sqrt{d})$$

holds with probability $1 - \frac{\delta}{4H}$ when (43) holds. Following the proof of Theorem 5.2, the analysis of (x) is the same as that of (vi) in (35), in which we use Lemma F.1 and B.7. With the same analysis, we have

$$\left\|\mathbb{S}_{h+1}\widehat{V}_{h+1} - \mathbb{S}_{h+1}V_{h+1}^*\right\|_\infty = \widetilde{O}\left(\frac{\sqrt{d}HR_\beta}{\sqrt{K}\kappa}\right), \text{ and } (\mathrm{x}) = \widetilde{O}\left(\frac{d^{3/2}HR_\beta}{\sqrt{K}\kappa\underline{\sigma}}\right)$$

with probability $1 - \frac{\delta}{4H}$ given that $K \geq \max\{512R_\beta^4\log(8dH/\delta)/(\underline{\sigma}^4\kappa^2), 4\lambda R_\beta/\kappa\}$. Together with the assumption that $K \geq \widetilde{\Omega}(d^2H^2R_\beta^4/(\kappa\underline{\sigma}^4))$, we have $(\mathrm{x}) = \widetilde{O}(\sqrt{d})$ and thus (50) holds for all $h \in [H]$ with probability at least $1 - H \cdot (\frac{\delta}{4H} + \frac{\delta}{4H} + \frac{\delta}{2H}) = 1 - \delta$. In other words, with Lemma B.2, Lemma B.3 and Lemma B.5, we have

$$V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1) \leq \widetilde{O}\left(\frac{\sqrt{d}}{|\beta|}\right)\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\|\phi(s_h, a_h)\|_{\Sigma_h^{-1}}\Big|s_1 = s\right], \tag{55}$$

with probability $1 - \delta$. Moreover, from (52), we have $\widehat{\sigma}_h^2 \leq \frac{3}{2}\sigma_h^2 \leq \frac{3}{2}R_\beta^2$, which implies that

$$\|\phi(s_h, a_h)\|_{\Sigma_h^{-1}} \leq \sqrt{\frac{3}{2}}\|\phi(s_h, a_h)\|_{(\Sigma_h^*)^{-1}} \leq \sqrt{\frac{3}{2}}R_\beta\|\phi(s_h, a_h)\|_{\Lambda_h^{-1}}. \tag{56}$$

Plugging (56) into (55), the proof of Theorem 5.3 is completed, and we can see that Theorem 5.3 produces a tighter bound compared to Theorem 5.2. $\qquad\square$

E. Proof of Lemmas for Main Theorems

E.1. Proof of Lemma B.1

Proof. By mean value theorem, for any x, y satisfying $0 \le y \le x \le H$,

$$e^{\beta x} - e^{\beta y} = \beta e^{\beta z}(x - y)$$

for some $z \in [x, y] \in [0, H]$. When $\beta > 0$, it implies that $e^{\beta x} - e^{\beta y} \ge \beta(x - y)$, together with $V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1) \ge 0$ from the definition of V_1^*, we have

$$V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1) \le \frac{1}{\beta}(e^{\beta V_1^*(s_1)} - e^{\beta V_1^{\widehat{\pi}}(s_1)}) = \frac{1}{|\beta|}(\mathbb{S}_1 V_1^*(s_1) - \mathbb{S}_1 V_1^{\widehat{\pi}}(s_1)).$$

When $\beta < 0$, the mean value theorem gives $e^{\beta x} - e^{\beta y} \le \beta e^{\beta H}(x - y)$ and thus

$$V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1) \le \frac{e^{-\beta H}}{\beta}(e^{\beta V_1^*(s_1)} - e^{\beta V_1^{\widehat{\pi}}(s_1)}) = \frac{1}{|\beta|}(\mathbb{S}_1 V_1^*(s_1) - \mathbb{S}_1 V_1^{\widehat{\pi}}(s_1)).$$

The proof of Lemma B.1 is thus concluded. $\square$

E.2. Proof of Lemma B.2

Proof. We divide our proof into two cases.

Case 1: $\beta > 0$. Given the definition of $\mathbb{S}_h$ and ι_h, multiplying both sides of (13) and (14) by $e^{\beta(h-1)}$, we obtain (16) and (17) directly. Using them iteratively with $V_{H+1}^* = \widehat{V}_{H+1} = V_{H+1}^{\widehat{\pi}} = 0$, we have

$$\mathbb{S}_1 V_1^*(s_1) - \mathbb{S}_1 V_1^{\widehat{\pi}}(s_1) \le \mathbb{S}_1 V_1^*(s_1) - \mathbb{S}_1 \widehat{V}_1(s_1)$$
$$\le \mathbb{E}_{\pi^*}[\iota_1(s_1, a_1)|s_1] + \mathbb{E}_{\pi^*}[\mathbb{S}_2 V_2^*(s_2) - \mathbb{S}_2 \widehat{V}_2(s_2)|s_1]$$
$$\vdots$$
$$\le \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\iota_h(s_h, a_h)|s_1].$$

Case 2: $\beta < 0$. Similarly, multiplying both sides of (13) and (14) by $-e^{\beta H}$ leads to the desired results. This completes the proof. $\square$

E.3. Proof of Lemma B.3

Proof. We divide our proof into two cases.

Case 1: $\beta > 0$. For positive β, we have

$$e^{\beta(\widehat{Q}_h + h - 1)}(s, a) = \{e^{\beta(\widehat{r}_h(s,a) - 1)}(\phi(s, a)^\top \widehat{w}_h + e^{\beta h}) - \Gamma_h(s, a)\}_{[e^{\beta(h-1)}, e^{\beta H}]}$$
$$= \{e^{\beta(\widehat{r}_h(s,a) - 1)}(\widehat{\mathbb{P}}_h(\mathbb{S}_{h+1} \widehat{V}_{h+1}) + e^{\beta h}) - \Gamma_h(s, a)\}_{[e^{\beta(h-1)}, e^{\beta H}]}$$

and

$$\iota_h(s, a) = e^{\beta(r_h(s,a) - 1)}(\mathbb{P}_h(\mathbb{S}_{h+1} \widehat{V}_{h+1})(s, a) + e^{\beta h}) - e^{\beta(\widehat{Q}_h + h - 1)}(s, a).$$

We show that $\iota_h(s, a) \ge 0$ first. If $e^{\beta(\widehat{Q}_h + h - 1)}(s, a) \le e^{\beta(h-1)}$, it is straight forward that $\iota_h(s, a) \ge 0$ as $\widehat{r}_h(s, a) \in [0, 1]$ and $\mathbb{S}_{h+1} \widehat{V}_{h+1} \ge e^{\beta h}$. Otherwise, by the definition of $\mathbb{B}_h$ and $\widehat{\mathbb{B}}_h$ in (18), on the event $\mathcal{E}_h$, we have

$$e^{\beta(\widehat{Q}_h + h - 1)}(s, a) \le e^{\beta(\widehat{r}_h(s,a) - 1)}(\widehat{\mathbb{P}}_h(\mathbb{S}_{h+1} \widehat{V}_{h+1})(s, a) + e^{\beta h}) - \Gamma_h(\cdot, \cdot)$$
$$\le e^{\beta(r_h(s,a) - 1)}(\mathbb{P}_h(\mathbb{S}_{h+1} \widehat{V}_{h+1})(s, a) + e^{\beta h}).$$

Equivalent, we have $\iota_h(s, a) \geq 0$. For the upper bound of ι_h, we have

$$e^{\beta(\hat{r}_h(s,a)-1)}\big(\widehat{\mathbb{P}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a) + e^{\beta h}\big) - \Gamma_h(s,a)$$
$$\leq e^{\beta(r_h(s,a)-1)}\big(\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a) + e^{\beta h}\big)$$
$$\leq e^{\beta H}$$

on $\mathcal{E}_h$, which leads to

$$e^{\beta(\widehat{Q}_h+h-1)}(s,a) \geq e^{\beta(\hat{r}_h(s,a)-1)}\big(\widehat{\mathbb{P}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a) + e^{\beta h}\big) - \Gamma_h(s,a)$$

and thus $\iota_h(s, a) \leq 2\Gamma_h(s, a)$.

Case 2: $\beta < 0$. Similarly, following the above proof, the proof for negative β can be established by

$$e^{\beta(\widehat{Q}_h-H)}(s,a) = \{e^{\beta\hat{r}_h(s,a)}(e^{-\beta H} - \widehat{\mathbb{P}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})) + \Gamma_h(s,a)\}_{[e^{-\beta(h-1)}, e^{-\beta H}]}$$

and

$$\iota_h(s,a) = e^{\beta r_h(s,a)}(e^{-\beta H} - \mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a)) - e^{\beta(\widehat{Q}_h-H)}(s,a).$$

This completes the proof. $\qquad\square$

E.4. Proof of Lemma B.5

Proof. This is a standard result for ridge regression. From the definition of $\widehat{\mathbb{P}}_h$, we have

$$\mathbb{P}_h g_{h+1}(s,a) - \widehat{\mathbb{P}}_h g_{h+1}(s,a)$$
$$= \phi(s,a)^\top w_h - \phi(s,a)^\top \Sigma_h^{-1}(\Sigma_h - \lambda \cdot I)w_h + \phi(s,a)^\top \Sigma_h^{-1}(\Sigma_h - \lambda \cdot I)w_h$$
$$- \phi(s,a)^\top \Sigma_h^{-1}\left(\sum_\tau \frac{\phi_h^\tau \cdot g_{h+1}(s_{h+1}^\tau)}{\hat{\sigma}_h^2(s_h^\tau, a_h^\tau)}\right)$$
$$= \lambda\phi(s,a)^\top \Sigma_h^{-1} w_h + \phi(s,a)^\top \Sigma_h^{-1}\left(\sum_\tau \frac{\phi_h^\tau \cdot (\mathbb{P}_h g_{h+1}(s_h^\tau, a_h^\tau) - g_{h+1}(s_{h+1}^\tau))}{\hat{\sigma}_h^2(s_h^\tau, a_h^\tau)}\right),$$

which implies that

$$\left|\mathbb{P}_h g_{h+1}(s,a) - \widehat{\mathbb{P}}_h g_{h+1}(s,a)\right|$$
$$\leq \lambda \|w_h\|_{\Sigma_h^{-1}} \|\phi(s,a)\|_{\Sigma_h^{-1}} + \left\|\sum_\tau \frac{\phi_h^\tau}{\hat{\sigma}_h(s_h^\tau, a_h^\tau)} \cdot \widetilde{\epsilon}_h^\tau(g_{h+1})\right\|_{\Sigma_h^{-1}} \|\phi(s,a)\|_{\Sigma_h^{-1}}. \tag{57}$$

As $\|w_h\| = \left\|\int_{\mathcal{S}} g_{h+1}(s')\mu_h(s')ds'\right\| \leq R\sqrt{d}$ from the assumptions of linear MDP, we have

$$\lambda \|w_h\|_{\Sigma_h^{-1}} = \lambda\left\|\Sigma_h^{-1/2}w_h\right\| \leq \lambda\left\|\Sigma_h^{-1/2}\right\|_2 \|w_h\| \leq R\sqrt{d\lambda}.$$

Plugging this inequality into (57), we obtain the desired result for $\left|\mathbb{P}_h g_{h+1}(s,a) - \widehat{\mathbb{P}}_h g_{h+1}(s,a)\right|$. The proof of $|r_h(s,a) - \hat{r}_h(s,a)|$ can be viewed as the case that $\widetilde{\epsilon}_h^\tau(\cdot) = 0$ and is omitted here. This completes the proof. $\qquad\square$

E.5. Proof of Lemma B.4

Proof. Again we divide the proof into the following two cases.

Case 1: $\beta > 0$. By the definition in (18), we have

$$\left|e^{\beta(r_h(s,a)-1)}(\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a) + e^{\beta h}) - e^{\beta(\hat{r}_h(s,a)-1)}(\widehat{\mathbb{P}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a) + e^{\beta h})\right|$$
$$\leq e^{\beta H}\left|e^{\beta(r_h(s,a)-1)} - e^{\beta(\hat{r}_h(s,a)-1)}\right| + \left|\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}) - \widehat{\mathbb{P}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})\right|$$
$$\leq e^{\beta H}\beta |r_h - \hat{r}_h| + \left|\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1}) - \widehat{\mathbb{P}}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})\right|.$$

The first inequality follows $\left|\mathbb{P}_h(\mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a) + e^{\beta h}\right| \le e^{\beta H}$, $r_h(s,a) \in [0,1]$ and the triangular inequality. The second inequality follows the fact that $\left|e^{\beta x} - e^{\beta y}\right| \le \beta\,|x - y|$ for $\beta > 0$ and $-1 \le x, y \le 0$, which can be obtained by the mean value theorem as in the proof of Lemma B.1.

Case 2: $\beta < 0$. The proof of negative β is similar with the fact that $\left|\mathbb{P}_h(e^{-\beta H} - \mathbb{S}_{h+1}\widehat{V}_{h+1})(s,a)\right| \le e^{\beta H}$ and $r_h(s,a) \in [0,1]$.

This completes the proof. $\qquad\square$

E.6. Proof of Lemma B.6

Proof. By the definition of $\widehat{w}_h$ in Algorithm 2, we have

$$\|\widehat{w}_h\| = \left\|\Sigma_h^{-1}\left(\sum_{\tau=1}^{K}\frac{\phi_h^\tau \cdot \mathbb{S}_{h+1}\widehat{V}_{h+1}(s_{h+1}^\tau)}{\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau)}\right)\right\| \le \sum_{\tau=1}^{K}\left\|\Sigma_h^{-1}\phi_h^\tau \cdot \mathbb{S}_{h+1}\widehat{V}_{h+1}(s_{h+1}^\tau)\right\|.$$

Note that $\left|\mathbb{S}_{h+1}\widehat{V}_{h+1}(s_{h+1}^\tau)\right| \le R_\beta$, we have

$$\|\widehat{w}_h\| \le R_\beta \sum_{\tau=1}^{K}\left\|\Sigma_h^{-1/2}\right\|\|\phi_h^\tau\|_{\Lambda_h^{-1}} \le \frac{R_\beta}{\sqrt{\lambda}}\sum_{\tau=1}^{K}\|\phi_h^\tau\|_{\Lambda_h^{-1}} = \frac{R_\beta}{\sqrt{\lambda}}\sum_{\tau=1}^{K}\sqrt{(\phi_h^\tau)^\top \Sigma_h^{-1}\phi_h^\tau}.$$

By the Cauchy-Schwarz inequality, we have

$$\|\widehat{w}_h\| \le \frac{R_\beta\sqrt{K}}{\sqrt{\lambda}}\sqrt{\sum_{\tau=1}^{K}(\phi_h^\tau)^\top \Sigma_h^{-1}\phi_h^\tau} = \frac{R_\beta\sqrt{K}}{\sqrt{\lambda}}\sqrt{\mathrm{Tr}\left(\Sigma_h^{-1}\sum_{\tau=1}^{K}(\phi_h^\tau)^\top \phi_h^\tau\right)}. \tag{58}$$

Weights $\widehat{w}_h$ in Algorithm 1. If $\widehat{\sigma}_h = 1$, we are using the standard ridge regression without variance information, then $\widehat{w}$ becomes that in Algorithm 1. In this case, $\Sigma_h = \sum_{\tau=1}^{K}(\phi_h^\tau)^\top \phi_h^\tau + \lambda \cdot I$ and (58) leads to

$$\|\widehat{w}_h\| \le \frac{R_\beta\sqrt{K}}{\sqrt{\lambda}}\sqrt{\mathrm{Tr}\left(\Sigma_h^{-1}(\Sigma_h - \lambda \cdot I)\right)} \le \frac{R_\beta\sqrt{K}}{\sqrt{\lambda}}\sqrt{\mathrm{Tr}\left(\Sigma_h^{-1}\Sigma_h\right)} = R_\beta\sqrt{Kd/\lambda}.$$

Weights $\widehat{w}_h$ in Algorithm 2. In Algorithm 2, as $\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau) \le R_\beta^2$, we have

$$\Sigma_h = \sum_{\tau=1}^{K}(\phi_h^\tau)^\top \phi_h^\tau/\widehat{\sigma}_h^2(s_h^\tau, a_h^\tau) + \lambda \cdot I \gtrsim \sum_{\tau=1}^{K}(\phi_h^\tau)^\top \phi_h^\tau/R_\beta^2 + \lambda \cdot I = \frac{1}{R_\beta^2}\Phi_h + \lambda \cdot I,$$

where $\Phi_h = \sum_{\tau=1}^{K}(\phi_h^\tau)^\top \phi_h^\tau$. Then, by (58), we have

$$\|\widehat{w}_h\| \le \frac{R_\beta\sqrt{K}}{\sqrt{\lambda}}\sqrt{\mathrm{Tr}\left(\left(\frac{1}{R_\beta^2}\Phi_h + \lambda \cdot I\right)^{-1}\Phi_h\right)} \le R_\beta^2\sqrt{Kd/\lambda}.$$

The proof of $\widehat{\theta}_h$ is similar. With $r_h(s,a) \in [0,1]$, we have the desired result in Lemma B.6. This completes the proof. $\quad\square$

E.7. Proof of Lemma B.7

Proof. We first more formally introduce the corresponding function classes and covering numbers for when $\beta < 0$. As $\widehat{\mathbb{B}}_h$ will be different for $\beta > 0$ and $\beta < 0$, we consider two corresponding function classes.

Case 1: $\beta > 0$. We consider the function class

$$\begin{aligned}&\mathcal{U}_h(L_\theta, L_w, L_\gamma, \lambda)\\&\quad = \{U_h(s; \theta, w, \gamma, \Sigma) : \mathcal{S} \to [0, R_\beta] \text{ with } \|\theta\| \le L_\theta, \|w\| \le L_w, \gamma \in [0, L_\gamma], \Sigma \gtrsim \lambda \cdot I\},\end{aligned}$$

where

$$
\begin{aligned}
& U_h(s; \theta, w, \gamma, \Sigma) \\
& = \max_a \left\{ e^{\beta[\{\phi(s,a)^\top \theta\}_{[0,1]} - 1]} (\phi(s,a)^\top w + e^{\beta h}) - \gamma \sqrt{\phi(s,a)^\top \Sigma^{-1} \phi(s,a)} \right\}_{[e^{\beta(h-1)}, e^{\beta H}]}.
\end{aligned}
$$

Case 2: $\beta < 0$. We consider the function class

$$
\begin{aligned}
& \mathcal{U}'_h(L_\theta, L_w, L_\gamma, \lambda) \\
& = \{ U'_h(s; \theta, w, \gamma, \Sigma) : \mathcal{S} \to [0, R_\beta] \text{ with } \|\theta\| \leq L_\theta, \|w\| \leq L_w, \gamma \in [0, L_\gamma], \Sigma \gtrsim \lambda \cdot I \},
\end{aligned}
$$

where

$$
\begin{aligned}
& U'_h(s; \theta, w, \gamma, \Sigma) \\
& = \max_a \left\{ e^{\beta\{\phi(s,a)^\top \theta\}_{[0,1]}} (e^{-\beta H} - \phi(s,a)^\top w) + \gamma \sqrt{\phi(s,a)^\top \Sigma^{-1} \phi(s,a)} \right\}_{[e^{\beta(h-1)}, e^{\beta H}]}.
\end{aligned} \tag{59}
$$

Even though the representation of $\mathcal{U}_h$ and $\mathcal{U}'_h$ are slightly different.

We now provide bounds for the covering numbers, again dividing the proof into two cases.

Case 1: $\beta > 0$. For any two functions U_1 and U_2 from $\mathcal{U}$, let them take parameters $(\theta_1, w_1, \gamma_1, \Sigma_1)$ and $(\theta_2, w_2, \gamma_2, \Sigma_2)$, respectively. As $\{\cdot\}_{a,b}$ and $\max_a$ are contraction map, we have

$$
\begin{aligned}
\|U_1 - U_2\|_\infty &\leq \underbrace{\sup_{\|\phi\| \leq 1} \left| e^{\beta[\{\phi^\top \theta_1\}_{[0,1]} - 1]} (\phi^\top w_1 + e^{\beta h}) - e^{\beta[\{\phi^\top \theta_2\}_{[0,1]} - 1]} (\phi^\top w_2 + e^{\beta h}) \right|}_{(a)} \\
& \quad + \sup_{\|\phi\| \leq 1} \left| \sqrt{\gamma_1^2 \phi^\top \Sigma_1^{-1} \phi} - \sqrt{\gamma_2^2 \phi^\top \Sigma_2^{-1} \phi} \right| \\
&\leq (L_w + e^{\beta h}) \beta \sup_{\|\phi\| \leq 1} \left| \phi^\top \theta_1 - \phi^\top \theta_2 \right| + \sup_{\|\phi\| \leq 1} \left| \phi^\top w_1 - \phi^\top w_2 \right| \\
& \quad + \sup_{\|\phi\| \leq 1} \left| \sqrt{\gamma_1^2 \phi^\top \Sigma_1^{-1} \phi} - \sqrt{\gamma_2^2 \phi^\top \Sigma_2^{-1} \phi} \right| \\
&\leq \beta(L_w + e^{\beta H}) \|\theta_1 - \theta_2\| + \|w_1 - w_2\| + \sqrt{\left\| \gamma_1^2 \Sigma_1^{-1} - \gamma_2^2 \Sigma_2^{-1} \right\|_F}.
\end{aligned} \tag{60}
$$

The second step follows from $\left| \phi^\top w_1 + e^{\beta h} \right| \leq L_w + e^{\beta H}$, $e^{\beta[\{\phi^\top \theta_1\}_{[0,1]} - 1]} \leq 1$ and the fact that $\left| e^{\beta x} - e^{\beta y} \right| \leq \beta |x - y|$ for $\beta > 0$ and $-1 \leq x, y \leq 0$. Let $\mathcal{C}_w$ be the $\varepsilon/4$-cover of $\{w \in \mathbb{R}^d : \|w\| \leq L_w\}$, $\mathcal{C}_\theta$ be the $\varepsilon/(4\beta(L_w + e^{\beta H}))$-cover of $\{\theta \in \mathbb{R}^d : \|\theta\| \leq L_\theta\}$ and $\mathcal{C}_A$ be the $\varepsilon^2/4$-cover of $\{A \in \mathbb{R}^{d \times d} : \|A\|_F \leq d^{1/2} L_\gamma^2 \lambda^{-1}\}$. As we have $\left\| \gamma^2 \Sigma^{-1} \right\|_F \leq d^{1/2} L_\gamma^2 \lambda^{-1}$, we can see that $\mathcal{C}_w$, $\mathcal{C}_\theta$ and $\mathcal{C}_A$ provide a ε−cover of $\mathcal{U}_h$. Together with Lemma F.4 for the covering number of the Euclidean ball, we have

$$
\log \mathcal{N}(\varepsilon) \leq d \log(1 + 8L_w/\varepsilon) + d \log(1 + 8\beta L_\theta(L_w + e^{\beta H})/\varepsilon) + d^2 \log(1 + 8d^{1/2} L_\gamma^2 / (\lambda \varepsilon^2)).
$$

Case 2: $\beta < 0$. In the proof of negative β, the term (a) in (60) becomes

$$
\sup_{\|\phi\| \leq 1} \left| e^{\beta\{\phi^\top \theta_1\}_{[0,1]}} (e^{-\beta H} - \phi^\top w_1) - e^{\beta\{\phi^\top \theta_2\}_{[0,1]}} (e^{-\beta H} - \phi^\top w_2) \right|.
$$

As $\left| e^{-\beta H} - \phi^\top w_1 \right| \leq e^{|\beta| H} + L_w$ and $e^{\beta\{\phi^\top \theta_1\}_{[0,1]}} \leq 1$, the second step in (60) remains valid if we replace β by $|\beta|$. Therefore, we will reach the same upper bound for $\mathcal{N}'_h(\varepsilon)$. This completes the proof. $\square$

F. Additional Technical Lemmas

In this section, we provide technical lemmas which are widely used in theoretical reinforcement learning. The proofs of them are omitted and we refer interested readers to the cited sources.

Lemma F.1 (Lemma B.2 of (Jin et al., 2021)). *Let $f : \mathcal{S} \to [0, R]$ be any fixed function, for any $\delta \in (0, 1)$, we have*

$$\mathbb{P}\left(\left\|\sum_\tau \phi_h^\tau \epsilon_h^\tau(f)\right\|_{\Lambda_h^{-1}}^2 \geq R^2(2\log(1/\delta) + d\log(1 + K/\lambda))\right) \leq \delta.$$

Lemma F.2 (Bernstein-type inequality for self-normalized process in (Zhou et al., 2021a)). *Let $\{\eta_t\}_{t=1}^\infty$ be a real-valued stochastic process and let $\{\mathcal{F}_t\}_{t=1}^\infty$ be a filtration such that η_t is $\mathcal{F}_t$-measurable. Let $\{x_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process where x_t is $\mathcal{F}_{t-1}$ measurable and $\|x_t\| \leq L$. Let $\Lambda_t = \lambda I_d + \sum_{s=1}^t x_s x_s^\top$. Assume that*

$$|\eta_t| \leq R, \ \mathbb{E}[\eta_t|\mathcal{F}_{t-1}] = 0, \ \mathbb{E}[\eta_t^2|\mathcal{F}_{t-1}] \leq \sigma^2.$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t > 0$, we have

$$\left\|\sum_{s=1}^t x_s \eta_s\right\|_{\Lambda_h^{-1}} \leq 8\sigma\sqrt{d\log\left(1 + \frac{tL^2}{\lambda d}\right) \cdot \log\left(\frac{4t^2}{\delta}\right)} + 4R\log\left(\frac{4t^2}{\delta}\right) = \widetilde{O}(\sigma\sqrt{d} + R).$$

Lemma F.3 (Lemma H.5 of (Min et al., 2021)). *Let $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ satisfying $\|\phi(s, a)\| \leq C$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. For any $K > 0$ and $\lambda > 0$, define $\bar{\mathbb{G}}_K = \sum_{k=1}^K \phi(s_k, a_k)\phi(s_k, a_k)^\top + \lambda I_d$ where (s_k, a_k)'s are i.i.d, samples from some distribution ν over $\mathcal{S} \times \mathcal{A}$. Let $\mathbb{G} = \mathbb{E}_\nu[\phi(s, a)\phi(s, a)^\top]$. Then, for any $\delta \in (0, 1)$, if K satisfies that*

$$K \geq \max\left\{512C^4 \left\|\mathbb{G}^{-1}\right\|^2 \log\left(\frac{2d}{\delta}\right), 4\lambda \left\|\mathbb{G}^{-1}\right\|\right\},$$

with probability at least $1 - \delta$, it holds simultaneously for all $u \in \mathbb{R}^d$ that

$$\|u\|_{\bar{\mathbb{G}}_\mathbb{K}^{-1}} \leq \frac{2}{\sqrt{K}}\|u\|_{\mathbb{G}^{-1}}.$$

Lemma F.4 (Lemma D.5 of (Jin et al., 2020)). *For any $\varepsilon > 0$, the ε-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is upper bounded by $(1 + 2R/\varepsilon)^d$.*

G. Numerical Simulations

For completeness, we examine a variant of the ModelWin MDP introduced in Thomas & Brunskill (2016) to verify theoretical findings. This particular instance contains 3 states (S_1, S_2, S_3) and 2 actions (a_1, a_2). Each episode starts from the state S_1, where the agent must choose between two actions. Action a_1 transitions the agent to either S_2 or S_3 with probability 0.5 to each state. In contrast, the second action, a_2, causes the agent to stay at S_1 with probability 0.6; otherwise, the agent will transit to S_2 or S_3 with equal probability. In states S_2 and S_3, the agent still has two possible actions, but both always produce a deterministic transition back to S_1. The rewards are state-dependent, with S_1, S_2, and S_3 yielding 0.5, 1, and 0, respectively. Formally,

$$\mathbb{P}(S_2|S_1, a_1) = \mathbb{P}(S_3|S_1, a_1) = 0.5, \quad \mathbb{P}(S_1|S_1, a_1) = 0$$
$$\mathbb{P}(S_2|S_1, a_2) = \mathbb{P}(S_3|S_1, a_2) = 0.2, \quad \mathbb{P}(S_1|S_1, a_2) = 0.6$$
$$\mathbb{P}(S_1|S_3, a) = \mathbb{P}(S_1|S_2, a) = 1 \text{ for } a = a_1 \text{ or } a_2.$$

and

$$r(S_1, a) = 0.5, r(S_2, a) = 1, r(S_1, a) = 0 \text{ for } a = a_1 \text{ or } a_2.$$

In this MDP, the expectation of the total reward is consistently $\mathbb{E}(\sum_{h=1}^H r_h) = 0.5H$, indicating that all policies are equally optimal when $\beta = 0$. Consequently, all policies have a suboptimality of 0. However, this MDP is intriguing if we consider how policy choices can impact the reward variance. Staying at S_1 guarantees a 0.5 reward, but leaving S_1 yields a 0.5 chance of obtaining either 1 or 0. One can imagine that a_1 is a more risky action compared to a_2 as a_2 is more likely to keep the agent at S_1. The behavior policy we use to generate the offline data is taking a_1 and a_2 randomly with equal

probability. Considering the entropic risk measure we study, the agent is risk-seeking when $\beta > 0$ and risk-averse when $\beta < 0$. Therefore, the optimal policy will be consistently taking action a_1 when $\beta > 0$ and consistently taking a_2 when $\beta < 0$. We take our first algorithm as an example and conduct experiments using the above environment based on this algorithm.

We evaluate the scenarios $H = 5, 10, 15, 20$ and $\beta = 0.5, 1$ in the experiment. The suboptimality results are reported in Figure 1. We can see that with a larger K, the suboptimality goes to 0, which serves as simulation evidence for our algorithm. Moreover, Figure 1 illustrates that with an increase in H and $|\beta|$, there is a corresponding rise in the suboptimality gap, aligning with our theoretical result.
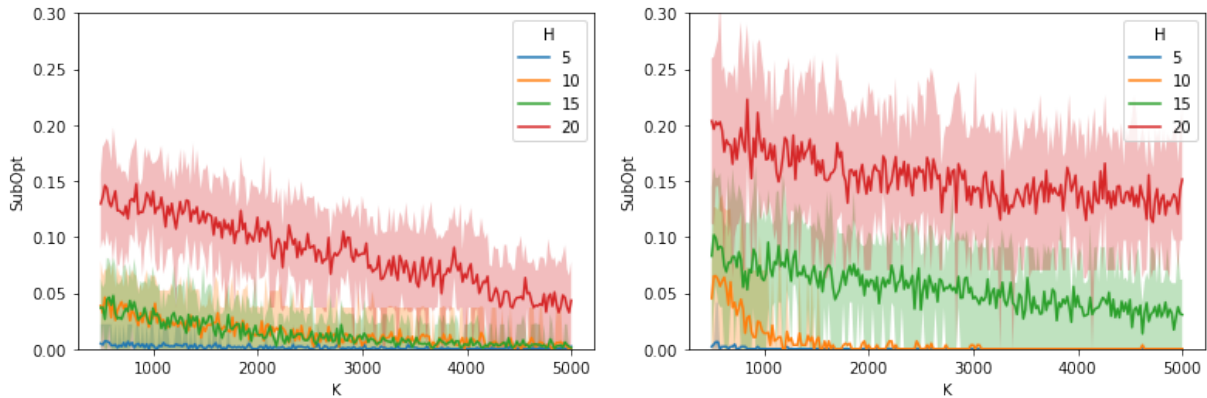

Figure 1. Each panel reports the suboptimality of the learned policy from Algorithm 1 for different K and h. $\beta = 0.5$ (left) and $\beta = 1$ (right). The results are averaged over 20 independent trails, and the mean results are plotted as solid lines. The error bar area corresponds to the 80% confidence interval.