Towards Robust Model-Based Reinforcement Learning Against Adversarial Corruption

Chenlu Ye [*1] **Jiafan He** [*2] **Quanquan Gu** [2] **Tong Zhang** [3]

Abstract

This study tackles the challenges of adversarial corruption in model-based reinforcement learning (RL), where the transition dynamics can be corrupted by an adversary. Existing studies on corruption-robust RL mostly focus on the setting of model-free RL, where robust least-square regression is often employed for value function estimation. However, the uncertainty weighting techniques cannot be directly applied to model-based RL. In this paper, we focus on model-based RL and take the maximum likelihood estimation (MLE) approach to learn transition model. Our work encompasses both online and offline settings. In the online setting, we introduce an algorithm called corruption-robust optimistic MLE (CR-OMLE), which leverages total-variation (TV)-based information ratios as uncertainty weights for MLE. We prove that CR-OMLE achieves a regret of $\tilde{\mathcal{O}}(\sqrt{T}+C)$, where C denotes the cumulative corruption level after T episodes. We also prove a lower bound to show that the additive dependence on C is optimal. We extend our weighting technique to the offline setting, and propose an algorithm named corruption-robust pessimistic MLE (CR-PMLE). Under a uniform coverage condition, CR-PMLE exhibits suboptimality worsened by $\mathcal{O}(C/n)$, nearly matching the lower bound. To the best of our knowledge, this is the first work on corruption-robust model-based RL algorithms with provable guarantees.

[*]Equal contribution [1]The Hong Kong University of Science and Technology. [2]University of California, Los Angeles. [3]University of Illinois Urbana-Champaign.. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>, Tong Zhang <tongzhang@tongzhang-ml.org>.

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

1. Introduction

Reinforcement learning (RL) seeks to find the optimal policy within an unknown environment associated with rewards and transition dynamics. A representative model for RL is the Markov decision process (MDP) (Sutton & Barto, 2018). While numerous studies assume static rewards and transitions, the environments in real-world scenarios are often non-stationary and vulnerable to adversarial corruption. For instance, autonomous vehicles frequently fall victim to misled navigation caused by hacked maps and adversarially contaminated traffic signs (Eykholt et al., 2018). Similarly, in smart healthcare systems, an adversary with partial knowledge can easily manipulate patient statuses (Nanayakkara et al., 2022). Under this situation, standard RL algorithms often fail to find policies robust to such adversarial corruption. Therefore, how to identify the optimal policies against adversarial corruption has witnessed a flurry of recent investigations.

In this work, we focus on the scenario where the adversary can manipulate the transitions before the agent observes the next state. Achieving a sub-linear regret bound under transition corruption has been shown to be computationally challenging with full information feedback (Abbasi Yadkori et al., 2013), and even information-theoretically challenging with bandit feedback (Tian et al., 2021). Therefore, a series of studies have introduced constraints on the level of corruption, such as limiting the fraction of corrupted samples (Zhang et al., 2022) or the cumulative sum of corruptions over T rounds (Lykouris et al., 2018; Gupta et al., 2019; He et al., 2022; Ye et al., 2023a;b; Yang et al., 2023). While these works exhibit a sub-linear regret bound, to the best of our knowledge, existing works all focus on the setting of model-free RL, where the agent directly learns a policy or a value function from the experiences gained through interactions.

In contrast to model-free RL, in model-based RL, the agent learns an explicit model of the environment and utilizes this model for decision-making. This paradigm not only exempts from Bellman completeness (Jin et al., 2021) (which is a standard assumption for model-free RL) but also has demonstrated impressive sample efficiency in both theories (Kearns & Singh, 2002; Brafman & Tennenholtz, 2002;

Auer et al., 2008; Sun et al., 2019; Agarwal & Zhang, 2022) and applications (Chua et al., 2018; Nagabandi et al., 2020; Schrittwieser et al., 2020), such as robotics (Polydoros & Nalpantidis, 2017) and autonomous driving (Wu et al., 2021). In the online setting, one of the most representative frameworks is Optimistic Maximum Likelihood (OMLE) (Liu et al., 2023a). This framework establishes a confidence set based on log-likelihood and selects the most optimistic model within the confidence set. However, how to make model-based RL provably robust against adversarial corruption remains an open problem.

In this paper, we resolve this open problem in both online and offline settings with a general function approximation. For simplicity, we assume that the reward is known and mainly focuses on learning the transition, which is both unknown and subject to corruption. In particular, we first introduce a cumulative measure for the corruption level of the transition probabilities.

In the online setting, to enhance the resilience of exploitation and exploration to potential corruption, we integrate OMLE with a novel uncertainty weighting technique. Distinct from previous works (He et al., 2022; Ye et al., 2023a;b) in the bandits or model-free RL that characterize uncertainty with rewards or value functions, we characterize the uncertainty with the probability measure of transitions. More specifically, we define the uncertainty as a total-variation (TV)-based information ratio (IR) between the current sample and historical samples. Notably, the introduced IR resembles the eluder coefficient in Zhang (2023). The samples with higher uncertainty are down-weighted since they are more vulnerable to corruption. Additionally, similar to Liu et al. (2023a), we quantify the complexity of the model class $\mathcal{M}$ with TV-based eluder dimension and establish a connection between the TV-based eluder dimension and the cumulative TV-based IR following Ye et al. (2023a) (by considering uncertainty weights).

In addition to the online setting, we further apply the uncertainty weighting technique to the offline learning scenario, and employ pessimistic MLE with uncertainty weights.

Contributions. We summarize our contributions as follows.

- For the online setting, we propose an algorithm CR-OMLE (Corruption-Robust Optimistic MLE) by integrating uncertainty weights with OMLE. This algorithm enjoys a regret bound of $\tilde{\mathcal{O}}(H\log B\sqrt{T\log|\mathcal{M}|\mathrm{ED}}+CH\cdot\mathrm{ED})$, where H is the episode length, B is an upper bound of transition ratios, T is the number of episodes, $|\mathcal{M}|$ is the cardinality of the model class, ED is the eluder dimension, and C is the corruption level over T episodes. Moreover, we construct a novel lower bound of $\Omega(HdC)$ for linear MDPs (Jin et al., 2020) with dimension d. As a result, the corruption-dependent term in the upper bound matches the lower bound, when reduced to the linear setting. These results collectively suggest that our algorithm is not only robust to adversarial corruption but also near-optimal with respect to the corruption level C.
- For the offline setting, we propose an algorithm CR-PMLE (Corruption-Robust Pessimistic MLE) and demonstrate that, given a corrupted dataset with an uniform coverage condition, the sub-optimality of the policy generated by this algorithm against the optimal policy can be upper bounded by $\tilde{\mathcal{O}}(H\log B\sqrt{\log|\mathcal{M}|/(T\mathrm{Cov}(\mathcal{M},\mathcal{D}))}+CH/(T\mathrm{Cov}(\mathcal{M},\mathcal{D})))$, where T is the number of trajectories, $\mathrm{Cov}(\mathcal{M},\mathcal{D})$ is the coverage coefficient with respect to model class $\mathcal{M}$ and dataset $\mathcal{D}$. Furthermore, we establish a lower bound of $\Omega\big(C/(T\cdot\mathrm{Cov}(\mathcal{M},\mathcal{D}))\big)$ for the suboptimality within the linear MDP region. Remarkably, the corruption-dependent term in the upper bound aligns closely with the lower bound up to H. It is also worth noting that this is also the first provable model-based offline RL algorithm even without considering the corruption.

Notation. For two probabilities P_1, P_2, we denote the total variation and the Hellinger distance of P_1 and P_2 by $\mathrm{TV}(P_1\|P_2)=\|P_1-P_2\|_1/2$ and $H(P_1\|P_2)^2=\mathbb{E}_{z\sim P_1}(\sqrt{\mathrm{d}P_2/\mathrm{d}P_1}-1)^2$. We use the short-hand notation $z=(x,a)$ for $(x,a)\in\mathcal{X}\times\mathcal{A}$ when there is no confusion. For two positive sequences $\{f(n)\}_{n=1}^{\infty}$, $\{g(n)\}_{n=1}^{\infty}$, we say $f(n)=\mathcal{O}(g(n))$ if there exists a constant $C>0$ such that $f(n)\le Cg(n)$ for all $n\ge 1$, and $f(n)=\Omega(g(n))$ if there exists a constant $C>0$ such that $f(n)\ge Cg(n)$ for all $n\ge 1$. We use $\tilde{\mathcal{O}}(\cdot)$ to omit polylogarithmic factors.

2. Related Work

Corruption-Robust Bandits and RL. The adversarial corruption is first brought into multi-armed bandit problems by Lykouris et al. (2018), where rewards face attacks by c_t in each round, with the cumulative corruption level over T rounds represented as $C=\sum_{t=1}^{T}|c_t|$. A regret lower bound of $o(T)+O(C)$ was constructed by Gupta et al. (2019), illustrating that an "entangled" relationship between T and C is ideal. Subsequently, the corruption was extended to linear contextual bandits by Bogunovic et al. (2021); Ding et al. (2022); Foster et al. (2020); Lee et al. (2021); Zhao et al. (2021); Kang et al. (2023). However, these approaches either obtain sub-optimal regret bounds or necessitate additional assumptions. He et al. (2022) first achieved a regret bound that matches the lower bound by utilizing an uncertainty weighting technique. Turning to MDPs, the corruption in transitions triggers considerable interest. Wu et al. (2021) first studied corruption on rewards and corruption simultaneously for tabular MPDs.

Later, a unified framework to deal with unknown corruption was established by Wei et al. (2022), involving a weak adversary who decides the corruption amount for each action before observing the agent's action. Subsequently, Ye et al. (2023a;b) extended the uncertainty weighting technique from He et al. (2022) to RL with general function approximation for both online and offline settings respectively, aligning with the lower bound in terms of the corruption-related term. However, previous works all focused on the model-free setting, where models are learned via regression. Notably, corruption in the model-based setting, where the models are learned via maximum likelihood estimation (MLE), remains an unexplored area in the literature.

Model-Based RL. There is an emerging body of literature addressing model-based RL problems, from tabular MDPs (Kearns & Singh, 2002; Brafman & Tennenholtz, 2002; Auer et al., 2008) to rich observation spaces with linear function approximation (Yang & Wang, 2020; Ayoub et al., 2020). Extending beyond linear settings, recent studies (Sun et al., 2019; Agarwal et al., 2020; Uehara et al., 2021; Du et al., 2021; Agarwal & Zhang, 2022; Wang et al., 2023; Liu et al., 2023a; Foster et al., 2021; Zhong et al., 2022b; Chen et al., 2022) have explored general function approximations. Simultaneously, a parallel line of research has focused on the model-free approach with general function approximations (Jin et al., 2021; Jiang et al., 2017; Chen et al., 2022; Zhang, 2023; Agarwal et al., 2023; Zhao et al., 2023; Di et al., 2023; Liu et al., 2023b). Among the works, the most relevant one to ours is Liu et al. (2023a), who makes optimistic estimations via a log-likelihood approach and generalizes an eluder-type condition from Russo & Van Roy (2013). Despite numerous works on online model-based RL problems, to the best of our knowledge, there is still a gap in the literature regarding a theoretical guarantee for offline model-based RL.

Distributional Robust RL. Our work shares the same goal of robustness with another line of work, called distributional robust RL, which formulates the uncertainty of the transitions as an uncertainty set (Roy et al., 2017; Badrinath & Kalathil, 2021; Wang & Zou, 2021) or a set of distributions surrounding the nominal models (Zhou et al., 2021; Shi et al., 2024; Clavier et al., 2023). In this setting, there is no corruption during the training process or exploration phase, and the aim is to find a robust policy that maintains a near-optimal policy when there exists a small distributional shift between the training and real environments. In comparison, our work focuses on cases where the training data is corrupted, and we aim to identify the optimal policy for the hidden environment even with a few corrupted observations. Hence, these two categories of works study different notions of robustness and have different challenges.

3. Preliminaries

Consider an episodic MDP$(\mathcal{X}, \mathcal{A}, H, \mathbb{P}, r)$ specified by a state space $\mathcal{X}$, an action space $\mathcal{A}$, the number of transition steps H, a group of transition models $P = \{P^h : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})\}_{h=1}^H$ and a reward function $r = \{r^h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}\}_{h=1}^H$. Given a policy $\pi = \{\pi^h : \mathcal{X} \to \Delta(\mathcal{A})\}_{h=1}^H$, we define the Q-value and V-value functions as the cumulative sum of rewards from the h-th step for policy π: $Q_\pi^h(x,a) = \sum_{h'=h}^H \mathbb{E}_\pi^*[r^{h'}(x^{h'}, a^{h'}) \,|\, x^h = x, a^h = a]$, $V_\pi^h(x) = \sum_{h'=h}^H \mathbb{E}_\pi^*[r^{h'}(x^{h'}, a^{h'}) \,|\, x^h = x]$. For simplicity, we assume that the reward function r is known and is normalized: $\sum_{h=1}^H r^h(x^h, a^h) \in [0,1]$. Hence, we can delve into learning the transition model by interacting with the system online or using offline data.

In *model-based* RL, we consider an MDP model class $\mathcal{M}$. Without loss of generality, we assume that the class $\mathcal{M}$ has finite elements. Note that the finite class assumption is only to simplify the analysis. We can extend the proof to an infinite model class by taking a finite covering. Each model $M \in \mathcal{M}$ depicts transition probability $P_M^h(x^{h+1}|x,a)$. We use $\mathbb{E}^M[\cdot]$ to represent the expectation over the trajectory under transition probability P_M^π, V_M^π and Q_M^π to denote the V-value and Q-value function of model M and policy π, and $\pi_M = \pi_{Q_M}$ to denote the optimal policy for the model M. Then, we short-notate $V_M^{\pi_M,h}$ and $Q_M^{\pi_M,h}$ as V_M^h and Q_M^h. Additionally, we suppose that there exists an underlying true model $M^* \in \mathcal{M}$ such that the true transition probability is $P_{M_*}^h(x^{h+1}|x^h,a^h)$, and use the short-hand notation $P_*^h = P_{M^*}^h$ and $\mathbb{E}^*[\cdot] = \mathbb{E}^{M_*}[\cdot]$. Given a model M, the model-based Bellman error is defined as

$$\begin{aligned}
&\mathcal{E}^h(M, x^h, a^h) \\
&= Q_M^h(x^h,a^h) - R_*^h(x^h,a^h) - \mathbb{E}^*[V_M^{h+1}(x^{h+1})|x^h,a^h] \\
&= \mathbb{E}^M[V_M^{h+1}(x^{h+1})|x^h,a^h] - \mathbb{E}^*[V_M^{h+1}(x^{h+1})|x^h,a^h].
\end{aligned}$$

For analysis, we assume that for any model $M \in \mathcal{M}$, the ratio between transition dynamic P_M and P_* is upper-bounded.

Assumption 3.1. There exists a constant $B > 0$ such that

$$\sup_{h\in[H], M\in\mathcal{M}} \max\left\{ \left\| \frac{P_*^h}{P_M^h} \right\|_\infty \left\| \frac{P_M^h}{P_*^h} \right\|_\infty \right\} \le B.$$

Online Learning. In online learning, an agent interacts with the environment iteratively for T rounds with the **goal** of learning a sequence of policy $\{\pi_t\}_{t=1}^T$ that minimize the cumulative suboptimality: $\mathrm{Reg}(T) = \sum_{t=1}^T \left[V_*^1(x_t^1) - V_{\pi_t}^1(x_t^1)\right]$.

To depict the structure of the model space $\mathcal{M}$, we introduce the model-based version of the eluder dimension similar to Russo & Van Roy (2013); Liu et al. (2023a).

Definition 3.2 (ϵ-Dependence)**.** For an $\epsilon > 0$, we say that a point z is ϵ-dependent on a set $\mathcal{Z}$ with respect to $\mathcal{M}$ if there exists $M, M' \in \mathcal{M}$ such that $\sum_{z' \in \mathcal{Z}} |\mathrm{tv}(P_M^h(\cdot|z') \| P_{M'}^h(\cdot|z'))|^2 \leq \epsilon^2$ implies $|\mathrm{tv}(P_M^h(\cdot|z) \| P_{M'}^h(\cdot|z))| \leq \epsilon$.

This dependence means that a small in-sample error leads to a small out-of-sample error. Accordingly, we say that z is ϵ-independent of $\mathcal{Z}$ if it is not ϵ-dependent of $\mathcal{Z}$.

Definition 3.3 (TV-Eluder Dimension)**.** For a model class $\mathcal{M}$, an $\epsilon > 0$, the TV-eluder dimension $\mathrm{ED}(\mathcal{M}, \epsilon)$ is the length n of the longest sequence $\{z_1, \ldots, z_n\} \subset \mathcal{X} \times \mathcal{A}$ such that for some $\epsilon' \geq \epsilon$ and all $h \in [H]$, $i \in [n]$, z_i is ϵ' independent of its predecessors.

Particularly, Liu et al. (2023a) also defines the TV-eluder condition. The distinction lies in their consideration of a sequence of policies, while we focus on a sequence of state-action samples. In Theorems D.1 and D.2, we offer examples of tabular and linear MDPs where the TV-eluder dimension can be bounded. To facilitate analysis, we formulate the TV-norm version of the eluder coefficient drawing inspiration from the approach of Zhang (2023). Given a model class $\mathcal{M}'$, sample set $\mathcal{S}_t^h = \{z_s^h\}_{s=1}^t$ and some estimator $\bar{M}_t$ from $\mathcal{S}_t^h$ (will be specified by later algorithms), the information ratio (IR) between the estimation error and the training error within $\mathcal{M}'$ with respect to $\bar{M}_t$ is

$$I^h(\lambda, \mathcal{M}', \mathcal{S}_t^h) = \min\left\{1, \sup_{M \in \mathcal{M}'} \frac{l_t^h(M, \bar{M}_t)}{\sqrt{\lambda + \sum_{s=1}^{t-1} l_s^h(M, \bar{M}_t)^2}}\right\}, \tag{1}$$

where $l_t^h(M, M')$ denotes $\mathrm{TV}(P_M^h(\cdot|z_t^h) \| P_{M'}^h(\cdot|z_t^h))$. We use the linear MDP as an example to illustrate IR. If the transition can be embedded as $P_M^h(x^{h+1}|z^h) = \nu^h(M, x^{h+1})^\top \phi^h(z^h)$, according to Example D.3, the IR can be reduced to

$$\min\left\{1, \|\phi^h(z^h)\|_{(\Sigma_t^h)^{-1}}\right\},$$

where $\Sigma_t^h = \sum_{s=1}^{t-1} \phi^h(z_s^h)\phi^h(z_s^h)^\top$. Intuitively, this quantity represents the uncertainty of vector $\nu^h(M, x^{h+1})$ in the direction of $\phi(z_t^h)$. After observing the state-action pair (x_t^h, a_t^h) in each round, an adversary corrupts the transition dynamics from P_*^h to P_t^h and the agent receives the next state x_t^{h+1} induced by $P_t^h(\cdot|x_t^h, a_t^h)$. To measure the corruption level, we define cumulative corruption.

Definition 3.4 (Corruption Level)**.** We define the corruption level C as the minimum value that satisfies the following property: For the sequence $\{x_t^h, a_t^h\}_{t,h=1}^{T,H}$ chosen by the agent and each stage $h \in [H]$,

$$c_t^h = c_t^h(x_t^h, a_t^h) = \sup_{x^{h+1} \in \Delta_t(\mathcal{X})} \left|\frac{P_t^h(x^{h+1}|x_t^h, a_t^h)}{P_*(x^{h+1}|x_t^h, a_t^h)} - 1\right|,$$

$$\sum_{t=1}^T c_t^h \leq C,$$

where $\Delta_t(\mathcal{X})$ is the support of $P_*^h(\cdot|x_t^h, a_t^h)$.

We use $\mathbb{E}^t$ to denote expectations evaluated in the corrupted transition probability P_t. Note that $P_t^h(\cdot|x_t^h, a_t^h)$ and $P_*^h(\cdot|x_t^h, a_t^h)$ have the same support, i.e., the adversary cannot make the agent transfer to a state that is impossible to be visited under P_*.

Offline Setting. For offline environment, the agent independently collects T trajectories $\mathcal{D} = (x_t^h, a_t^h, r_t^h)_{i,h=1}^{T,H}$ during the data collection process. For each trajectory $t \in [T]$, an adversary corrupts the transition probability to $P_t^h(\cdot|x_t^h, a_t^h)$ after observing (x_t^h, a_t^h). The corruption level C is measured in the same way as in Definition 3.4. The only difference is that corruption occurs during the collection process. The **goal** is to learn a policy $\hat{\pi}$ such that the suboptimality with respect to the uncorrupted transition is sufficiently small:

$$\mathrm{SubOpt}(\hat{\pi}, x^1) = V_*^h(x^1) - V_{\hat{\pi}}^h(x^1).$$

4. Online Model-based RL

In this section, we will first present a novel uncertainty weighting technique tailored for model-based RL, followed by the corruption-robust model-based online RL algorithm and its analysis.

4.1. Uncertainty Weighting for Model-based RL

In this subsection, we will illustrate how our uncertainty-weighting technique differs from the one for model-free RL. The main difficulty arises from the different estimators applied by these two settings. In model-free RL, we estimate the value functions using a value target regression. In contrast, for model-based RL, we directly estimate the hidden transition probability with MLE.

In detail, for model-free RL, let f^* be the uncorrupted true value function and $\hat{f}$ be the least squares estimator with weights σ_s. The corruption term can be directly decomposed as the multiplication between uncertainty $U_s = \sup_{f \in \mathcal{F}} \frac{|\hat{f}(z_s) - f(z_s)|}{\sqrt{\lambda + \sum_{j=1}^t (\hat{f}(z_j) - f(z_j))^2/\sigma_j}}$ and corruption (Ye et al., 2023a;b):

$$\sum_{s=1}^t \frac{(\hat{f}(z_s) - f^*(z_s))^2}{\sigma_s} = \sum_{s=1}^t \frac{(\hat{f}(z_s) - y_s)^2 - (f^*(z_s) - y_s)^2}{\sigma_s} + 2\sum_{s=1}^t \frac{(\hat{f}(z_s) - f^*(z_s))\epsilon_s}{\sigma_s} + \underbrace{2\sum_{s=1}^t \frac{U_s c_s}{\sigma_s} \cdot \beta}_{\text{Corruption term}}.$$

By choosing $\sigma_s \geq U_s/\alpha$, we can convert the corruption term to $\alpha C\beta$. To conclude, only if we transform the corruption term into the multiplication between uncertainty U_s

and corruption c_s, can the uncertainty-related weights cancel U_s and bring in a small hyper-parameter α.

While for model-based RL, it is difficult to decompose the corruption term as the multiplication between uncertainty and corruption terms from the log-likelihood, especially when uncertainty is based on the total variation (TV) norm. Specifically, let P_M be the estimated probability, P_* be the true probability, and P_t be the corrupted probability at round t. Neglecting the upscript h for convenience, we write:

$$\mathbb{E}\frac{1}{\sigma_s}\log\sqrt{\frac{\mathrm{d}P_{\widehat{M}}(x|z_s)}{\mathrm{d}P_*(x|z_s)}} = \underbrace{\frac{1}{\sigma_s}\int \mathrm{d}P_*(x|z_s)\log\sqrt{\frac{\mathrm{d}P_{\widehat{M}}(x|z_s)}{\mathrm{d}P_*(x|z_s)}}}_{\text{Uncorrupted term}}$$
$$+\underbrace{\frac{1}{\sigma_s}\int(\mathrm{d}P_s(x|z_s)-\mathrm{d}P_*(x|z_s))\log\sqrt{\frac{\mathrm{d}P_{\widehat{M}}(x|z_s)}{\mathrm{d}P_*(x|z_s)}}}_{\text{Corrupted term}}.$$

To make the corrupted term a multiplication between uncertainty and corruption, we use $\log x \le x-1$ and decompose the integration into two regions according to whether $\mathrm{d}P_s - \mathrm{d}P_*$ is positive or not. Then, we deal with the variance similarly and use Assumption 3.1 and the Freedman inequality to obtain

$$\mathbb{E}\sum_{s=1}^{t-1}\frac{1}{\sigma_s^h}\log\sqrt{\frac{\mathrm{d}P^h_{\bar{M}_t}(x^{h+1}|z_s^h)}{\mathrm{d}P^h_*(x^{h+1}|z_s^h)}}$$
$$\lesssim -\sum_{s=1}^{t-1}\frac{\mathrm{TV}\big(P^h_*(\cdot|z_s^h)\|P^h_{\bar{M}_t}(\cdot|z_s^h)\big)^2}{\sigma_s^h}+\underbrace{\sum_{s=1}^{t-1}\frac{c_s^h U_s(z_s^h)\beta}{\sigma_s^h}}_{\text{Corruption Term}},$$

where we use $\lesssim$ to omit constants for conciseness, and U_s is the uncertainty defined in (4). Therefore, our new technique makes it possible for an TV-based uncertainty weight to control the corruption term.

4.2. Algorithm: CR-OMLE

We propose Algorithm 1 for the online episodic learning setting. In each round t, after observing an initial state x_t^1, the agent follows the principle of optimism and selects the model M_t from the confidence set $\mathcal{M}_t$ to maximize the value function $V_M^1(x_t^1)$. Subsequently, the agent follows the greedy policy π_t to collect a trajectory $\{(x_t^h, a_t^h, r_t^h)\}_{h=1}^H$. In constructing the confidence set, the agent initially learns $\bar{M}_{t+1}$ by maximizing the weighted log-likelihood, where the weight σ_t^h is a truncated variant of the Information Ratio (IR) (1), referred to as uncertainty

$$\sigma_t^h = \max\Big\{1, \frac{1}{\alpha}U_t(x_t^h, a_t^h)\Big\}, \tag{3}$$

where the hyper-parameter $\alpha > 0$ is inversely proportional to the corruption level, and we define uncertainty as the

Algorithm 1 Corruption-Robust Optimistic MLE (CR-OMLE)

1: **Input:** $\mathcal{M}_1 = \mathcal{M}$ and $\mathcal{D} = \{\}$.
2: **for** t=1,…,T **do**
3: Observe x_t^1;
4: Construct $M_t = \mathrm{argmax}_{M\in\mathcal{M}_t} V_M^1(x_t^1)$;
5: Let π_t be the greedy policy of $V_{M_t}^1$;
6: Collect new trajectory $\{x_t^1, a_t^1, r_t^1, \ldots, x_t^H, a_t^H, r_t^H\}$ and update it into $\mathcal{D}$;
7: Set σ_t^h as (3), and calculate

$$\bar{M}_{t+1} = \operatorname*{argmax}_{M'\in\mathcal{M}_t}\sum_{s=1}^{t}\sum_{h=1}^{H}\frac{\log P^h_{M'}(x_s^{h+1}|x_s^h, a_s^h)}{\sigma_s^h};$$

8: Find β and construct the confidence set $\mathcal{M}_{t+1}$ as

$$\Big\{M\in\mathcal{M}_t : \forall\, h\in[H],\ \sum_{s=1}^{t}\frac{\log P^h_M(x_s^{h+1}|x_s^h,a_s^h)}{\sigma_s^h} \ge \sum_{s=1}^{t}\frac{\log P^h_{\bar{M}_{t+1}}(x_s^{h+1}|x_s^h,a_s^h)}{\sigma_s^h} - \beta^2\Big\}. \tag{2}$$

9: **end for**

Information Ratio (IR) with weights:

$$U_t(x_t^h, a_t^h) = \sup_{M\in\mathcal{M}_t}\frac{\mathrm{TV}\big(P^h_M(\cdot|x_t^h,a_t^h)\|P^h_{\bar{M}_t}(\cdot|x_t^h,a_t^h)\big)}{\sqrt{\lambda+\sum_{s=1}^{t-1}\mathrm{TV}\big(P^h_M(\cdot|x_s^h,a_s^h)\|P^h_{\bar{M}_t}(\cdot|x_s^h,a_s^h)\big)^2/\sigma_s^h}}. \tag{4}$$

Subsequently, the confidence set is a subset of $\mathcal{M}_t$ that introduces a β^2-relaxation to the maximum of the log-likelihood as shown in (2).

Computation Efficiency. If the model class is finite or has a finite cover with cardinality M, the computational complexity is MTH since in Line 4 and 7 of Algorithm 1, we have to search over also the components in the model class. If the model class is complicated, we have to acknowledge that methods like version space methods (Liu et al., 2023a; Jiang et al., 2017; Wang et al., 2023; Jin et al., 2021), which construct a confidence set, face computational drawbacks as they are computationally inefficient. However, practical algorithms may only need to leverage the insight that optimism is helpful and may not require such thorough exploration. Instead, we could construct a bonus and add it to the value function, and then choose the greedy policy for the optimistic value function (Curi et al., 2020).

When computing the weights, calculating the uncertainty quantity in practical scenarios is the main difficulty. Inspired by Ye et al. (2023b), we can approximate the uncertainty by "bootstrapped uncertainty". The intuition is that if the transition probability can be embedded into a

vector space $P_M^h(x^{h+1}|z^h) = \nu^h(M, x^{h+1})^\top \phi^h(z^h)$. According to Example D.3, the uncertainty can be reduced to $\|\phi^h(z^h)\|_{(\Sigma_t^h)^{-1}}$, where the covariance matrix $\Sigma_t^h = \sum_{s=1}^{t-1} \phi^h(z_s^h)\phi^h(z_s^h)^\top$. Hence, from Ye et al. (2023b), the bootstrapped variance (called bootstrapped uncertainty) is an estimation of the uncertainty. To compute the bootstrapped uncertainty, we first learn K transition probabilities independently with different seeds. Then, we take the uncertainty estimation as $\sqrt{\mathrm{Var}_{i=1,\dots,K}[P_{M_k}(z)]}$. We leave the development of practical algorithms as future work.

4.3. Regret Bounds

The following two theorems offer theoretical guarantees for the upper and lower bounds of regret under the CR-OMLE algorithm.

Theorem 4.1 (Upper Bound)**.** *Under Assumption 3.1, given a finite eluder dimension* $\mathrm{ED}(\mathcal{M}, \epsilon)$*, if we choose* $\beta = 5\sqrt{\log(|\mathcal{M}|/\delta)\log^2 B} + 7\alpha C$ *and* $\lambda = \log|\mathcal{M}|$, $\alpha = \sqrt{\log|\mathcal{M}|\log^2 B}/C$*, with probability at least* $1-\delta$*, the regret of Algorithm 1 is upper bounded by*

$$\mathrm{Reg}(T) = \tilde{\mathcal{O}}\Big(H\log B\sqrt{T\log|\mathcal{M}|\mathrm{ED}(\mathcal{M}, \sqrt{\lambda/T})} + CH\cdot\mathrm{ED}(\mathcal{M}, \sqrt{\lambda/T})\Big).$$

This regret bound consists of two parts. The main part $H(T\log B\log|\mathcal{M}|\mathrm{ED}(\mathcal{M}, \sqrt{\lambda/T}))^{1/2}$ is unrelated to corruption and matches the bound for OMLE (Liu et al., 2023a). According to Theorem D.2, in the linear MDP setting with d dimensions, $\mathrm{ED}(\mathcal{M}, \sqrt{\lambda/T}) = O(d)$. Thus, the regret reduces to $\tilde{\mathcal{O}}(H\log B\sqrt{Td\log|\mathcal{M}|} + CHd)$, where the corruption part CHd also matches the lower bound from Theorem 4.2. We can achieve sub-linear regret when the corruption level is sub-linear. We highlight the proof sketch in the sequel and delay the detailed proof to Appendix A.1.

Theorem 4.2 (Lower Bound)**.** *For any dimension* $d \geq 2$*, stage* $H \geq 3$ *and a known corruption level* $C > 0$*, if the number of episode* T *satisfied* $T \geq \Omega(dCH + H^2)$*, there exists an instance such that any algorithm must incur* $(H-2)(d-1)C/64$ *expected regret.*

Remark 4.3. It is noteworthy that the lower bound in Theorem 4.2 matches the corruption term dCH in our upper bound, up to logarithmic factors. This result suggests that our algorithm is optimal for defending against potential adversarial attacks. Furthermore, for any algorithm, the regret across the first T episodes is limited to no more than $\Omega(T)$. In this situation, the requirement that the number of episodes $T > \Omega(dCH)$ is necessary to achieve a lower bound of $\tilde{\Omega}(dCH)$. The proof is available in Appendix A.2.

Unknown Corruption Level Since it is hard to know the corruption level ahead of time, we discuss the solution for the unknown corruption level. Note that only the choice of parameter $\alpha = \sqrt{\log|\mathcal{M}|\log^2 B}/C$ requires the knowledge of C, so we following He et al. (2022) to replace C in α by a predefined corruption tolerance threshold $\bar{C}$. As long as the actually corruption level C is no more then than the tolerance threshold $\bar{C}$, we can still obtain a non-trivial regret bound.

Theorem 4.4 (Unknown Corruption Level)**.** *Under the same conditions as Theorem 4.1, let* $\beta_t^h = \Theta(\sqrt{\log(|\mathcal{M}|/\delta)\log^2 B})$ *and* $\alpha = \sqrt{\log|\mathcal{M}|}/\bar{C}$*. If* $C \leq \bar{C}$*, we have with probability at least* $1-\delta$*,*

$$\mathrm{Reg}(T) = \tilde{\mathcal{O}}\Big(H\log B\sqrt{T\log|\mathcal{M}|\mathrm{ED}(\mathcal{M}, \sqrt{\lambda/T})} + \bar{C}H\cdot\mathrm{ED}(\mathcal{M}, \sqrt{\lambda/T})\Big).$$

On the other hand, if $C > \bar{C}$*, we have* $\mathrm{Reg}(T) = \tilde{\mathcal{O}}(T)$.

Remark 4.5. Theorem 4.4 establishes a trade-off between adversarial defense and algorithmic performance. Specifically, a higher predefined corruption tolerance threshold $\bar{C}$ can effectively fortify against a broader spectrum of attacks. However, this advantage comes at the cost of a more substantial regret guarantee. For a notable case, when we set $\bar{C} = \sqrt{T\log|\mathcal{M}|/\mathrm{ED}(\mathcal{M}, \sqrt{\lambda/T})}$, our algorithm demonstrates a similar performance as the no corruption case, even in the absence of prior information, within the specified region of $C \leq \bar{C}$. We defer the proof to Appendix A.3.

4.4. Proof Sketch.

We provide the proof sketch for Theorem 4.1, which consists of three key steps.

Step I. Regret Decomposition. Following the optimism principle, we can decompose regret as the sum of Bellman errors and corruption:

$$\mathrm{Reg}(T) \leq \sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\sum_{h=1}^{H}\mathcal{E}^h(M_t, z_t^h) + \sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\sum_{h=1}^{H}c_t^h(x^h, a^h).$$

The summation of corruption occurs because, when transforming the regret into Bellman error, we must consider the Bellman error under the distribution of real data, i.e., the corrupted one.

Step II. Confidence Set for Optimism. To ensure that the true, uncorrupted model M^* belongs to the confidence

set $\mathcal{M}_t$, we demonstrate in the following lemma that M^* satisfies (2). Moreover, for any $M \in \mathcal{M}_t$, the in-sample error is bounded.

Lemma 4.6 (Confidence Set). *Under Assumption 3.1, if we choose* $\beta = 5\sqrt{\log(|\mathcal{M}|/\delta)\log^2 B} + 7\alpha C$ *in Algorithm 1, then with probability at least* $1-\delta$, *for all* $h \in [H]$ *and all* $t \in [T]$, $M_* \in \mathcal{M}_t$, *we have*

$$\sum_{s=1}^{t-1} \mathrm{TV}(P_*^h(\cdot|x_s^h, a_s^h) \| P_{\bar{M}_t}^h(\cdot|x_s^h, a_s^h))^2/\sigma_s^h \le 2\beta^2.$$

Moreover, for any $M \in \mathcal{M}_t$, *with probability at least* $1-\delta$, *we have*

$$\sum_{s=1}^{t-1} \mathrm{TV}(P_M^h(\cdot|x_s^h, a_s^h) \| P_{\bar{M}_t}^h(\cdot|x_s^h, a_s^h))^2/\sigma_s^h \le 4\beta^2.$$

The key ideas of the analysis are presented in Subsection 4.1, and we postpone the detailed explanation to Appendix C.1.

Step III. Bounding the sum of Bellman Errors. Finally, by combining the results derived from the first two steps and categorizing the samples into two classes based on whether $\sigma_t^h > 1$, we obtain an upper bound on the regret:

$$\tilde{\mathcal{O}}\Bigg(\sqrt{TH\log|\mathcal{M}|\sum_{h=1}^{H}\sup_{\mathcal{S}_T}\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\big(I_\sigma^h(\lambda,\mathcal{M}_t,\mathcal{S}_t)\big)^2} + C\sum_{h=1}^{H}\sup_{\mathcal{S}_T}\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\big(I_\sigma^h(\lambda,\mathcal{M}_t,\mathcal{S}_t)\big)^2\Bigg),$$

where we define the weighted form of IR:

$$I_\sigma^h(\lambda,\mathcal{M},\mathcal{S}_t^h) = \min\Bigg\{1, \sup_{M\in\mathcal{M}_t} \frac{\mathrm{TV}(P_M^h(\cdot|z_t^h)\|P_{\bar{M}_t}^h(\cdot|z_t^h))/(\sigma_t^h)^{1/2}}{\sqrt{\lambda+\sum_{s=1}^{t-1}\mathrm{TV}(P_M^h(\cdot|z_s^h)\|P_{\bar{M}_t}^h(\cdot|z_s^h))^2/\sigma_s^h}}\Bigg\}. \quad (5)$$

Moreover, to establish an instance-independent bound, we follow Ye et al. (2023a) to demonstrate the relationship between the sum of weighted Information Ratio (IR) concerning dataset $\mathcal{S}_t = {z_s^h}_{s=1}^t$ and the eluder dimension:

$$\sup_{\mathcal{S}_T^h}\sum_{t=1}^{T} I_\sigma^h(\lambda,\mathcal{M},\mathcal{S}_t)^2 = \tilde{O}(\mathrm{ED}(\mathcal{M},\sqrt{\lambda/T})),$$

which leads to the final bound.

5. Offline Model-based RL

In this section, we will extend the uncertainty weighting technique proposed in 4 to the setting of offline RL, and propose a corruption-robust model-based offline RL algorithm and its analysis.

5.1. Algorithm: CR-PMLE

Algorithm 2 Corruption-Robust Pessimistic MLE (CR-PMLE)

1: **Input:** $\mathcal{D} = \{(x_t^h, a_t^h, r_t^h)\}_{t,h=1}^{T,H}, \mathcal{M}$.
2: For $h = 1,\dots,H$, choose weights $\{\sigma_t^h\}_{t=1}^T$ by proceeding Algorithm 3 with inputs $\{(x_t^h, a_t^h)\}_{t=1}^T, \mathcal{M}, \alpha$;
3: Let

$$\bar{M} = \operatorname*{argmax}_{M\in\mathcal{M}} \sum_{t=1}^{T}\sum_{h=1}^{H}(\sigma_t^h)^{-1}\log P_M^h(x_t^{h+1}|x_t^h, a_t^h);$$

4: Find β and construct confidence set $\widehat{\mathcal{M}}$

$$\Big\{M\in\mathcal{M}: \forall\, h\in[H],\ \sum_{t=1}^{T}\frac{\log P_M^h(x_t^{h+1}|x_t^h,a_t^h)}{\sigma_t^h} \ge \sum_{t=1}^{T}\frac{\log P_{\bar{M}}^h(x_t^{h+1}|x_t^h,a_t^h)}{\sigma_t^h} - \beta^2\Big\};$$

5: Set $(\hat{\pi}, \widehat{M}) = \operatorname{argmax}_{\pi\in\Pi}\min_{M\in\widehat{\mathcal{M}}} V_{M,\pi}^1$;
6: **Output:** $\{\hat{\pi}^h\}_{h=1}^H$.

In the offline learning setting, we introduce our algorithm, Corruption-Robust Pessimistic MLE, shown in Algorithm 2. In this algorithm, we also estimate the model by maximizing the log-likelihood with uncertainty weights and constructing the confidence set $\widehat{\mathcal{M}}$. To address the challenge that weights cannot be computed iteratively in rounds, as in the online setting, we adopt the weight iteration algorithm (Algorithm 3) from Ye et al. (2023b) to obtain an approximated uncertainty. They prove that this iteration will converge since the weights are monotonically increasing and upper-bounded. Moreover, as long as the weights σ_t^h are in the same order as the truncated uncertainty, the learning remains robust to corruption. For completeness, we present the convergence result in Lemma B.1.

Algorithm 3 Uncertainty Weight Iteration

1: **Input:** $\{(x_t^h, a_t^h)\}_{t=1}^T$,$\mathcal{M}$,$\alpha > 0$.
2: **Initialization:** $k = 0$, $\sigma_t^0 = 1, t = 1,\dots,T$.
3: **repeat**
4: $k \leftarrow k+1$;
5: For $t = 1,\dots,T$, let

$$\sigma_t^k \leftarrow \max\Big(1, \sup_{M,M'\in\mathcal{M}} \frac{l_t^h(M,M')/\alpha}{\sqrt{\lambda+\sum_{s=1}^{T} l_s^h(M,M')^2/\sigma_s^{k-1}}}\Big);$$

6: **until** $\max_{t\in[T]}\sigma_t^k/\sigma_t^{k-1} \le 2$;
7: **Output:** $\{\sigma_t^k\}_{t=1}^T$.

To consider the case where the offline data lacks full cov-

erage, we proceed with pessimism and tackle a minimax optimization

$$(\hat{\pi}, \widehat{M}) = \operatorname*{argmax}_{\pi\in\Pi} \min_{M\in\widehat{\mathcal{M}}} V^{\pi}_{M,\pi},$$

which shares a similar spirit with the model-free literature (Xie et al., 2021).

5.2. Analysis

In this subsection, we first provide an instance-dependence suboptimality upper bound. Then, we can obtain an instance-independent bound under the uniform coverage condition. Finally, we present a lower bound under the uniform coverage condition, which aligns with the corruption term in the upper bound.

Instance-Dependent Bound. To facilitate analysis, we define the offline variant of IR (1) as follows.

Definition 5.1. Given an offline dataset $\mathcal{D}$, for any initial state $x^1 = x \in \mathcal{X}$, the information coefficient with respect to space $\widehat{\mathcal{M}}$ is:

$$\begin{aligned}&\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D})\\ &= \max_{h\in[H]} \mathbb{E}_{\pi_*}\Big[\sup_{M,M'\in\widehat{\mathcal{M}}} \frac{T\cdot l^h(M,M',z^h)^2/\sigma^h(z^h)}{\lambda+\sum_{t=1}^{T} l^h_t(M,M')^2/\sigma^h_t}\Big],\end{aligned}$$

where we define $l^h(M, M', z^h) = \mathrm{TV}(P^h_M(\cdot|z^h)\|P^h_{M'}(\cdot|z^h))$, $\mathbb{E}_{\pi_*}$ is taken with respect to (x^h, a^h) induced by policy π_* for the uncorrupted transition, and we define

$$\sigma^h(z^h) = \max\Big\{1, \sup_{M,M'\in\widehat{\mathcal{M}}} \frac{l^h(M,M',z^h)/\alpha}{\sqrt{\lambda+\sum_{t=1}^{T} l^h_t(M,M')^2/\sigma^h_t}}\Big\}.$$

This coefficient depicts the information ratio of the trajectory for the optimal policy π_* and the dataset. Now, we can demonstrate an instance-dependent bound, where uniform data coverage is not required.

Theorem 5.2 (Instance-Dependent Bound)**.** *Suppose that Assumption 3.1 holds. Under Algorithm 2, if we choose* $\lambda = \log|\mathcal{M}|$, $\alpha = \sqrt{\log|\mathcal{M}|\log^2 B}/C$ *and*

$$\beta = 5\sqrt{\log(|\mathcal{M}|/\delta)\log^2 B} + 7\alpha C.$$

Then, with probability at least $1-2\delta$*, the sub-optimality* $\mathrm{SubOpt}(\hat{\pi}, x)$ *is bounded by*

$$\tilde{\mathcal{O}}\Big(H\log B\sqrt{\frac{\mathrm{IC}^{\sigma}(\lambda,\widehat{\mathcal{M}},\mathcal{D})\log|\mathcal{M}|}{T}} + \frac{\mathrm{IC}^{\sigma}(\lambda,\widehat{\mathcal{M}},\mathcal{D})CH}{T}\Big),$$

where the weighted information coefficient $\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D})$ *is defined in Definition 5.1.*

This bound depends on the information coefficient with uncertainty weights along the trajectory induced by π_*. To remove the weights σ^h_t from the suboptimality bound, we need a stronger coverage condition.

We present the proof in Appendix B.1, and highlight key points here. The proof also contains three steps. First of all, with pessimism, the suboptimality can be decomposed as

$$\mathbb{E}_{\pi_*}\sum_{h=1}^{H}\mathcal{E}^h(\widehat{M}, z^h) \le \mathbb{E}_{\pi_*}\sum_{h=1}^{H}\mathrm{TV}(P^h_{\widehat{M}}(\cdot|z^h)\|P^h_{M_*}(\cdot|z^h)).$$

Given the proximity of the weights σ^h_t to uncertainty measures, we can modify Lemma 4.6 to achieve $\beta = \Theta(\sqrt{\log(|\mathcal{M}|/\delta)\log^2 B} + \alpha C)$ for $\delta > 0$. This modification allows us to establish the high-probability inclusion of M_* in $\widehat{\mathcal{M}}$. Ultimately, by integrating the initial two steps and leveraging the relationship between uncertainty weights and the information coefficient, we derive the bound.

Instance-Independent Bound. To obtain an instance-independent upper bound, we follow Ye et al. (2023b) to introduce a TV-norm-based version of the coverage condition for the dataset.

Assumption 5.3 (Uniform Data Coverage)**.** Define the measure for any $M, M' \in \mathcal{M}$, $\rho^h(M, M') = \sup_z \mathrm{TV}(P^h_M(\cdot|z)\|P^h_{M'}(\cdot|z))$. There exists a constant $\mathrm{Cov}(\mathcal{M}, \mathcal{D})$ such that for any $h \in [H]$, and two distinct $M, M' \in \mathcal{M}$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{TV}(P^h_M(\cdot|z^h_t)\|P^h_{M'}(\cdot|z^h_t))^2 \ge \mathrm{Cov}(\mathcal{M},\mathcal{D})\big(\rho^h(M,M')\big)^2.$$

This is a generally adopted condition in offline literature (Duan et al., 2020; Wang et al., 2020; Xiong et al., 2022; Di et al., 2023). Intuitively, this condition depicts that the dataset explores each direction of the space. In linear case where $\mathcal{M}^h = \{\langle \nu^h(M, x^{h+1}), \phi^h(\cdot)\rangle : \mathcal{X} \to \mathbb{R}\}$, according to Lemma D.4, this assumption is implied by the condition that the covariance matrix $T^{-1}\sum_{t=1}^{T}\phi(z^h_t)\phi(z^h_t)^\top$ is strictly positive definite, and $\mathrm{Cov}(\mathcal{M}, \mathcal{D})$ has a $\Theta(d^{-1})$ dependence on d. Under this condition, the suboptimality upper bound yielded by Algorithm 2 is guaranteed.

Theorem 5.4 (Instance-Independent Bound)**.** *Supposing that Assumption 3.1 and 5.3 hold, if we choose* $\lambda = \log|\mathcal{M}|$, $\alpha = \sqrt{\log|\mathcal{M}|\log^2 B}/C$ *and*

$$\beta = 5\sqrt{\log(|\mathcal{M}|/\delta)\log^2 B} + 7\alpha C.$$

Then, with probability at least $1-2\delta$*, the sub-optimality* $\mathrm{SubOpt}(\hat{\pi}, x)$ *of Algorithm 2 is bounded by*

$$\tilde{\mathcal{O}}\Big(H\log B\sqrt{\frac{\log|\mathcal{M}|}{T\mathrm{Cov}(\mathcal{M},\mathcal{D})}} + \frac{CH}{T\mathrm{Cov}(\mathcal{M},\mathcal{D})}\Big).$$

We present the proof in Appendix B.2. Additionally, when the corruption level is unknown, we can take the hyperparameter α as a tuning parameter and attain a similar result as Theorem 4.4.

Remark 5.5 (Comparison between Online and Offline Results). Both of the online and offline bounds are composed of the main term (uncorrupted error) and the corruption error, and the corruption error is sublinear as long as the corruption level C is sublinear w.r.t sample size T. The main difference is that the online bound is affected by the eluder dimension and the offline bound is affected by the covering coefficient. This difference comes from the ability to continue interacting with the environment. In the online setting, the agent can explore the model space extensively, thus, the final regret bound is controlled by the complexity of the model space. In contrast, exploration in the offline setting is constrained by a given dataset, thus the final bound is determined by the coverage quality of the dataset.

Regardless of the different conditions between online and offline cases, this bound roughly aligns with the online upper bound. Specifically, achieving an ϵ-suboptimality under both online and offline algorithms requires a $\tilde{\mathcal{O}}(\epsilon^{-2}H\log B\sqrt{\log|\mathcal{M}|\dim}+\epsilon^{-1}CH\dim)$ sample complexity. The dim denotes the eluder dimension for the online setting and inverse of coverage $(\mathrm{Cov}(\mathcal{M},\mathcal{D}))^{-1}$ for the offline setting. Particularly, for the linear case, the $\dim = d$ for both settings. Additionally, the corruption component of the bound closely corresponds to the demonstrated lower bound below.

Theorem 5.6 (Lower Bound)**.** *For a given corruption level C, data coverage coefficient $\mathrm{Cov}(\mathcal{M},\mathcal{D})$, dimension $d>3$ and episode length $H>2$, it the data size T satisfied $T>\Omega\big(\mathrm{Cov}(\mathcal{M},\mathcal{D})/C\big)$, there exist a hard to learn instance such that the suboptimality for any algorithm is lower bounded by*

$$\mathbb{E}[\mathrm{SubOpt}(\hat{\pi},x)]\geq\Omega\Big(\frac{C}{\mathrm{Cov}(\mathcal{M},\mathcal{D})T}\Big).$$

Remark 5.7. The offline lower bound in Theorem 5.6 matches the corruption term $CH/\big(T\mathrm{Cov}(\mathcal{M},\mathcal{D})\big)$ in our upper bound, up to a factor of H. This result demonstrates that our algorithm achieves near-optimal suboptimality guarantee for defending against adversarial attacks.

In comparison to the online lower bound presented in Theorem 4.1, a discrepancy of H emerges, and we believe this variance comes from the data coverage coefficient $\mathrm{Cov}(\mathcal{M},\mathcal{D})$. In general, meeting the uniform data coverage assumption becomes more challenging with an increased episode length H, resulting in a smaller data coverage coefficient $\mathrm{Cov}(\mathcal{M},\mathcal{D})$. Furthermore, for the hard-to-learn instances that constructed in the lower bound, the coefficient exhibits an inverse dependence on the episode length H, specifically, $\mathrm{Cov}(\mathcal{M},\mathcal{D})=O(1/H)$. Under this situation, our lower bound does not incur additional dependencies on H and we leave it as a potential future work. More details can be found in Appendix B.3.

6. Conclusion and Future Work

We delve into adversarial corruption for model-based RL, encompassing both online and offline settings. Our approach involves quantifying corruption by summing the differences between true and corrupted transitions. In the online setting, our algorithm, CR-OMLE, combines optimism and weighted log-likelihood principles, with weights employing a TV-norm-based uncertainty. Our analysis yields a $O(\sqrt{T}+C)$ regret upper bound for CR-OMLE and establishes a new lower bound for MDPS with transition corruption, aligning in terms of corruption-dependent terms. In the offline realm, we present CR-PMLE, a fusion of pessimism and MLE with uncertainty weights, offering theoretical assurances through instance-dependent and instance-independent upper bounds. Notably, the instance-independent bound necessitates a uniform coverage condition, where the corruption-dependent term nearly matches the lower bound. We anticipate that our findings will contribute theoretical insights to address practical challenges in model-based RL, particularly for adversarial corruption.

Several future works are worth exploring: (1) Extending the uncertainty weighting technique to representation learning problems in RL (Jiang et al., 2017; Liu et al., 2022) would be an intriguing direction; and (2) In situations where the corruption level C is unknown, our method relies on an optimistic estimation of corruption denoted by $\bar{C}$. There is a need for further investigation into the applicability of the model selection technique introduced by Wei et al. (2022) to the realm of model-based RL.

Acknowledgements

The authors would like to thank the anonymous reviewers for many insightful comments and suggestions. JH and QG are supported by the National Science Foundation CAREER Award 1906169 and research fund from UCLA-Amazon Science Hub. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

Impact Statement

This paper introduces research aimed at enhancing the resilience of model-based reinforcement learning (RL) when confronted with adversarial corruption and malicious attacks. The enhanced robustness achieved by this work not only enhances the reliability of RL systems but also con-

tributes to the development of more secure and trustworthy AI technologies. It will also promote greater public trust in AI applications, thereby promoting their responsible and ethical deployment across various domains.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Abbasi Yadkori, Y., Bartlett, P. L., Kanade, V., Seldin, Y., and Szepesvári, C. Online learning in markov decision processes with adversarially chosen transition probability distributions. *Advances in neural information processing systems*, 26, 2013.

Agarwal, A. and Zhang, T. Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity. *Advances in Neural Information Processing Systems*, 35:35284–35297, 2022.

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

Agarwal, A., Jin, Y., and Zhang, T. Vo q l: Towards optimal regret in model-free rl with nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 987–1063. PMLR, 2023.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

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

Badrinath, K. P. and Kalathil, D. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pp. 511–520. PMLR, 2021.

Bogunovic, I., Losalka, A., Krause, A., and Scarlett, J. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 991–999. PMLR, 2021.

Brafman, R. I. and Tennenholtz, M. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct): 213–231, 2002.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

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

Chua, K., Calandra, R., McAllister, R., and Levine, S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

Clavier, P., Pennec, E. L., and Geist, M. Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*, 2023.

Curi, S., Berkenkamp, F., and Krause, A. Efficient model-based reinforcement learning through optimistic policy search and planning. *Advances in Neural Information Processing Systems*, 33:14156–14170, 2020.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. 2008.

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.

Ding, Q., Hsieh, C.-J., and Sharpnack, J. Robust stochastic linear contextual bandits under adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 7111–7123. PMLR, 2022.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pp. 2826–2836. PMLR, 2021.

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.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018.

Foster, D. J., Gentile, C., Mohri, M., and Zimmert, J. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33: 11478–11489, 2020.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Gupta, A., Koren, T., and Talwar, K. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pp. 1562–1578. PMLR, 2019.

He, J., Zhou, D., Zhang, T., and Gu, Q. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in Neural Information Processing Systems*, 35:34614–34625, 2022.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

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

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Kang, Y., Hsieh, C.-J., and Lee, T. Robust lipschitz bandits to adversarial corruptions. *arXiv preprint arXiv:2305.18543*, 2023.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

Lee, C.-W., Luo, H., Wei, C.-Y., Zhang, M., and Zhang, X. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, pp. 6142–6151. PMLR, 2021.

Liu, Q., Chung, A., Szepesvári, C., and Jin, C. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pp. 5175–5220. PMLR, 2022.

Liu, Q., Netrapalli, P., Szepesvari, C., and Jin, C. Optimistic mle: A generic model-based algorithm for partially observable sequential decision making. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 363–376, 2023a.

Liu, Z., Lu, M., Xiong, W., Zhong, H., Hu, H., Zhang, S., Zheng, S., Yang, Z., and Wang, Z. One objective to rule them all: A maximization objective fusing estimation and planning for exploration. *arXiv preprint arXiv:2305.18258*, 2023b.

Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.

Nagabandi, A., Konolige, K., Levine, S., and Kumar, V. Deep dynamics models for learning dexterous manipulation. In *Conference on Robot Learning*, pp. 1101–1112. PMLR, 2020.

Nanayakkara, T., Clermont, G., Langmead, C. J., and Swigon, D. Unifying cardiovascular modelling with deep reinforcement learning for uncertainty aware control of sepsis treatment. *PLOS Digital Health*, 1(2): e0000012, 2022.

Polydoros, A. S. and Nalpantidis, L. Survey of model-based reinforcement learning: Applications on robotics. *Journal of Intelligent & Robotic Systems*, 86(2):153–173, 2017.

Roy, A., Xu, H., and Pokutta, S. Reinforcement learning under model mismatch. *Advances in neural information processing systems*, 30, 2017.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

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

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588 (7839):604–609, 2020.

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 36, 2024.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

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

Tian, Y., Wang, Y., Yu, T., and Sra, S. Online learning in unknown markov games. In *International conference on machine learning*, pp. 10279–10288. PMLR, 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.

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

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*, 2020.

Wang, Y. and Zou, S. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

Wei, C.-Y., Dann, C., and Zimmert, J. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pp. 1043–1096. PMLR, 2022.

Wu, T., Yang, Y., Du, S., and Wang, L. On reinforcement learning with adversarial corruption and its application to block mdp. In *International Conference on Machine Learning*, pp. 11296–11306. PMLR, 2021.

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

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. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pp. 10746–10756. PMLR, 2020.

Yang, R., Zhong, H., Xu, J., Zhang, A., Zhang, C., Han, L., and Zhang, T. Towards robust offline reinforcement learning under diverse data corruption. *arXiv preprint arXiv:2310.12955*, 2023.

Ye, C., Xiong, W., Gu, Q., and Zhang, T. Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes. In *International Conference on Machine Learning*, pp. 39834–39863. PMLR, 2023a.

Ye, C., Yang, R., Gu, Q., and Zhang, T. Corruption-robust offline reinforcement learning with general function approximation. *arXiv preprint arXiv:2310.14550*, 2023b.

Zhang, T. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

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, 2022.

Zhao, H., Zhou, D., and Gu, Q. Linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2110.12615*, 2021.

Zhao, H., He, J., and Gu, Q. A nearly optimal and low-switching algorithm for reinforcement learning with general function approximation. *arXiv preprint arXiv:2311.15238*, 2023.

Zhong, H., Xiong, W., Tan, J., Wang, L., Zhang, T., Wang, Z., and Yang, Z. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *International Conference on Machine Learning*, pp. 27117–27142. PMLR, 2022a.

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

Zhou, Z., Zhou, Z., Bai, Q., Qiu, L., Blanchet, J., and Glynn, P. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339. PMLR, 2021.

A. Proof of Online Setting

A.1. Proof of Theorem 4.1

We delay the proof of the supporting lemmas in Appendix C.1

Lemma A.1 (Bellman Decomposition). *For any $M \in \mathcal{M}$ and any $t \in [t]$, we have*

$$V_M^1(x^1) - V_{\pi_M}^1(x^1) \le \mathbb{E}_{\pi_M}^t \sum_{h=1}^H \big[\mathcal{E}^h(M, x^h, a^h) + c_t^h(x^h, a^h)\big],$$

and

$$V_M^1(x^1) - V_{\pi_M}^1(x^1) \ge \mathbb{E}_{\pi_M}^t \sum_{h=1}^H \big[\mathcal{E}^h(M, x^h, a^h) - c_t^h(x^h, a^h)\big].$$

We demonstrate the relationship between the weighted IR defined in (5) and eluder dimension in Definition 3.3. For simplicity, we consider a single step in the following lemma.

Lemma A.2 (Relation between Information Ratio and Eluder Dimension). *Consider a model class $\mathcal{M}$ and sample set $\mathcal{S}_T = \{z_t\}_{t=1}^T$. Let $\alpha = \sqrt{\log|\mathcal{M}|}/C$ and $\lambda = \log|\mathcal{M}|$. The square sum of weighted IR with weight established in Algorithm 1*

$$I_\sigma(\lambda, \mathcal{M}, \mathcal{S}_t) = \sup_{M \in \mathcal{M}_t} \min\left\{1, \frac{\mathrm{TV}(P_M(\cdot|z_t)\|P_{\bar{M}_t}(\cdot|z_t))/\sigma_t^{1/2}}{\sqrt{\lambda + \sum_{s=1}^{t-1} \mathrm{TV}(P_M(\cdot|z_s)\|P_{\bar{M}_t}(\cdot|z_s))^2/\sigma_s}}\right\},$$

$$\sup_{\mathcal{S}_T} \dim_{E,\sigma}(\lambda, \mathcal{M}, \mathcal{S}_T) = \sup_{\mathcal{S}_T} \sum_{t=1}^T I_\sigma(\lambda, \mathcal{M}, \mathcal{S}_t)^2 = \tilde{O}(\mathrm{ED}(\mathcal{M}, \sqrt{\lambda/T}))$$

Ultimately, we can prove Theorem 4.1.

Proof of Theorem 4.1. Define the event

$$A_1 = \Big\{M_* \in \mathcal{M}_t, \text{ and } \sum_{s=1}^{t-1} \mathrm{TV}(P_*^h(\cdot|z_s^h)\|P_{\bar{M}_t}^h(\cdot|z_s^h))^2/\sigma_s^h \le 2\beta^2, \ \forall\, t \in [T]\Big\}.$$

According to Lemma A.1, we know that A_1 holds with probability at least $1-\delta$. Assuming that A_1 holds and using Lemma A.1, we have

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^T V_*^1(x_t^1) - V_{M_t}^1(x_t^1) + V_{M_t}^1(x_t^1) - V_{\pi_t}^1(x_t^1) \\
&\le \sum_{t=1}^T V_{M_t}^1(x_t^1) - V_{\pi_t}^1(x_t^1) \\
&\le \sum_{t=1}^T \mathbb{E}_{\pi_t}^t \sum_{h=1}^H \mathcal{E}^h(M_t, z_t^h) + \sum_{t=1}^T \mathbb{E}_{\pi_t}^t \sum_{h=1}^H c_t^h(x^h, a^h) \\
&\le \sum_{t=1}^T \mathbb{E}_{\pi_t}^t \sum_{h=1}^H \big[\mathbb{E}^{M_t}[V_{M_t}^{h+1}(x^{h+1})|z_t^h] - \mathbb{E}^*[V_{M_t}^{h+1}(x^{h+1})|z_t^h]\big] + CH \\
&\le \sum_{t=1}^T \mathbb{E}_{\pi_t}^t \sum_{h=1}^H \mathrm{TV}\big(P_{M_t}^h(\cdot|z_t^h)\|P_{\bar{M}_t}^h(\cdot|z_t^h)\big) + \sum_{t=1}^T \mathbb{E}_{\pi_t}^t \sum_{h=1}^H \mathrm{TV}\big(P_*^h(\cdot|z_t^h)\|P_{\bar{M}_t}^h(\cdot|z_t^h)\big) + CH,
\end{aligned}$$

where the first inequality uses $M_* \in \mathcal{M}_t$ and $M_t = \operatorname{argmax}_{M\in\mathcal{M}_t} V_M^1(x_t^1)$, and the last inequality uses $V_{M_t}^{h+1}(\cdot) \in [0,1]$. For any $M \in \mathcal{M}_t$, we have with probability at least $1-\delta$,

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\sum_{h=1}^{H}\mathrm{TV}\big(P_M^h(\cdot|z_t^h)\|P_{\bar{M}_t}^h(\cdot|z_t^h)\big)\\
&\quad\le \sum_{h=1}^{H}\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\min\left\{1,\frac{\mathrm{TV}\big(P_M^h(\cdot|z_t^h)\|P_{\bar{M}_t}^h(\cdot|z_t^h)\big)}{\sqrt{\lambda+\sum_{s=1}^{t-1}\mathrm{TV}\big(P_M^h(\cdot|z_s^h)\|P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2/\sigma_s^h}}\sqrt{\lambda+\sum_{s=1}^{t-1}\frac{\mathrm{TV}\big(P_M^h(\cdot|z_s^h)\|P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2}{\sigma_s^h}}\right\}\\
&\quad\le 3\sum_{h=1}^{H}\beta\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\min\left\{1,\frac{\mathrm{TV}\big(P_M^h(\cdot|z_t^h)\|P_{\bar{M}_t}^h(\cdot|z_t^h)\big)}{\sqrt{\lambda+\sum_{s=1}^{t-1}\mathrm{TV}\big(P_M^h(\cdot|z_s^h)\|P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2/\sigma_s^h}}\right\}\\
&\quad= \underbrace{3\sum_{h=1}^{H}\beta\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\min\left\{1,\frac{\mathrm{TV}\big(P_M^h(\cdot|z_t^h)\|P_{\bar{M}_t}^h(\cdot|z_t^h)\big)}{\sqrt{\lambda+\sum_{s=1}^{t-1}\mathrm{TV}\big(P_M^h(\cdot|z_s^h)\|P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2/\sigma_s^h}}\mathbb{1}\{\sigma_t^h=1\}\right\}}_{P_1}\\
&\qquad+\underbrace{3\sum_{h=1}^{H}\beta\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\min\left\{1,\frac{\mathrm{TV}\big(P_M^h(\cdot|z_t^h)\|P_{\bar{M}_t}^h(\cdot|z_t^h)\big)}{\sqrt{\lambda+\sum_{s=1}^{t-1}\mathrm{TV}\big(P_M^h(\cdot|z_s^h)\|P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2/\sigma_s^h}}\mathbb{1}\{\sigma_t^h>1\}\right\}}_{P_2},
\end{aligned}
\tag{6}
$$

where the second inequality holds by invoking Lemma 4.6:

$$
\sum_{s=1}^{t-1}\mathrm{TV}\big(P_M^h(\cdot|z_s^h)\|P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2/\sigma_s^h \le 4\beta^2.
$$

Then, we bound terms P_1 and P_2 separately. For term P_1,

$$
\begin{aligned}
P_1 &= \sum_{h=1}^{H}\sum_{t=1}^{T}\beta\mathbb{E}_{\pi_t}^t\min\left\{1,\frac{\mathrm{TV}\big(P_M^h(\cdot|z_t^h)\|P_{\bar{M}_t}^h(\cdot|z_t^h)\big)/(\sigma_t^h)^{1/2}}{\sqrt{\lambda+\sum_{s=1}^{t-1}\mathrm{TV}\big(P_M^h(\cdot|z_s^h)\|P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2/\sigma_s^h}}\mathbb{1}\{\sigma_t^h=1\}\right\}\\
&\le \sqrt{\sum_{h=1}^{H}\sum_{t=1}^{T}\beta^2}\cdot\sqrt{\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\sum_{h=1}^{H}\big(I_\sigma^h(\lambda,\mathcal{M}_t,\mathcal{S}_t)\big)^2}\\
&\le \sqrt{TH}\Big(5\sqrt{\log(|\mathcal{M}|/\delta)\log^2 B}+7\alpha C\Big)\sqrt{\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\sum_{h=1}^{H}\big(I_\sigma^h(\lambda,\mathcal{M}_t,\mathcal{S}_t)\big)^2}.
\end{aligned}
\tag{7}
$$

For term P_2,

$$
\begin{aligned}
P_2 &= \sum_{h=1}^{H}\beta\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\min\left\{1,\frac{\mathrm{TV}\big(P_M^h(\cdot|z_t^h)\|P_{\bar{M}_t}^h(\cdot|z_t^h)\big)/(\sigma_t^h)^{1/2}}{\sqrt{\lambda+\sum_{s=1}^{t-1}\mathrm{TV}\big(P_M^h(\cdot|z_s^h)\|P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2/\sigma_s^h}}(\sigma_t^h)^{1/2}\mathbb{1}\{\sigma_t^h>1\}\right\}\\
&\le \sum_{h=1}^{H}\beta\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\min\left\{1,\sup_{M\in\mathcal{M}_t^h}\left(\frac{\mathrm{TV}\big(P_M^h(\cdot|z_t^h)\|P_{\bar{M}_t}^h(\cdot|z_t^h)\big)/(\sigma_t^h)^{1/2}}{\sqrt{\lambda+\sum_{s=1}^{t-1}\mathrm{TV}\big(P_M^h(\cdot|z_s^h)\|P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2/\sigma_s^h}}\right)^2\frac{1}{\alpha}\right\}\\
&= \frac{1}{\alpha}\sum_{h=1}^{H}\beta\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\big(I_\sigma^h(\lambda,\mathcal{M}_t,\mathcal{S}_t)\big)^2\\
&= \frac{1}{\alpha}\Big(5\sqrt{\log(|\mathcal{M}|/\delta)\log^2 B}+7\alpha C\Big)\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\sum_{h=1}^{H}\big(I_\sigma^h(\lambda,\mathcal{M}_t,\mathcal{S}_t)\big)^2,
\end{aligned}
\tag{8}
$$

where the inequality uses the definition of σ_t^h when $\sigma_t^h > 1$:

$$(\sigma_t^h)^{1/2} = \frac{1}{\alpha} \sup_{M \in \mathcal{M}_t} \frac{\mathrm{TV}\big(P_M^h(\cdot|x_t^h, a_t^h) \| P_{M_t}^h(\cdot|x_t^h, a_t^h)\big)/(\sigma_t^h)^{1/2}}{\sqrt{\lambda + \sum_{s=1}^{t-1} \mathrm{TV}\big(P_M^h(\cdot|x_s^h, a_s^h) \| P_{M_t}^h(\cdot|x_s^h, a_s^h)\big)^2/\sigma_s^h}}$$

By taking (7) and (8) back into (6) and taking M as M_* and M_t, respectively, we obtain with probability at least $1-\delta$,

$$\begin{aligned}
\mathrm{Reg}(T) &\le 6\sqrt{TH}\Big(5\sqrt{\log(|\mathcal{M}|/\delta)\log^2 B} + 7\alpha C\Big)\sqrt{\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t \sum_{h=1}^{H}\big(I_\sigma^h(\lambda, \mathcal{M}_t, \mathcal{S}_t)\big)^2} \\
&\quad + \frac{6}{\alpha}\Big(5\sqrt{\log(|\mathcal{M}|/\delta)\log^2 B} + 7\alpha C\Big)\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t \sum_{h=1}^{H}\big(I_\sigma^h(\lambda, \mathcal{M}_t, \mathcal{S}_t)\big)^2 + CH \\
&= \tilde{\mathcal{O}}\left(\sqrt{TH\log|\mathcal{M}|\sum_{h=1}^{H}\sup_{\mathcal{S}_T}\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\big(I_\sigma^h(\lambda, \mathcal{M}_t, \mathcal{S}_t)\big)^2} + C\sum_{h=1}^{H}\sup_{\mathcal{S}_T}\sum_{t=1}^{T}\mathbb{E}_{\pi_t}^t\big(I_\sigma^h(\lambda, \mathcal{M}_t, \mathcal{S}_t)\big)^2\right) \\
&= \tilde{\mathcal{O}}\left(H\sqrt{T\log|\mathcal{M}|\mathrm{ED}(\mathcal{M}, \sqrt{\lambda/T})} + CH\cdot\mathrm{ED}(\mathcal{M}, \sqrt{\lambda/T})\right),
\end{aligned}$$

where we take $\alpha = \sqrt{\log|\mathcal{M}|\log^2 B}/C$. □

A.2. Proof of Theorem 4.2

To prove the lower bound, we first create a series of hard-to-learn tabular MDP $(\mathcal{X}, \mathcal{A}, H, \mathbb{P}, r)$ with 4 different state x_0, x_1, x_2, x_3, where x_2, x_3 are absorbing states, and d different actions $\mathcal{A} = \{a_1, ..., a_d\}$. For initial state x_0 and each stage $h \in [H]$, the agent will stay at the state x_0 with probability $1 - 1/H$ and transition to the state x_1 with probability $1/H$. For state x_1 and each stage $h \in [H]$, the agent will only transit to state x_2 or x_3. In addition, an optimal action a_h^* is selected from the action set $\mathcal{A} = \{a_1, ..., a_d\}$ and the transition probability at stage h can be denoted by

$$P_{a_h^*}^h(x_2|x_1, a) = \begin{cases} \frac{3}{4}, & a = a_h^* \\ \frac{1}{4}, & a \neq a_h^*, \end{cases}$$

The agent is rewarded with 1 at the absorbing state x_2 during the final stage H, while receiving zero reward otherwise. Consequently, state x_2 can be regarded as the goal state and the optimal strategy at stage h involves taking action a_h^* to maximize the probability of reaching the goal state x_2. Consequently, taking a not-optimal action $a_h \neq a_h^*$ at state x_1 for stage $2 \le h \le H-1$ will incur a $1/2$ regret in a episode.

For these hard-to-learn MDPs, we can divide the episodes to several groups $\mathcal{T}_1, .., \mathcal{T}_H$, where episodes can be categorized into distinct groups, denoted as $\mathcal{T}_1, \ldots, \mathcal{T}_H$, where $\mathcal{T}_i (i \in [H-1])$ comprises episodes transitioning to state x_1 after taking action at stage i, and $\mathcal{T}_H$ includes episodes where the agent never reaches state x_1. Intuitively, the learning problem can be seen as the task of learning H distinct linear bandit problems, with the goal of determining the optimal action a_h^* for each stage $h \in [H]$. Now, considering a fixed stage $2 \le h \le H$, let $t_1, t_2, ..., t_{dC/2} \in \mathcal{T}_{h-1}$ be the first $dC/2$ episodes that transition to state x_1 after taking action at stage $h-1$. To ensure an adequate number of episodes in each group $\mathcal{T}_{h-1}$, we continue the learning problem indefinitely, while our analysis focuses only on the regret across the first T episodes. In this scenario, it is guaranteed, with probability 1, that each group $\mathcal{T}_{h-1}$ consists of at least $dC/2$ episodes. The following lemma provides a lower bound for learning the optimal action a_h^* in the presence of adversarial corruption.

Lemma A.3. *For each fixed stage $2 \le h \le H-1$ and any algorithm **Alg** with the knowledge of corruption level C, if the optimal action a_h^* is uniformed random selected from the action set $\mathcal{A} = \{a_1, ..., a_d\}$, then the expected regret across episode $t_1, ..., t_{dC/2}$ is at least $(d-1)C/16$.*

Proof of Lemma A.3. To proof the lower bound, we construct an auxiliary transition probability function P_0^h such that

$$P_0^h(x_2|x_1, a) = \frac{1}{4}, \forall a \in \mathcal{A}.$$

Now, the following corruption strategy is employed for the the transition probability $P^h_{a^*_h}$: if the optimal action a^*_h is selected at stage h and the total corruption level for stage h up to the previous step is no more than $C-2$, then the adversary corrupts the transition probability $P^h_{a^*_h}$ to P^h_0 and the corruption level c^h_t in this episode is $c^h_t = |P^h_{a^*_h}(x_2|x_1, a^*_h)/P^h_0(x_2|x_1, a^*_h) - 1| = 2$. In this case, regardless of the actual optimal action a^*_h, there is no distinction between $Pa^{*\,h}_h$ and P^h_0, unless the agent selects the optimal action a^*_h at least $C/2$ times, and the adversary lacks sufficient corruption level to corrupt the transition probability. Now, let's consider the execution of the algorithm **Alg** on the uncorrupted transition probability P^h_0. By the pigeonhole principle, there exist at least $d/2$ different arms, whose expected selected time is less than $C/2$ times. Without loss of generality, we assume these actions are $a_1, ..., a_{d/2}$. Then for each action $a_i (i \leq d/2)$, according to Markov inequality, with probability at least $1/2$, the number of episodes selecting a_i at stage h is less than $C/2$. Under this scenario, the performance of **Alg** on transition probability P^h_0 is equivalent to its performance on transition probability $P^h_{a_i}$ and the regret in these episode is at least $(dC/2 - C/2) \times 1/2 = (d-1)C/4$. Therefore, if the optimal action a^*_h is uniform randomly selected from the action set $\mathcal{A} = \{a_1, ..., a_d\}$, then the expected regret across episodes $t_1, .., t_{dC/2}$ is lower bound by

$$\mathbb{E}\Big[\sum_{i=1}^{dC/2} \mathbb{1}(a^h_{t_i} \neq a^*_h) \cdot 1/2\Big] \geq \frac{1}{2} \times \frac{1}{2} \times \frac{(d-1)C}{4} = \frac{(d-1)C}{16}.$$

Thus, we complete the proof of Lemma A.3. □

Lemma A.3 provides a lower bound on the expected regret for each stage $2 \leq h \leq H-1$ over the first $dC/2$ visits to state x_1. Regrettably, Lemma A.3 does not directly yield a lower bound for the regret over the initial T episodes, as the agent may not visit the state x_1 $dC/2$ times for some stage $h \in [H]$. The following lemma posits that as the number of episodes T increases, the occurrence of this event becomes highly improbable.

Lemma A.4. *For the proposed hard-to-learn MDPs, if the number of episodes T satisfies $T > edCH + 2e^2H^2\log(1/\delta)$, then for each stage $2 \leq h \leq H-1$, with a probability of at least $1-\delta$, the agent visits state x_1 at least $dC/2$ times.*

Proof of Lemma A.4. For any episode T and stage $h \in [H]$, the agent, starting from the current state x_0, will transition to state x_1 with a probability of $1/H$, regardless of the selected action. Consequently, the probability that the agent visits state x_1 at stage h can be expressed as:

$$P(x_h = x_1) = \left(1 - \frac{1}{H}\right)^{h-1} \cdot \frac{1}{H} \geq \frac{1}{eH}, \forall 2 \leq h \leq H-1.$$

For a fixed stage $2 \leq h \leq H-1$, we define the random variable $y_i = \mathbb{1}(x^h_i = x_1)$ as the indicator function of visiting the state x_1 at stage h of episode i. Subsequently, leveraging the Azuma–Hoeffding inequality (Lemma E.1), with a probability of at least $1-\delta$, we obtain:

$$\sum_{i=1}^{T} y_i \geq \frac{T}{eH} - \sqrt{2T\log(1/\delta)} \geq \frac{T}{2eH} - eH\log(1/\delta),$$

where the last inequality holds due to $x^2 + y^2 \geq 2xy$. Therefore, for $T > edCH + 2e^2H^2\log(1/\delta)$, with probability at least $1-\delta$, we have

$$\sum_{i=1}^{T} y_i \geq dC/2,$$

which completes the proof of Lemma A.4. □

With the help of these lemmas, we are able to prove Theorem 4.2.

Proof of Theorem 4.2. For each stage $2 \leq h \leq H-1$, we use $\mathcal{E}_h$ to denote the events that the agent visit the state x_1 at least $dC/2$ times for stage h across the first T episodes. Under this situation, we use $t_{h,1}, ..., t_{h,dC/2}$ denotes the first $dC/2$ episodes that visit the state x_1 at stage h.

Then for any algorithm, the expected regret can be lower bounded by

$$\begin{aligned}
\mathbb{E}[\text{Regret(T)}] &= \mathbb{E}\Big[\sum_{t=1}^{T}\sum_{h=2}^{H-1}\mathbb{1}(x_t^h = x_1)\cdot\mathbb{1}(a_t^h \neq a_h^*)\cdot 1/2\Big] \\
&\geq \sum_{h=2}^{H-1}\mathbb{1}(\mathcal{E}_h)\mathbb{E}\Big[\sum_{t=1}^{T}\mathbb{1}(x_t^h = x_1)\cdot\mathbb{1}(a_t^h \neq a_h^*)\cdot 1/2|\mathcal{E}_h\Big] \\
&\geq \sum_{h=2}^{H-1}\mathbb{1}(\mathcal{E}_h)\mathbb{E}\Big[\sum_{i=1}^{dC/2}\mathbb{1}(a_{t_{h,i}}^h \neq a_h^*)\cdot 1/2|\mathcal{E}_h\Big] \\
&= \sum_{h=2}^{H-1}\mathbb{E}\Big[\sum_{i=1}^{dC/2}\mathbb{1}(a_{t_{h,i}}^h \neq a_h^*)\cdot 1/2\Big] - \sum_{h=2}^{H-1}\mathbb{1}(\neg\mathcal{E}_h)\mathbb{E}\Big[\sum_{i=1}^{dC/2}\mathbb{1}(a_{t_{h,i}}^h \neq a_h^*)\cdot 1/2|\neg\mathcal{E}_h\Big] \\
&\geq \frac{(H-2)(d-1)C}{32} - \sum_{h=2}^{H-1}\mathbb{1}(\neg\mathcal{E}_h)\mathbb{E}\Big[\sum_{i=1}^{dC/2}\mathbb{1}(a_{t_{h,i}}^h \neq a_h^*)\cdot 1/2|\neg\mathcal{E}_h\Big] \\
&\geq \frac{(H-2)(d-1)C}{32} - \frac{(H-2)\delta dC}{4}
\end{aligned} \tag{9}$$

where the first equation holds due to the construction of these hard-to-learn MDPs, the first inequality holds due to the fact that $\mathbb{E}[x] \leq \mathbb{E}[x|\mathcal{E}]\cdot\Pr(\mathcal{E})$, the second inequality holds due to the definition of event $\mathcal{E}_h$, the third inequality holds due to Lemma A.3 and the second inequality holds due to Lemma A.4. Finally, setting the probability $\delta = 1/32$, we have

$$\mathbb{E}[\text{Regret(T)}] \geq \frac{(H-2)(d-1)C}{64}.$$

Thus, we complete the proof of Theorem 4.2. □

A.3. Unknown Corruption Level

Proof of Theorem 4.4. First of all, we consider the case when $C \leq \bar{C}$. Since only the hyper-parameter α is modified by replacing C with $\bar{C}$ and $C \leq \bar{C}$, we can follow the analysis of Theorem 4.1 to derive the suboptimality bound with probability at least $1-\delta$,

$$\begin{aligned}
\mathrm{Reg}(T) =& \tilde{\mathcal{O}}\Big(H\log B\sqrt{T\log|\mathcal{M}|\mathrm{ED}(\mathcal{M},\sqrt{\lambda/T})} \\
&+ \bar{C}H\cdot\mathrm{ED}(\mathcal{M},\sqrt{\lambda/T})\Big).
\end{aligned}$$

Additionally, we can also demonstrate the relationship between the sum of weighted information ratio and the eluder dimension as Lemma A.2 by discussing the value of $\bar{C}$ in the first step.

For the case when $C > \bar{C}$, we simply take the trivial bound T. □

B. Proof of Offline Setting

B.1. Proof of Theorem 5.2

First of all, we demonstrate that for the uncertainty weight iteration in Algorithm 3 converges and the output solution approximates the real uncertainty quantity. Since the convergence has been illustrated in Ye et al. (2023b), we restate the result in the following lemma.

Lemma B.1 (Lemma 3.1 of Ye et al. (2023a)). *There exists a T such that the output of Algorithm 3 $\{\sigma_t := \sigma_t^{K+1}\}_{t=1}^T$ satisfies:*

$$\sigma_t \geq \max\big\{1, \psi(z_t)/2\big\}, \quad \sigma_t \leq \max\big\{1, \psi(z_t)\big\}, \tag{10}$$

where

$$\psi(z_t) = \sup_{M,M'\in\mathcal{M}} \frac{\mathrm{TV}(P_M^h(\cdot|z_t)\|P_{M'}^h(\cdot|z_t))/\alpha}{\sqrt{\lambda + \sum_{s=1}^T \mathrm{TV}(P_M^h(\cdot|z_s)\|P_{M'}^h(\cdot|z_s))^2/\sigma_s}}.$$

We provide the proof in Appendix C.2.

Lemma B.2. *Under Assumption 3.1 and Algorithm 2, if we choose*

$$\beta = 5\sqrt{\log(|\mathcal{M}|/\delta)\log^2 B} + 7\alpha C,$$

we have with probability at least $1-\delta$*, for all* $h \in [H]$ *and all* $t \in [T]$*,* $M_* \in \widehat{\mathcal{M}}$ *and*

$$\sum_{t=1}^{T} \mathrm{TV}(P_*^h(\cdot|z_t^h) \| P_{\widehat{M}}^h(\cdot|z_t^h))^2/\sigma_t^h \le 2\beta^2.$$

Moreover, for any $M \in \mathcal{M}_t$*, we have with probability at least* $1-\delta$*,*

$$\sum_{t=1}^{T} \mathrm{TV}(P_M^h(\cdot|z_t^h) \| P_{\widehat{M}}^h(\cdot|z_t^h))^2/\sigma_t^h \le 4\beta^2.$$

We show the proof in Appendix C.2.

Proof of Theorem 5.2. Define the event

$$A_2 = \Big\{ M_* \in \mathcal{M}_t, \text{ and } \sum_{t=1}^{T} \mathrm{TV}(P_*^h(\cdot|z_s^h) \| P_{M_t}^h(\cdot|z_s^h))^2/\sigma_s^h \le 2\beta^2 \Big\}.$$

According to Lemma A.1, we know that A_1 holds with probability at least $1-\delta$. First of all, we obtain

$$\begin{aligned}
\mathrm{SubOpt}(\hat{\pi}, x^1) =& V_*^1(x^1) - V_{\widehat{M}}^1(x^1) + V_{\widehat{M}}^1(x^1) - V_{\hat{\pi}}^1(x^1) \\
\le& V_*^1(x^1) - V_{\widehat{M}}^1(x^1) \\
\le& \mathbb{E}_{\pi_*} \sum_{h=1}^{H} \big(V_{\widehat{M},\pi_*}^h(x^h) - V_{\widehat{M},\hat{\pi}}^h(x^h) \big) - \mathbb{E}_{\pi_*} \sum_{h=1}^{H} \mathcal{E}^h(\widehat{M}, z^h) \\
\le& -\mathbb{E}_{\pi_*} \sum_{h=1}^{H} \big(\mathbb{E}^{\widehat{M}}[V_{\widehat{M}}^{h+1}(x^{h+1})|z^h] - \mathbb{E}^{M_*}[V_{\widehat{M}}^{h+1}(x^{h+1})|z^h] \big) \\
\le& \mathbb{E}_{\pi_*} \sum_{h=1}^{H} \mathrm{TV}(P_{\widehat{M}}^h(\cdot|z^h) \| P_{\widehat{M}}^h(\cdot|z^h)) + \mathbb{E}_{\pi_*} \sum_{h=1}^{H} \mathrm{TV}(P_{M_*}^h(\cdot|z^h) \| P_{\widehat{M}}^h(\cdot|z^h))
\end{aligned}$$

and the first inequality uses the condition that $M_* \in \widehat{\mathcal{M}}$, which implies that $V_{\widehat{\mathcal{M}},\hat{\pi}}^1(x^1) \le V_{*,\hat{\pi}}^1(x^1)$, and the second inequality applies Lemma A.1 with $c_t^h = 0$.

Hence, for any $M \in \widehat{\mathcal{M}}$, we have with probability at least $1-\delta$,

$$
\begin{aligned}
&\mathbb{E}_{\pi_*} \sum_{h=1}^{H} \mathrm{TV}(P_M^h(\cdot|z^h) \| P_{\bar{M}}^h(\cdot|z^h)) \\
&\quad \le \mathbb{E}_{\pi_*} \sum_{h=1}^{H} \min\left\{1, \frac{\mathrm{TV}\big(P_M^h(\cdot|z^h) \| P_{\bar{M}}^h(\cdot|z^h)\big)}{\sqrt{\lambda + \sum_{t=1}^{T} \mathrm{TV}\big(P_M^h(\cdot|z_t^h) \| P_{\bar{M}}^h(\cdot|z_t^h)\big)^2/\sigma_t^h}} \sqrt{\lambda + \sum_{t=1}^{T} \frac{\mathrm{TV}\big(P_M^h(\cdot|z_t^h) \| P_{\bar{M}}^h(\cdot|z_t^h)\big)^2}{\sigma_t^h}}\right\} \\
&\quad \le 3\beta \mathbb{E}_{\pi_*} \sum_{h=1}^{H} \frac{\mathrm{TV}\big(P_M^h(\cdot|z^h) \| P_{\bar{M}}^h(\cdot|z^h)\big)}{\sqrt{\lambda + \sum_{t=1}^{T} \mathrm{TV}\big(P_M^h(\cdot|z_t^h) \| P_{\bar{M}}^h(\cdot|z_t^h)\big)^2/\sigma_t^h}} \\
&\quad = 3\beta \mathbb{E}_{\pi_*} \sum_{h=1}^{H} \Bigg[\frac{\mathrm{TV}\big(P_M^h(\cdot|z^h) \| P_{\bar{M}}^h(\cdot|z^h)\big)}{\sqrt{\lambda + \sum_{t=1}^{T} \mathrm{TV}\big(P_M^h(\cdot|z_t^h) \| P_{\bar{M}}^h(\cdot|z_t^h)\big)^2/\sigma_t^h}} \mathbb{1}(\sigma^h(z^h) = 1) \\
&\qquad\quad + \frac{\mathrm{TV}\big(P_M^h(\cdot|z^h) \| P_{\bar{M}}^h(\cdot|z^h)\big)/\sigma^h(z^h)}{\sqrt{\lambda + \sum_{t=1}^{T} \mathrm{TV}\big(P_M^h(\cdot|z_t^h) \| P_{\bar{M}}^h(\cdot|z_t^h)\big)^2/\sigma_t^h}} \sigma^h(z^h) \mathbb{1}(\sigma^h(z^h) > 1) \Bigg] \\
&\quad = 3\beta \mathbb{E}_{\pi_*} \sum_{h=1}^{H} \Bigg[\sup_{M, M' \in \widehat{\mathcal{M}}} \frac{\mathrm{TV}\big(P_M^h(\cdot|z^h) \| P_{\bar{M}}^h(\cdot|z^h)\big)}{\sqrt{\lambda + \sum_{t=1}^{T} \mathrm{TV}\big(P_M^h(\cdot|z_t^h) \| P_{\bar{M}}^h(\cdot|z_t^h)\big)^2/\sigma_t^h}} \\
&\qquad\quad + \frac{1}{\alpha} \Big(\sup_{M, M' \in \widehat{\mathcal{M}}} \frac{\mathrm{TV}\big(P_M^h(\cdot|z^h) \| P_{\bar{M}}^h(\cdot|z^h)\big)}{\sqrt{\lambda + \sum_{t=1}^{T} \mathrm{TV}\big(P_M^h(\cdot|z_t^h) \| P_{\bar{M}}^h(\cdot|z_t^h)\big)^2/\sigma_t^h}} \Big)^2 \Bigg],
\end{aligned} \tag{11}
$$

where the second inequality uses Lemma B.2, and the last inequality holds due to the definition of $\sigma^h(z^h)$ when $\sigma^h(z^h) > 1$:

$$
\sigma^h(z^h) = \frac{1}{\sigma^h(z^h)} \sup_{M, M' \in \widehat{\mathcal{M}}} \frac{\mathrm{TV}\big(P_M^h(\cdot|z^h) \| P_{M'}^h(\cdot|z^h)\big)/\alpha}{\sqrt{\lambda + \sum_{t=1}^{T} \mathrm{TV}\big(P_M^h(\cdot|z_t^h) \| P_{\bar{M}}^h(\cdot|z_t^h)\big)^2/\sigma_t^h}}.
$$

Further, by defining the weighted form of information coefficient

$$
\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D}) = \sup_{M, M' \in \widehat{\mathcal{M}}} \mathbb{E}_{\pi_*} \left[\frac{T \cdot \mathrm{TV}(P_M^h(\cdot|z^h) \| P_{M'}^h(\cdot|z^h))/\sigma^h(z^h)^{1/2}}{\lambda + \sum_{t=1}^{T} \mathrm{TV}(P_M^h(\cdot|z_t^h) \| P_{M'}^h(\cdot|z_t^h))^2/\sigma_t^h} \middle| x^1 = x \right],
$$

we deduce that

$$
\mathbb{E}_{\pi_*} \sum_{h=1}^{H} \mathrm{TV}(P_M^h(\cdot|z^h) \| P_{\bar{M}}^h(\cdot|z^h)) \le 3\beta H \left[\sqrt{\frac{\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D})}{T}} + \frac{\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D})}{\alpha T} \right].
$$

Then, we use the inequality above with $M = M_*$ and $\widehat{M}$ to get

$$
\begin{aligned}
\mathrm{SubOpt}(\hat{\pi}, x^1) \le & 6\beta H \left[\sqrt{\frac{\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D})}{T}} + \frac{\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D})}{\alpha T} \right] \\
= & 6\big(5\sqrt{\log(|\mathcal{M}|/\delta) \log^2 B} + 7\alpha C\big) H \left[\sqrt{\frac{\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D})}{T}} + \frac{\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D})}{\alpha T} \right] \\
= & \tilde{O}\Big(\frac{H \sqrt{\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D}) \log(|\mathcal{M}|/\delta) \log^2 B}}{T} + \frac{\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D}) C H}{T} \Big),
\end{aligned}
$$

where we take $\alpha = \sqrt{\log |\mathcal{M}| \log^2 B}/C$. Therefore, we complete the proof. □

B.2. Proof of Theorem 5.4

Then, we will analyze the relationship between the weighted information coefficient $\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D})$ and $\mathrm{Cov}(\mathcal{M}, \mathcal{D})$.

Lemma B.3. *Under Assumption 5.3, we have*

$$\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D}) \leq \frac{1}{\mathrm{Cov}(\mathcal{M}, \mathcal{D})}.$$

The proof is presented in Appendix C.2.

Now, we are ready to prove Theorem 5.4.

Proof of Theorem 5.4. According to Theorem 5.2, we have with probability at least $1-2\delta$,

$$\mathrm{SubOpt}(\hat{\pi}, x^1) = \tilde{O}\Big(\frac{H\sqrt{\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D})\log(|\mathcal{M}|/\delta)\log^2 B}}{T} + \frac{\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D})CH}{T}\Big).$$

By invoking Lemma B.3, we complete the proof. □

B.3. Offline Lower bound

In this section, we provide the lower bound for offline learning. Here, we used the hard-to-learn instance in Section A.2, while modified the transition probability at stage h from $1/4$ or $3/4$ to

$$P^h_{a^*_h}(x_2|x_1, a) = \begin{cases} \frac{1}{2}+\eta, & a = a^*_h \\ \frac{1}{2}-\eta, & a \neq a^*_h, \end{cases}$$

for different corruption level C and data-size T.

For the offline data collection process, we employ the behavior policy π^v, which with probability $1-\epsilon$ will select the action a_d, and with probability ϵ, uniform select an action from $\{a_1, .., a_{d-1}\}$. For each stage $h \in [H]$, the optimal action a^*_h is uniform randomly selected from the action set $\mathcal{A} = \{a_1, ..., a_{d-1}\}$ and we construct an auxiliary transition probability function P^h_0 without optimal action such that

$$P^h_0(x_2|x_1, a) = \frac{1}{2} - \eta, \forall a \in \mathcal{A}.$$

Then the following lemma provides upper and lower bounds for the visiting times of state-action pair (x_1, a).

Lemma B.4. *For each stage $h \in [H]$ and fixed action $a \in \{a_1, ..., a_{d-1}\}$, if the data-size T satisfied $T > 4e^2(d-1)^2H^2\log(1/\delta)/\epsilon^2$,then with probability at least $1-\delta$, the behavior policy visit the state x_1 and take action a no less than times $\epsilon T/\big(4eH(d-1)\big)$ during the data collection process. In addition, with probability at least $1-\delta$, the visiting times is no more than $3\epsilon T/\big(H(d-1)\big)$.*

Proof of Lemma B.4. For any episode T and stage $h \in [H]$, we define the random variable $y_i = \mathbb{1}(x^h_i = x_1, a^h_i = a)$ as the indicator function of visiting the state x_i and taking action a at stage h of episode i. Since the agent, starting from the current state x_0, will transition to state x_1 with a probability of $1/H$, regardless of the selected action, the probability that the agent visits state x_1 at stage h and take action a is upper bounded by

$$\Pr(y_i = 1) = \left(1 - \frac{1}{H}\right)^{h-1} \cdot \frac{1}{H} \cdot \frac{\epsilon}{d-1} \geq \frac{1}{eH} \cdot \frac{\epsilon}{d-1}.$$

Therefore, applying the Azuma–Hoeffding inequality (Lemma E.1), with a probability of at least $1-\delta$, we have:

$$\sum_{i=1}^{T} y_i \geq \frac{T}{eH} \cdot \frac{\epsilon}{d-1} - \sqrt{2T\log(1/\delta)} \geq \frac{T}{2eH} \cdot \frac{\epsilon}{d-1} - \frac{e(d-1)H}{\epsilon} \cdot \log(1/\delta),$$

where the last inequality holds due to $x^2 + y^2 \geq 2xy$. Therefore, for $T > 4e^2(d-1)^2H^2\log(1/\delta)/\epsilon^2$, with probability at least $1-\delta$, we have

$$\sum_{i=1}^{T} y_i \geq \frac{T}{4eH} \cdot \frac{\epsilon}{d-1},$$

which completes the proof of lower bounds in Lemma B.4.

On the other hand, for the upper bound, we have

$$\Pr(y_i = 1) = \left(1 - \frac{1}{H}\right)^{h-1} \cdot \frac{1}{H} \leq \frac{1}{H} \cdot \frac{\epsilon}{d-1}.$$

Similarly, applying the Azuma–Hoeffding inequality (Lemma E.1), with a probability of at least $1-\delta$, we have:

$$\sum_{i=1}^{T} y_i \leq \frac{T}{H} \cdot \frac{\epsilon}{d-1} + \sqrt{2T\log(1/\delta)} \leq \frac{2T}{H} \cdot \frac{\epsilon}{d-1} + \frac{(d-1)H}{2\epsilon} \cdot \log(1/\delta),$$

where the last inequality holds due to $x^2 + y^2 \geq 2xy$. Therefore, for $T > 4e^2(d-1)^2H^2\log(1/\delta)/\epsilon^2$, with probability at least $1-\delta$, we have

$$\sum_{i=1}^{T} y_i \leq \frac{3T}{H} \cdot \frac{\epsilon}{d-1},$$

which completes the proof of upper bounds in Lemma B.4. □

For simplicity, we denote $\mathcal{E}$ as the event where the high probability event in Lemma B.4 holds for all stage $h \in [H]$ and $a \in \{a_1, ..., a_{d-1}\}$. Now, for behavior policy π^v, the following corruption strategy is applied to the transition probability $P^h_{a^*_h}$: if the optimal action a^*_h is selected at stage h with the current state x_1, then the adversary corrupts the transition probability $P^h_{a^*_h}$ to P^h_0. Under this situation, the following lemma provides the upper bound of corruption level and lower bounds for the data coverage coefficient $\mathrm{Cov}(\mathcal{M}, \mathcal{D})$.

Lemma B.5. *Conditioned on the event $\mathcal{E}$, the corruption level is upper bounded by*

$$C \leq \frac{3\epsilon T}{(d-1)H} \cdot \frac{4\eta}{1-2\eta}.$$

In addition, the data coverage coefficient $\mathrm{Cov}(\mathcal{M}, \mathcal{D})$ is lower bounded by

$$\mathrm{Cov}(\mathcal{M}, \mathcal{D}) \geq \frac{\epsilon}{4eH(d-1)}.$$

Proof of Lemma B.5. According the definition of corruption strategy, if the optimal action a^*_h is selected at stage h with the current state x_1, the corresponding corruption level within this episode is denoted by

$$c^h_t = c^h_t(x_1, a^*_h) = \sup_{x^{h+1} \in \Delta_t(\mathcal{X})} \left| \frac{P^h_0(x^{h+1}|x_1, a^*_h)}{P^j_{a^*_h}(x^{h+1}|x_1, a^*_h)} - 1 \right| = \frac{4\eta}{1-2\eta}.$$

It is worth noting that the event $\mathcal{E}$ only focused on the transition before reaching the state x_1, and remains unaffected by the adversary corruption employed. Therefore, conditioned on the event $\mathcal{E}$, for each stage $h \in [H]$, the total corruption level throughout the offline data collection process is upper-bounded by

$$\sum_{i=1}^{T} c^h_t \leq \frac{3\epsilon T}{(d-1)H} \cdot \frac{4\eta}{1-2\eta}.$$

Thus, the corruption level for the employed strategy satisfied

$$C \leq \frac{3\epsilon T}{(d-1)H} \cdot \frac{4\eta}{1-2\eta}.$$

Regarding the data coverage coefficient $\mathrm{Cov}(\mathcal{M}, \mathcal{D})$, it is worth noting that all models $M \in \mathcal{M}$ share identical transition probabilities across states x_0, x_2, x_3. Consequently, we only need to focus on state x_1 and action $a \in \{a_1, ..., a_{d-1}\}$. Conditioned on event $\mathcal{E}$, we have

$$\begin{aligned}\rho^h(M, M')^2 &\leq \sum_{a \in \{a_1,\ldots,a_{d-1}\}} \mathrm{TV}(P_M^h(\cdot|x_1, a) \| P_{M'}^h(\cdot|x_1, a))^2 \\ &\leq \frac{4eH(d-1)}{\epsilon} \cdot \frac{1}{T} \sum_{t=1}^{T} \mathrm{TV}(P_M^h(\cdot|z_t^h) \| P_{M'}^h(\cdot|z_t^h))^2,\end{aligned}$$

where the second inequality holds due to the definition of events $\mathcal{E}$. Thus, we have

$$\mathrm{Cov}(\mathcal{M}, \mathcal{D}) \geq \frac{\epsilon}{4eH(d-1)},$$

and we complete the proof of Lemma B.5. □

With the help of these lemmas, we can start the proof of Theorem 5.6.

Proof of Theorem 5.6. According to the corruption strategy, regardless of the actual optimal action a_h^*, during the offline collection process, the behavior of the transition probability function is same as P_0^h. Rough speaking, the agent cannot outperform random guessing of the optimal action a_h^* and subsequently outputting the corresponding optimal policy $\hat{\pi}^h$. Thus, when the optimal action a_h^* is uniform randomly selected from the action set $\mathcal{A} = \{a_1, ..., a_{d-1}\}$, the sub-optimally gap of policy $\hat{\pi}$ can be denoted by

$$\begin{aligned}\mathbb{E}[\mathrm{SubOpt}(\hat{\pi}, x)] &= \mathbb{E}\Big[\sum_{h=2}^{H-1} \mathbb{1}(x^h = x_1) \cdot \mathbb{1}(\hat{\pi}^h(x_1) \neq a_h^*) \cdot 2\eta\Big] \\ &\geq \mathbb{E}\Big[\sum_{h=2}^{H-1} \frac{1}{eH} \cdot \mathbb{1}(\hat{\pi}^h(x_1) \neq a_h^*) \cdot 2\eta\Big] \\ &= \frac{H-1}{eH} \cdot \frac{d-2}{d-1} \cdot 2\eta.\end{aligned}$$

Now, for a given dataset size T, corruption level C and data coverage coefficient $\mathrm{Cov}(\mathcal{M}, \mathcal{D})$, according to Lemma B.5, we can select the parameter as following:

$$\epsilon = \mathrm{Cov}(\mathcal{M}, \mathcal{D}) \cdot 4eH(d-1), \eta = \frac{\mathrm{Cov}(d-1)H}{24\epsilon T} = \frac{C}{96e\mathrm{Cov}(\mathcal{M}, \mathcal{D})T}.$$

Then, if the dataset size T satisfied $T > C/(24e\mathrm{Cov}(\mathcal{M}, \mathcal{D}))$ and $d > 3, H > 2$, then we have

$$\mathbb{E}[\mathrm{SubOpt}(\hat{\pi}, x)] \geq \Omega(\eta) = \Omega\Big(\frac{C}{\mathrm{Cov}(\mathcal{M}, \mathcal{D})T}\Big).$$

Thus, we complete the proof of Theorem 5.6. □

C. Proof of Supporting Lemmas

C.1. Lemmas for Online Setting

Proof of Lemma 4.6. For simplicity, we assume that class $\mathcal{M}$ has finite elements. Let

$$E_t = \sum_{s=1}^{t} \mathrm{TV}\big(P_*^h(\cdot|z_s^h) \| P_{\widehat{M}_{t+1}}^h(\cdot|z_s^h)\big)^2 / \sigma_s^h.$$

For each time $s \in [t-1]$, we define the region $\mathcal{D}_s^h = \{x^{h+1} \in \mathcal{X} : \ P_s^h(x^{h+1}|z_s^h) \le P_*^h(x^{h+1}|z_s^h)\}$. For any fixed $M \in \mathcal{M}$, we have

$$\begin{aligned}
&\mathbb{E}\sum_{s=1}^{t-1}\frac{1}{\sigma_s^h}\log\sqrt{\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}} \\
&\quad=\underbrace{\sum_{s=1}^{t-1}\frac{1}{\sigma_s^h}\int_{\mathcal{D}_s^h}\mathrm{d}P_s^h(x^{h+1}|z_s^h)\log\sqrt{\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}}}_{P_1}+\underbrace{\sum_{s=1}^{t-1}\frac{1}{\sigma_s^h}\int_{\bar{\mathcal{D}}_s^h}\mathrm{d}P_s^h(x^{h+1}|z_s^h)\log\sqrt{\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}}}_{P_2}. \qquad (12)
\end{aligned}$$

For the term P_1, we derive

$$\begin{aligned}
P_1 &\le \frac{1}{2}\sum_{s=1}^{t-1}\frac{1}{\sigma_s^h}\int_{\mathcal{D}_s^h}\mathrm{d}P_*^h(x^{h+1}|z_s^h)\log\Big(\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}\Big) \\
&= -\frac{1}{2}\sum_{s=1}^{t-1}\frac{1}{\sigma_s^h}\int_{\mathcal{D}_s^h}\mathrm{d}P_*^h(x^{h+1}|z_s^h)\log\Big(\frac{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}\Big). \qquad (13)
\end{aligned}$$

For the term P_2, we have

$$\begin{aligned}
P_2 &= \sum_{s=1}^{t-1}\underbrace{\frac{1}{2\sigma_s^h}\int_{\bar{\mathcal{D}}_s^h}\big(\mathrm{d}P_s^h(x^{h+1}|z_s^h)-\mathrm{d}P_*^h(x^{h+1}|z_s^h)\big)\log\Big(\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}\Big)}_{Q} \\
&\quad-\sum_{s=1}^{t-1}\frac{1}{2\sigma_s^h}\int_{\bar{\mathcal{D}}_s^h}\mathrm{d}P_*^h(x^{h+1}|z_s^h)\log\Big(\frac{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}\Big), \qquad (14)
\end{aligned}$$

from which we further bound the term Q by

$$\begin{aligned}
Q &\le \frac{1}{2\sigma_s^h}\int_{\bar{\mathcal{D}}_s^h}\big(\mathrm{d}P_s^h(x^{h+1}|z_s^h)-\mathrm{d}P_*^h(x^{h+1}|z_s^h)\big)\cdot\left(\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}-1\right) \\
&\le \frac{1}{2\sigma_s^h}\sup_{x^{h+1}\in\bar{\mathcal{D}}_s^h}\Big|\frac{\mathrm{d}P_s^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}-1\Big|\int_{\bar{\mathcal{D}}_s^h}\Big|\mathrm{d}P_M^h(x^{h+1}|z_s^h)-\mathrm{d}P_*^h(x^{h+1}|z_s^h)\Big| \\
&\le \frac{c_s^h}{2}\cdot\frac{\mathrm{TV}\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big)}{\sigma_s^h}. \qquad (15)
\end{aligned}$$

Therefore, by combining (13), (14) and (15) and take them back into (12), we obtain that

$$\begin{aligned}
&\mathbb{E}\sum_{s=1}^{t-1}\frac{1}{\sigma_s^h}\log\sqrt{\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}} \\
&\quad\le -\frac{1}{2}\sum_{s=1}^{t-1}\frac{1}{\sigma_s^h}H\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big)^2+\frac{1}{2}\sum_{s=1}^{t-1}\frac{c_s^h\mathrm{TV}\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big)}{\sigma_s^h}.
\end{aligned}$$

Then, we need to bound the Variance:

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{\sigma_s^h}\log\sqrt{\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}}\right] &\le \frac{1}{(\sigma_s^h)^2}\mathbb{E}\left[\log^2\sqrt{\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}}\right] \\
&\le \underbrace{\frac{1}{(\sigma_s^h)^2}\mathbb{E}_{x^{h+1}\sim|P_s^h(\cdot|z_s^h)-P_*^h(\cdot|z_s^h)|}\left[\log^2\sqrt{\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}}\right]}_{Q_1} \\
&\quad+\underbrace{\frac{1}{(\sigma_s^h)^2}\mathbb{E}_{x^{h+1}\sim P_*^h(\cdot|z_s^h)}\left[\log^2\sqrt{\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}}\right]}_{Q_2}. \qquad (16)
\end{aligned}$$

For the term Q_2, by invoking Lemma E.3 and using Assumption 3.1, we obtain

$$\begin{aligned} Q_2 &= \frac{1}{4(\sigma_s^h)^2}\mathbb{E}_{x^{h+1}\sim P_*^h(\cdot|z_s^h)}\left[\log^2\left(\frac{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}\right)\right] \\ &\le \frac{1+\log B}{2(\sigma_s^h)^2}\mathrm{KL}\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big) \\ &\le \frac{(1+\log B)(3+\log B)}{2(\sigma_s^h)^2}H\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big)^2. \end{aligned} \tag{17}$$

For the term Q_1,

$$\begin{aligned} Q_1 &= \frac{1}{4(\sigma_s^h)^2}\int\left|\mathrm{d}P_s^h(x^{h+1}|z_s^h)-\mathrm{d}P_*^h(x^{h+1}|z_s^h)\right|\cdot\log^2\left(\frac{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}\right) \\ &= \frac{1}{4(\sigma_s^h)^2}\sup_{x^{h+1}\in\mathcal{X}}\left|\frac{\mathrm{d}P_s^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}-1\right|\int\mathrm{d}P_*^h(x^{h+1}|z_s^h)\log^2\left(\frac{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}\right) \\ &\le \frac{(\log B+1)c_s^h}{2(\sigma_s^h)^2}\cdot\mathrm{KL}\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big) \\ &\le \frac{(\log B+1)(\log B+3)c_s^h}{2(\sigma_s^h)^2}\cdot H\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big)^2 \\ &\le \frac{(\log B+1)(\log B+3)c_s^h}{(\sigma_s^h)^2}\cdot\mathrm{TV}\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big), \end{aligned} \tag{18}$$

where the first inequality invokes Lemma E.3, the second inequality uses Lemma E.4, and the last inequality holds due to $H(P\|Q)^2\le 2\mathrm{TV}(P\|Q)$. Therefore, from (16), (17) and (18), the Variance is bounded as

$$\begin{aligned} \mathrm{Var}\left[\frac{1}{\sigma_s^h}\log\sqrt{\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}}\right] &\le \frac{(\log B+1)(\log B+3)}{2(\sigma_s^h)^2}H\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big)^2 \\ &\quad+\frac{(\log B+1)(\log B+3)c_s^h}{(\sigma_s^h)^2}\cdot\mathrm{TV}\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big). \end{aligned}$$

By applying Lemma E.2 with $\lambda_0<3/\log B$ and $b=\log B$, we get with probability at least $1-\delta$, for any $M\in\mathcal{M}$

$$\begin{aligned} &\sum_{s=1}^{t-1}\frac{1}{\sigma_s^h}\log\sqrt{\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}} \\ &\quad\le\frac{\log(|\mathcal{M}|/\delta)}{\lambda_0}-\frac{1}{2}\sum_{s=1}^{t-1}\frac{1}{\sigma_s^h}H\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big)^2+\frac{1}{2}\sum_{s=1}^{t-1}\frac{c_s^h\mathrm{TV}\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big)}{\sigma_s^h} \\ &\qquad+\frac{\lambda_0(\log B+1)(\log B+3)}{2(1-\lambda_0\log B/3)}\sum_{s=1}^{t-1}\left(\frac{1}{2(\sigma_s^h)^2}H\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big)^2+\frac{c_s^h}{(\sigma_s^h)^2}\mathrm{TV}\big(P_*^h(\cdot|z_s^h)\big\|P_M^h(\cdot|z_s^h)\big)\right). \end{aligned}$$

By taking $\lambda_0=3/(19\log^2 B)$ and $M=\bar{M}_t$, we further get

$$\begin{aligned} &\sum_{s=1}^{t-1}\frac{1}{\sigma_s^h}\log\sqrt{\frac{\mathrm{d}P_{\bar{M}_t}^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}} \\ &\quad\le\frac{19\log(|\mathcal{M}|/\delta)\log^2 B}{3}-\frac{1}{4}\sum_{s=1}^{t-1}\frac{1}{\sigma_s^h}H\big(P_*^h(\cdot|z_s^h)\big\|P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2 \\ &\qquad+\sum_{s=1}^{t-1}\frac{c_s^h\big(\mathrm{TV}\big(P_*^h(\cdot|z_s^h)\big\|P_{\bar{M}_s}^h(\cdot|z_s^h)\big)+\mathrm{TV}\big(P_{\bar{M}_s}^h(\cdot|z_s^h)\big\|P_{\bar{M}_t}^h(\cdot|z_s^h)\big)\big)}{\sigma_s^h} \\ &\quad\le\frac{19\log(|\mathcal{M}|/\delta)\log^2 B}{3}-\frac{1}{4}\sum_{s=1}^{t-1}\frac{1}{\sigma_s^h}\mathrm{TV}\big(P_*^h(\cdot|z_s^h)\big\|P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2+2\alpha C\max_s\sqrt{\lambda+E_{s-1}}, \end{aligned} \tag{19}$$

where the last inequality uses the induction that $M_* \in \mathcal{M}_s$ and $\bar{M}_t \in \mathcal{M}_{t-1} \subseteq \mathcal{M}_s$, and the definition of the weight:

$$\begin{aligned}\sigma_s^h \geq & \frac{1}{\alpha} \cdot \sup_{M \in \mathcal{M}_s} \frac{\mathrm{TV}\big(P_M^h(\cdot|z_s^h) \| P_{\bar{M}_s}^h(\cdot|z_s^h)\big)}{\sqrt{\lambda + \sum_{\tau=1}^{s-1} \mathrm{TV}\big(P_M^h(\cdot|z_\tau^h) \| P_{\bar{M}_s}^h(\cdot|z_\tau^h)\big)^2/\sigma_\tau^h}} \\ \geq & \frac{1}{\alpha} \cdot \frac{\mathrm{TV}\big(P_{M_*}^h(\cdot|z_s^h) \| P_{\bar{M}_s}^h(\cdot|z_s^h)\big)}{\sqrt{\lambda + \sum_{\tau=1}^{s-1} \mathrm{TV}\big(P_{M_*}^h(\cdot|z_\tau^h) \| P_{\bar{M}_s}^h(\cdot|z_\tau^h)\big)^2/\sigma_\tau^h}}.\end{aligned}$$

Since $\bar{M}_t$ is the maximizer of the log-likelihood,

$$\sum_{s=1}^{t-1} \frac{1}{\sigma_s^h} \mathrm{TV}\big(P_*^h(\cdot|z_s^h) \big\| P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2 \leq \frac{76 \log(|\mathcal{M}|/\delta) \log^2 B}{3} + 8\alpha C \max_s \sqrt{\lambda + E_{s-1}}.$$

Additionally, we use the non-negativity of TV distance to get from (19) that

$$\sum_{s=1}^{t-1} \frac{1}{\sigma_s^h} \log P_*^h(x_s^{h+1}|z_s^h) \geq \sum_{s=1}^{t} \frac{1}{\sigma_s^h} \log P_{\bar{M}_t}^h(x_s^{h+1}|z_s^h) - \frac{38 \log(|\mathcal{M}|/\delta) \log^2 B}{3} - 4\alpha C \max_s \sqrt{\lambda + E_{s-1}},$$

which implies $M_* \in \mathcal{M}_t$.

Moreover, for any $M \in \mathcal{M}_t$ satisfying that

$$\sum_{s=1}^{t-1} \frac{1}{\sigma_s^h} \log P_M^h(x_s^{h+1}|z_s^h) \geq \sum_{s=1}^{t} \frac{1}{\sigma_s^h} \log P_{\bar{M}_t}^h(x_s^{h+1}|z_s^h) - \beta^2, \tag{20}$$

by taking this back into (19) with a general M, we have

$$\begin{aligned}&\frac{\beta^2}{2} - \frac{1}{4} \sum_{s=1}^{t-1} \frac{1}{\sigma_s^h} \mathrm{TV}\big(P_*^h(\cdot|z_s^h) \big\| P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2 \\ &\geq \sum_{s=1}^{t-1} \frac{1}{\sigma_s^h} \log \sqrt{\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}} \\ &= \sum_{s=1}^{t-1} \frac{1}{\sigma_s^h} \log \sqrt{\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_{\bar{M}_t}^h(x^{h+1}|z_s^h)}} + \sum_{s=1}^{t-1} \frac{1}{\sigma_s^h} \log \sqrt{\frac{\mathrm{d}P_{\bar{M}_t}^h(x^{h+1}|z_s^h)}{\mathrm{d}P_*^h(x^{h+1}|z_s^h)}} \\ &\geq \sum_{s=1}^{t-1} \frac{1}{\sigma_s^h} \log \sqrt{\frac{\mathrm{d}P_M^h(x^{h+1}|z_s^h)}{\mathrm{d}P_{\bar{M}_t}^h(x^{h+1}|z_s^h)}} \geq -\frac{\beta^2}{2},\end{aligned}$$

which indicates that

$$\sum_{s=1}^{t-1} \frac{1}{\sigma_s^h} \mathrm{TV}\big(P_*^h(\cdot|z_s^h) \big\| P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2 \leq 4\beta^2.$$

Therefore, we complete the proof. □

Proof of Lemma A.1. Start at step $h = 1$:

$$\begin{aligned}&V_M^1(x^1) - V_{\pi_M}^1(x^1) \\ &\quad = \mathbb{E}_{a^1 \sim \pi_M}\big[\mathbb{E}^M V_M^2(x^2) - \mathbb{E}^* V_{\pi_M}^2(x^2)\big] \\ &\quad = \mathbb{E}_{a^1 \sim \pi_M}\big[(\mathbb{E}^M V_M^2(x^2) - \mathbb{E}^* V_M^2(x^2)) + (\mathbb{E}^* V_M^2(x^2) - \mathbb{E}^t V_M^2(x^2)) + (\mathbb{E}^t V_M^2(x^2) - \mathbb{E}^t V_{\pi_M}^2(x^2)) \\ &\qquad + (\mathbb{E}^t V_M^2(x^2) V_{\pi_M}^2(x^2) - \mathbb{E}^* V_{\pi_M}^2(x^2))\big] \\ &\quad \leq \mathbb{E}_{a^1 \sim \pi_M}\big[\mathbb{E}^M V_M^2(x^2) - \mathbb{E}^* V_M^2(x^2) + 2\mathrm{TV}(P_t^h(\cdot|x^1, a^1) \| P_*^h(\cdot|x^1, a^1)) + \mathbb{E}^t[V_M^2(x^2) - V_{\pi_M}^2(x^2)]\big] \\ &\quad \leq \mathbb{E}_{a^1 \sim \pi_M}\big[\mathcal{E}^1(M, x^1, a^1) + c_t^h(x^1, a^1) + \mathbb{E}^t[V_M^2(x^2) - V_{\pi_M}^2(x^2)]\big],\end{aligned}$$

where the first inequality holds from the fact that $V_M(\cdot), V_{\pi_M}(\cdot) \in [0,1]$, and the second inequality is due to the definition of bellman error and $\mathrm{TV}(P\|Q) \le 1/2 \cdot \sup_x |P(x)/Q(x) - 1|$. By further expanding the last term on the right-hand side of the inequality above, we complete the first inequality of the Lemma. Similarly, we can obtain the second inequality. □

Proof of Lemma A.2. To condense notations, we use the notation $l(M, \bar{M}_t, z) = \mathrm{TV}(P_M(\cdot|z)\|P_{\bar{M}_t}(\cdot|z))$. Now, we follow the three steps in the proof of Lemma 5.1 from Ye et al. (2023a).

Step I: Matched levels The first step is to divide the sample set $\mathcal{S}_T$ into $\log_2(T/\lambda)+1$ disjoint subsets

$$\mathcal{S}_T = \cup_{\iota=0}^{\log_2(T/\lambda)} \mathcal{S}^\iota.$$

For each $z_t \in \mathcal{S}_T$, let $M_{z_t} \in \mathcal{M}_t$ be the maximizer of

$$\frac{l(M_{z_t}, \bar{M}_t, z_t)/\sigma_t^{1/2}}{\sqrt{\lambda + \sum_{s=1}^{t-1} l(M_{z_t}, \bar{M}_t, z_s)^2/\sigma_s}}.$$

Since $l(M_{z_t}, \bar{M}_t, z_t) \in [0,1]$, we can decompose $\mathcal{S}_T$ into $\log_2(T/\lambda)$ disjoint subsequences:

$$\mathcal{S}^\iota = \{z_t \in \mathcal{S}_T \,|\, l(M_{z_t}, \bar{M}_t, z_t)^2 \in (2^{-\iota-1}, 2^{-\iota}]\},$$

and

$$\mathcal{S}^{\log_2(T/\lambda)} = \{z_t \in \mathcal{S}_T \,|\, l(M_{z_t}, \bar{M}_t, z_t)^2 \in [0, \lambda/T]\}.$$

Correspondingly, we also divide $\mathbb{R}^+$ into $\log_2(T/\lambda)+1$ disjoint subsets:

$$\mathbb{R}^+ = \cup_{\iota=-1}^{\log_2(T/\lambda)} \mathcal{R}^\iota,$$

where we define

$$\mathcal{R}^\iota = [2^{\iota/2}\log|\mathcal{M}|, 2^{(\iota+1)/2}\log|\mathcal{M}|), \text{ for } \iota = 0, \dots, \log_2(T/\lambda)-1,$$
$$\mathcal{R}^{\log_2(T/\lambda)} = [\sqrt{T/\lambda}\log|\mathcal{M}|, +\infty),\ \mathcal{R}^{\log_2(T/\lambda)} = [0, \log|\mathcal{M}|).$$

Then, there exists an $\iota_0 \in \{-1, 0, \dots, \log_2(T/\lambda)\}$ such that $C \in \mathcal{R}^{\iota_0}$.

Step II: Control weights in each level For any $z_t \in \mathcal{S}^{\log_2(T/\lambda)}$, we have

$$\begin{aligned}\big(I_\sigma(\lambda, \mathcal{M}, \mathcal{S}_t)\big)^2 \le& \sup_{M\in\mathcal{M}_t} \frac{l(M, \bar{M}_t, z_t)^2/\sigma_t^2}{\lambda + \sum_{s=1}^{t-1} l(M_{z_t}, \bar{M}_t, z_s)^2/\sigma_s} \\ \le& \frac{l(M, \bar{M}_t, z_t)^2}{\lambda} \le \frac{1}{T},\end{aligned} \tag{21}$$

which implies that

$$\sum_{z_t \in \mathcal{S}^{\log_2(T/\lambda)}} \big(I_\sigma(\lambda, \mathcal{M}, \mathcal{S}_t)\big)^2 \le 1.$$

Moreover, for $\iota = 0, \dots, \log_2(T/\lambda)-1$, we aim to control the upper and lower bound of weights $\{\sigma_t : z_t \in \mathcal{S}^\iota\}$. We define for $t \in [T]$,

$$\begin{aligned}\psi_t =& \frac{1}{\alpha} \sup_{M\in\mathcal{M}_t} \frac{l(M, \bar{M}_t, z_t)}{\sqrt{\lambda + \sum_{s=1}^{t-1} l(M, \bar{M}_t, z_s)^2/\sigma_s}} \\ =& \frac{1}{\alpha} \frac{l(M_{z_t}, \bar{M}_t, z_t)}{\sqrt{\lambda + \sum_{s=1}^{t-1} l(M_{z_t}, \bar{M}_t, z_s)^2/\sigma_s}}.\end{aligned}$$

When $\iota > \iota_0$, we have

$$C < 2^{(\iota_0+1)/2}\log|\mathcal{M}| < 2^{\iota/2}\log|\mathcal{M}|.$$

For any $z_t \in \mathcal{S}^\iota$, we know that $l(M_{z_t}, \bar{M}_t, z_t)^2 < 2^{-\iota}$. Hence, it follows that

$$\psi_t \le \frac{1}{\alpha} \cdot \frac{2^{-\iota/2}}{\sqrt{\lambda}} \le \frac{C}{\sqrt{\log|\mathcal{M}|}} \cdot \frac{2^{-\iota/2}}{\sqrt{\log|\mathcal{M}|}} \le 1$$

Since $\sigma_t = \max\{1, \psi_t\}$, we get $\sigma_t = 1$ for all $z_t \in \mathcal{S}^\iota$.

When $\iota \le \iota_0$, for all $z_t \in \mathcal{S}^\iota$ we get $C \ge 2^{\iota_0/2} \log|\mathcal{M}| \ge 2^{\iota/2} \log|\mathcal{M}|$, and $l(M, \bar{M}_t, z_t)^2 \in (2^{-\iota-1}, 2^{-\iota})$. Then, we can verify that

$$\psi_t \le \frac{C}{\sqrt{\log|\mathcal{M}|}} \cdot \frac{2^{-\iota/2}}{\sqrt{\log|\mathcal{M}|}} = \frac{C}{2^{\iota/2}\log|\mathcal{M}|},$$
$$\psi_t \ge \frac{C}{\sqrt{\log|\mathcal{M}|}} \cdot \frac{2^{-(\iota+1)/2}}{\sqrt{c_0 \log|\mathcal{M}|}} = \frac{C}{\sqrt{2c_0}2^{\iota/2}\log|\mathcal{M}|},$$

where the inequality of the second row applies

$$\lambda + \sum_{s=1}^{t-1} l(M, \bar{M}_t, z_t) \le \lambda + \beta^2 \le c_0 \log|\mathcal{M}|.$$

Since $C/(2^{\iota/2}\log|\mathcal{M}|) \ge 1$, we further have for all $z_t \in \mathcal{S}^\iota$

$$\sigma_t^2 \in \Big[\frac{C}{\sqrt{2c_0}2^{\iota/2}\log|\mathcal{M}|}, \frac{C}{2^{\iota/2}\log|\mathcal{M}|}\Big].$$

Step III: Bound the sum In this step, we bound the sum $\sum_{z_t \in \mathcal{S}^\iota} (I_\sigma(\lambda, \mathcal{M}, \mathcal{S}_t))^2$ for each $\iota = 0, \ldots, \log_2(T/\lambda) - 1$. Fixing an ι, we can decompose $\mathcal{S}^\iota$ into $N^\iota + 1$ disjoint subsets:

$$\mathcal{S}^\iota = \cup_{j=1}^{N^\iota+1} \mathcal{S}_j^\iota,$$

where we define $N^\iota = |\mathcal{S}^\iota|/\mathrm{ED}(\mathcal{M}, 2^{(-\iota-1)/2})$. With a slight abuse of notation, we have $\mathcal{S}^\iota = \{z_i\}_{i=1}^{|\mathcal{S}^\iota|}$, where the elements are arranged in the same order as in the original set $\mathcal{S}_T$. Initially, let $\mathcal{S}_j^\iota = \{\}$ for all $j \in [N^\iota + 1]$. From $i = 1$ to $|\mathcal{S}^\iota|$, we find the smallest $j \in [N^\iota]$ such that z_i is $2^{(-\iota-1)/2}$-independent of $\mathcal{S}_j^\iota$ with respect to $\mathcal{M}$. If such a j does not exist, set $j = N^\iota + 1$. Then, let the choice of j for each z_i be $j(z_i)$. According to the design of the procedure, it is obvious that for all $z_i \in \mathcal{S}^\iota$, z_i is $2^{(-\iota-1)/2}$-dependent on each of $\mathcal{S}_{1,i}^\iota, \ldots, \mathcal{S}_{j(z_i)-1,i}^\iota$, where $\mathcal{S}_{k,i}^\iota = \mathcal{S}_t^\iota \cap \{z_1, \ldots, z_{i-1}\}$ for $k = 1, \ldots, j(z_i) - 1$.

For any $z_i \in \mathcal{S}^\iota$ indexed by t in $\mathcal{S}_T$, we have $l(M_{z_t}, \bar{M}_t, z_t)^2 \ge 2^{-\iota-1}$. Then, because z_i is $2^{(-\iota-1)/2}$-dependent on $\mathcal{S}_{1,i}^\iota, \ldots, \mathcal{S}_{j(z_i)-1,i}^\iota$, respectively, we get for each $k = 1, \ldots, j(z_i) - 1$,

$$\sum_{z \in \mathcal{S}_{k,i}^\iota} l(M_{z_t}, \bar{M}_t, z)^2 \ge 2^{-\iota-1}.$$

Then, we obtain

$$\frac{l(M_{z_t}, \bar{M}_t, z_t)^2/\sigma_t}{\lambda + \sum_{s=1}^{t-1} l(M_{z_t}, \bar{M}_t, z_s)^2/\sigma_s} \le \frac{2^{-\iota}/\sigma_t}{\lambda + \sum_{k=1}^{j(z_i)-1} \sum_{z_s \in \mathcal{S}_{k,i}^\iota} l(M_{z_t}, \bar{M}_t, z)^2/\sigma_s}.$$

When $\iota > \iota_0$, recall from step II that $\sigma_t = 1$ for all $z_t \in \mathcal{S}^\iota$. Thus, we get

$$\frac{l(M_{z_t}, \bar{M}_t, z_t)^2/\sigma_t}{\lambda + \sum_{s=1}^{t-1} l(M_{z_t}, \bar{M}_t, z_s)^2/\sigma_s} \le \frac{2^{-\iota}}{\lambda + (j(z_i) - 1)2^{-\iota-1}} = \frac{2}{j(z_i) - 1 + \lambda 2^{\iota+1}}.$$

By summing over all $z_t \in \mathcal{S}^\iota$, we obtain

$$\begin{aligned}\sum_{z_t\in\mathcal{S}^\iota}(I_\sigma(\lambda,\mathcal{M},\mathcal{S}_t))^2 &\le \sum_{j=1}^{N^\iota}\sum_{z_i\in\mathcal{S}_j^\iota}\frac{2}{j-1+\lambda 2^{\iota+1}}+\sum_{z_i\in\mathcal{S}_{N^\iota+1}^\iota}\frac{2}{N^\iota}\\ &\le \sum_{j=1}^{N^\iota}\frac{2|\mathcal{S}_j^\iota|}{j}+\frac{2|\mathcal{S}_{N^\iota+1}^\iota|}{N^\iota}\\ &\le 2\mathrm{ED}(\mathcal{M},2^{(-\iota-1)/2})\log N^\iota+2|\mathcal{S}^\iota|\cdot\frac{\mathrm{ED}(\mathcal{M},2^{(-\iota-1)/2})}{|\mathcal{S}^\iota|}\\ &\le 4\mathrm{ED}(\mathcal{M},2^{(-\iota-1)/2})\log N^\iota, \end{aligned} \tag{22}$$

where the third inequality is deduced since by the definition of eluder dimension, we have $|\mathcal{S}_j^\iota| \le \mathrm{ED}(\mathcal{M}, 2^{(-\iota-1)/2})$ for all $j \in [N^\iota]$.

When $\iota \le \iota_0$, we have from step II that $\sigma_t^2 \in [C/(\sqrt{2c_0}2^{\iota/2}\log N), C/(2^{\iota/2}\log N)]$ for all $z_t \in \mathcal{S}^\iota$, which indicates that their weights are roughly of the same order. Then, we obtain that

$$\begin{aligned}\frac{l(M_{z_t},\bar{M}_t,z_t)^2/\sigma_t}{\lambda+\sum_{s=1}^{t-1}l(M_{z_t},\bar{M}_t,z_s)^2/\sigma_s} &\le \frac{l(M_{z_t},\bar{M}_t,z_t)^2/\sigma_t}{\lambda+\sum_{s\in[t-1],z_s\in\mathcal{S}^\iota}l(M_{z_t},\bar{M}_t,z_s)/\sigma_s}\\ &\le \frac{2^{-\iota}\sqrt{2c_0}2^{\iota/2}\log N/C}{\lambda+(j(z_i)-1)2^{-\iota-1}\cdot 2^{\iota/2}\log N/C}\\ &\le \frac{\sqrt{8c_0}}{j(z_i)-1+\lambda 2^{\iota/2+1}C/\log N}\\ &\le \frac{\sqrt{8c_0}}{j(z_i)-1+\lambda 2^{\iota+1}},\end{aligned}$$

where the last inequality uses $C \ge 2^{\iota/2}\log N$. By summing over all $z_t \in \mathcal{S}^\iota$, we have

$$\begin{aligned}\sum_{z_t\in\mathcal{S}^\iota}(I_\sigma(\lambda,\mathcal{M},\mathcal{S}_t))^2 &\le \sum_{j=1}^{N^\iota}\sum_{z_i\in\mathcal{S}_j^\iota}\frac{\sqrt{8c_0}}{j-1+\lambda 2^{\iota+1}}+\sum_{z_i\in\mathcal{S}_{N^\iota+1}^\iota}\frac{2}{N^\iota}\\ &\le (\sqrt{8c_0}+2)\mathrm{ED}(\mathcal{M},2^{(-\iota-1)/2})\log N^\iota.\end{aligned} \tag{23}$$

Finally, by combining (21), (22) and (23), we have

$$\begin{aligned}&\sum_{t=1}^{T}(I_\sigma(\lambda,\mathcal{M},\mathcal{S}_t))^2\\ &= \sum_{\iota=0}^{\log_2(T/\lambda)}\sum_{z_t\in\mathcal{S}^\iota}(I_\sigma(\lambda,\mathcal{M},\mathcal{S}_t))^2\\ &\le \sum_{\iota=0}^{\iota_0}(\sqrt{8c_0}+2)\mathrm{ED}(\mathcal{M},2^{(-\iota-1)/2})\log N^\iota+\sum_{\iota=\iota_0+1}^{\log_2(T/\lambda)-1}4\mathrm{ED}(\mathcal{M},2^{(-\iota-1)/2})\log N^\iota+1\\ &\le (\sqrt{8c_0}+3)\mathrm{ED}(\mathcal{M},\sqrt{\lambda/T})\log_2(T/\lambda)\log T,\end{aligned}$$

where the last inequality uses the monotonicity of the eluder dimension. Note that if $\iota_0 = -1$, let the sum from 0 to -1 be 0. Eventually, we accomplish the proof due to the arbitrariness of Z_1^T. □

C.2. Lemmas for Offline Setting

Proof of Lemma B.2. For simplicity, we assume that class $\mathcal{M}$ has finite elements. This proof is the same with the proof of Lemma 4.6 except for the formulation of the weights. For conciseness, we only present the difference here. Let

$$E = \sum_{t=1}^{T} \mathrm{TV}(P_*^h(\cdot|z_t^h) \| P_{\bar{M}}^h(\cdot|z_t^h))^2 / \sigma_t^h.$$

Similar to (19), we can deduce that

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{1}{\sigma_t^h} \log \sqrt{\frac{\mathrm{d}P_{\bar{M}}^h(x^{h+1}|z_t^h)}{\mathrm{d}P_*^h(x^{h+1}|z_t^h)}} \\
&\quad \le \frac{19 \log(|\mathcal{M}|/\delta) \log^2 B}{3} - \frac{1}{4} \sum_{t=1}^{T} \frac{1}{\sigma_t^h} H\big(P_*^h(\cdot|z_t^h) \big\| P_{\bar{M}}^h(\cdot|z_t^h)\big)^2 + \sum_{t=1}^{T} \frac{c_t^h \mathrm{TV}\big(P_*^h(\cdot|z_t^h) \big\| P_{\bar{M}}^h(\cdot|z_t^h)\big)}{\sigma_t^h} \\
&\quad \le \frac{19 \log(|\mathcal{M}|/\delta) \log^2 B}{3} - \frac{1}{4} \sum_{s=1}^{t-1} \frac{1}{\sigma_s^h} \mathrm{TV}\big(P_*^h(\cdot|z_s^h) \big\| P_{\bar{M}_t}^h(\cdot|z_s^h)\big)^2 + 2\alpha C\sqrt{\lambda + E},
\end{aligned}$$

where the last inequality uses Lemma B.1:

$$\begin{aligned}
\sigma_t^h \ge& \frac{1}{2\alpha} \cdot \sup_{M, M' \in \mathcal{M}} \frac{\mathrm{TV}(P_M^h(\cdot|z_t) \| P_{M'}^h(\cdot|z_t))/\alpha}{\sqrt{\lambda + \sum_{s=1}^{T} \mathrm{TV}(P_M^h(\cdot|z_s) \| P_{M'}^h(\cdot|z_s))^2 / \sigma_s^2}} \\
\ge& \frac{1}{2\alpha} \cdot \frac{\mathrm{TV}(P_*^h(\cdot|z_t) \| P_{\bar{M}}^h(\cdot|z_t))/\alpha}{\sqrt{\lambda + \sum_{s=1}^{T} \mathrm{TV}(P_*^h(\cdot|z_s) \| P_{\bar{M}}^h(\cdot|z_s))^2 / \sigma_s^2}}.
\end{aligned}$$

Since $\bar{M}$ is the maximizer of the log-likelihood, we have

$$E \le \frac{74 \log(|\mathcal{M}|/\delta) \log^2 B}{3} + 8\alpha C\sqrt{\lambda + E},$$

which implies that

$$E \le 2\beta^2.$$

On the other hand, we get

$$\sum_{t=1}^{T} \frac{1}{\sigma_t^h} \log P_*^h(x_t^{h+1}|z_t^h) \ge \sum_{t=1}^{T} \frac{1}{\sigma_t^h} \log P_{\bar{M}}^h(x_t^{h+1}|z_t^h) - \frac{38 \log(|\mathcal{M}|/\delta) \log^2 B}{3} - 4\alpha C\sqrt{\lambda + E},$$

which implies $M_* \in \widehat{\mathcal{M}}$. □

The proof adapts the analysis of Lemma 4.1 in Ye et al. (2023b).

Proof of Lemma B.3. For convenience, we use the short-hand notation

$$l^h(M, M', z) = \mathrm{TV}(P_M^h(\cdot|z) \| P_{M'}^h(\cdot|z)).$$

Recall from Definition 5.1 that

$$\mathrm{IC}^{\sigma}(\lambda, \widehat{\mathcal{M}}, \mathcal{D}) = T \cdot \max_{h \in [H]} \mathbb{E}_{\pi_*} \left[\sup_{M, M' \in \widehat{\mathcal{M}}} \frac{l^h(M, M', z^h)^2 / \sigma^h(z^h)}{\lambda + \sum_{t=1}^{T} l^h(M, M', z_t^h)^2 / \sigma_t^h} \middle| x^1 = x \right],$$

where we define

$$\sigma^h(z^h) = \max \left\{ 1, \sup_{M, M' \in \widehat{\mathcal{M}}} \frac{l^h(M, M', z^h)/\alpha}{\sqrt{\lambda + \sum_{t=1}^{T} l^h(M, M', z_t^h)^2 / \sigma_t^h}} \right\}.$$

Besides, Assumption 5.3 states that for any $h \in [H]$, and two distinct $M, M' \in \mathcal{M}$,

$$\frac{1}{T}\sum_{t=1}^{T} l^h(M, M', z_t^h)^2 \geq \mathrm{Cov}(\mathcal{M}, \mathcal{D})\rho^h(M, M')^2,$$

where $\rho^h(M, M') = \sup_z \mathrm{TV}(P_M^h(\cdot|z) \| P_{M'}^h(\cdot|z))$.

Let M_{z^h}, M'_{z^h} be the models that maximize

$$\frac{l^h(M, M', z^h)^2/\sigma^h(z^h)}{\lambda + \sum_{t=1}^{T} l^h(M, M', z_t^h)^2/\sigma_t^h}.$$

We use the notation

$$\psi(z^h) = \frac{l^h(M_{z^h}, M'_{z^h}, z^h)^2/\sigma^h(z^h)}{\lambda + \sum_{t=1}^{T} l^h(M_{z^h}, M'_{z^h}, z^h)^2/\sigma_t^h}.$$

Since

$$\begin{aligned}
\sigma^h(z^h) \geq & \sup_{M, M' \in \widehat{\mathcal{M}}} \frac{l^h(M, M', z^h)/\alpha}{\sqrt{\lambda + \sum_{t=1}^{T} l^h(M, M', z_t^h)^2/\sigma_t^h}} \\
= & \frac{l^h(M_{z^h}, M'_{z^h}, z^h)/\alpha}{\sqrt{\lambda + \sum_{t=1}^{T} l^h(M_{z^h}, M'_{z^h}, z_t^h)^2/\sigma_t^h}},
\end{aligned}$$

we deduce that

$$\begin{aligned}
(\sigma^h(z^h))^{1/2} \geq & \frac{1}{\alpha} \cdot \frac{l^h(M_{z^h}, M'_{z^h}, z^h)/(\sigma^h(z^h))^{1/2}}{\sqrt{\lambda + \sum_{t=1}^{T} l^h(M_{z^h}, M'_{z^h}, z_t^h)^2/\sigma_t^h}} \\
= & \frac{\sqrt{\psi(z^h)}}{\alpha}.
\end{aligned} \tag{24}$$

Then, we will derive a uniform upper bound for σ_t^h for $t \in [T]$. For all $t \in [T]$, we have from Lemma B.1 that

$$\begin{aligned}
\sigma_t^h \leq & \max\left\{1, \sup_{M, M' \in \mathcal{M}} \frac{l^h(M, M', z_t^h)/\alpha}{\sqrt{\lambda + \sum_{s=1}^{T} l^h(M, M', z_t^h)^2/\sigma_s^h}}\right\} \\
\leq & \max\left\{1, \sup_{M, M' \in \mathcal{M}} \frac{l^h(M, M', z_t^h)/\alpha}{\sqrt{\lambda + \sum_{s=1}^{T} l^h(M, M', z_t^h)^2/\max_s \sigma_s^h}}\right\} \\
\leq & \max_s \sqrt{\sigma_s^h} \cdot \max\left\{1, \sup_{M, M' \in \mathcal{M}} \frac{l^h(M, M', z_t^h)/\alpha}{\sqrt{\lambda + \sum_{s=1}^{T} l^h(M, M', z_t^h)^2}}\right\} \\
\leq & \max_s \sqrt{\sigma_s^h} \cdot \max\left\{1, \sup_{M, M' \in \mathcal{M}} \frac{l^h(M, M', z_t^h)/\alpha}{\sqrt{\lambda + \sum_{s=1}^{T} l^h(M, M', z_t^h)^2}}\right\}.
\end{aligned}$$

By using Assumption 5.3, we further have

$$\begin{aligned}
\sigma_t^h \leq & \max_s \sqrt{\sigma_s^h} \cdot \max\left\{1, \sup_{M, M' \in \mathcal{M}} \frac{l^h(M, M', z_t^h)/\alpha}{\sqrt{\lambda + T\mathrm{Cov}(\mathcal{M}, \mathcal{D})\rho^h(M, M')^2}}\right\} \\
\leq & \max_s \sqrt{\sigma_s^h} \cdot \max\left\{1, \frac{1}{\alpha\sqrt{T\mathrm{Cov}(\mathcal{M}, \mathcal{D})}}\right\},
\end{aligned}$$

which implies that

$$\max_t \sigma_t^h \le \max\Big\{1, \frac{1}{\alpha^2 T\mathrm{Cov}(\mathcal{M},\mathcal{D})}\Big\}. \tag{25}$$

There are two situations. If $\alpha^2 T\mathrm{Cov}(\mathcal{M},\mathcal{D}) \ge 1$, we have from (25) that $\sigma_t^h = 1$ for all $t \in [T]$. Hence, since $\sigma^h(z^h) \ge 1$ it follows that

$$\begin{aligned}
\mathrm{IC}^\sigma(\lambda, \widehat{\mathcal{M}}, \mathcal{D}) \le & T \cdot \max_{h\in[H]} \mathbb{E}_{\pi_*}\Big[\sup_{M,M'\in\widehat{\mathcal{M}}} \frac{l^h(M,M',z^h)^2}{\lambda + \sum_{t=1}^T l^h(M,M',z_t^h)^2}\Big] \\
\le & T \cdot \max_{h\in[H]} \mathbb{E}_{\pi_*}\Big[\sup_{M,M'\in\widehat{\mathcal{M}}} \frac{l^h(M,M',z^h)^2}{\lambda + T\mathrm{Cov}(\mathcal{M},\mathcal{D})\rho(M,M')^2}\Big] \\
\le & T \cdot \frac{1}{T\mathrm{Cov}(\mathcal{M},\mathcal{D})} = \frac{1}{\mathrm{Cov}(\mathcal{M},\mathcal{D})}.
\end{aligned}$$

If $\alpha^2 T\mathrm{Cov}(\mathcal{M},\mathcal{D}) \ge 1$, we combine (24) and (25) to get

$$\begin{aligned}
\psi(z^h) \le & \frac{l^h(M_{z^h}, M'_{z^h}, z^h)^2 \cdot \alpha^2/\psi(z^h)}{\lambda + \sum_{t=1}^T l^h(M_{z^h}, M'_{z^h}, z^h)^2 \cdot \alpha^2 T\mathrm{Cov}(\mathcal{M},\mathcal{D})} \\
\le & \frac{1}{\psi(z^h)} \cdot \frac{l^h(M_{z^h}, M'_{z^h}, z^h)^2}{T\mathrm{Cov}(\mathcal{M},\mathcal{D}) \cdot T\mathrm{Cov}(\mathcal{M},\mathcal{D})\rho(M_{z^h}, M'_{z^h})^2} \\
\le & \frac{1}{\psi(z^h)} \cdot \frac{1}{(T\mathrm{Cov}(\mathcal{M},\mathcal{D}))^2},
\end{aligned}$$

which implies that

$$\psi(z^h) \le \frac{1}{T\mathrm{Cov}(\mathcal{M},\mathcal{D})}.$$

Therefore, we obtain

$$\mathrm{IC}^\sigma(\lambda, \widehat{\mathcal{M}}, \mathcal{D}) \le \frac{1}{\mathrm{Cov}(\mathcal{M},\mathcal{D})}.$$

□

D. Auxiliary Results

D.1. Bounding TV-Eluder Dimension for Tabular MDPs

Theorem D.1. *Consider a family of tabular MDPs with transition $P_M^h : \mathcal{X}\times\mathcal{A} \to [0,1]$ for any $M \in \mathcal{M}, h \in [H]$. Then, we have*

$$\mathrm{ED}(\mathcal{M},\epsilon) \le 48SA\log(1 + 8SA/\epsilon^2).$$

Proof. For any $i \in [n], z \in \mathcal{X}\times\mathcal{A}$, denote $\mathrm{TV}(P_M^h(\cdot|z)\|P_{M'}^h(\cdot|z))$ by $l_i(z)$. Let $\{(z_i, l_i)\}_{i=1}^n$ be any sequence such that for any $i \in [n]$, (z_i, l_i) is ϵ-independent of its predecessors. We can formulate z_i and l_i as SA-dimensional vectors, respectively. Let $\Sigma_i = \sum_{s=1}^{i-1} z_i z_i^\top + \lambda I$ for any $i \in [n]$. Then, we get

$$\begin{aligned}
l_i(z_i) = & z_i^\top l_i \le \|z_i\|_{\Sigma_i^{-1}} \|l_i\|_{\Sigma_i} \\
\le & \|z_i\|_{\Sigma_i^{-1}} \sqrt{l_i^\top \Big(\sum_{s=1}^{i-1} z_i z_i^\top + \lambda I\Big) l_i} \\
= & \|z_i\|_{\Sigma_i^{-1}} \sqrt{\sum_{s=1}^{i-1} (l_i^\top z_i)^2 + \lambda SA}.
\end{aligned} \tag{26}$$

On the one hand, we have

$$\sum_{i=1}^{n} l_i(z_i) > n\epsilon.$$

On the other hand, we get

$$\begin{aligned}
\sum_{i=1}^{n} l_i(z_i) \leq & \sum_{i=1}^{n} \|z_i\|_{\Sigma_i^{-1}} \sqrt{\sum_{s=1}^{i-1} (l_i^\top z_i)^2 + \lambda SA} \\
\leq & \sum_{i=1}^{n} \|z_i\|_{\Sigma_i^{-1}} \sqrt{\epsilon^2 + \lambda SA} \\
\leq & (\epsilon + \sqrt{\lambda SA}) \sum_{i=1}^{n} \|z_i\|_{\Sigma_i^{-1}} \\
\leq & 2\epsilon \sqrt{T \sum_{i=1}^{n} \|z_i\|^2_{\Sigma_i^{-1}}} \leq 2\epsilon \sqrt{2nSA \log(1 + n/\epsilon^2)},
\end{aligned}$$

where the first inequality uses (26), the second inequality holds due to the definition of ϵ-independence, and the last inequality applies elliptical potential lemma (Lemma E.5) and takes $\lambda = \epsilon^2/SA$. It follows that

$$n\epsilon \leq 2\epsilon \sqrt{2nSA \log(1 + n/\epsilon^2)},$$

which implies that

$$n \leq 8SA \log(1 + n/\epsilon^2).$$

According to Lemma G.5 of Wang et al. (2023), we obtain that

$$n \leq 48SA \log(1 + 8SA/\epsilon^2).$$

□

D.2. Bounding TV-Eluder Dimension for Linear MDPs

Theorem D.2. *Consider a family of linear MDPs, and there exist maps $\nu^h : \mathcal{M} \times \mathcal{X} \to \mathbb{R}^d$ and $\phi^h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ such that transition $P_M^h(x^{h+1}|z^h) = \nu^h(M, x^{h+1})^\top \phi^h(z^h)$, where $\|\phi(z)\|_2 \leq 1$ for any $z \in \mathcal{X} \times \mathcal{A}$. Then, we have*

$$\mathrm{ED}(\mathcal{M}, \epsilon) \leq 48d \log(1 + 8d/\epsilon^2)$$

Proof. For any $i \in [n], z \in \mathcal{X} \times \mathcal{A}$, denote $\mathrm{TV}(P_{M_i}^h(\cdot|z) \| P_{M_i'}^h(\cdot|z))$ by $l_i(z)$. Let $\{(z_i, l_i)\}_{i=1}^n$ be any sequence such that for any $i \in [n]$, (z_i, l_i) is ϵ-independent of its predecessors. Let $\Sigma_i = \sum_{s=1}^{i-1} \phi^h(z_s)\phi^h(z_s)^\top + \lambda I$. For any $M, M' \in \mathcal{M}$, we have

$$\begin{aligned}
\mathrm{TV}(P_M^h(\cdot|z_i) \| P_{M'}^h(\cdot|z_i)) = & \sup_{\mathcal{X}_0} \left| \int_{\mathcal{X}_0} \left(\nu^h(M, x^{h+1}) - \nu^h(M', x^{h+1})\right)^\top \phi^h(z_i) \mathrm{d}x^{h+1} \right| \\
\leq & \|\phi^h(z_i)\|_{\Sigma_i^{-1}} \sup_{\mathcal{X}_0} \left\| \int_{\mathcal{X}_0} \left(\nu^h(M, x^{h+1}) - \nu^h(M', x^{h+1})\right) \mathrm{d}x^{h+1} \right\|_{\Sigma_i}.
\end{aligned}$$

Since

$$\begin{aligned}
&\sup_{\mathcal{X}_0} \Big\| \int_{\mathcal{X}_0} \big(\nu^h(M, x^{h+1}) - \nu^h(M', x^{h+1})\big) \mathrm{d}x^{h+1} \Big\|_{\Sigma_i}^2 \\
&\quad \le \sup_{\mathcal{X}_0} \sum_{s=1}^{i-1} \Big(\int_{\mathcal{X}_0} \big(\nu^h(M, x^{h+1}) - \nu^h(M', x^{h+1})\big)^\top \phi^h(z_s) \Big)^2 + \lambda d \\
&\quad \le \sum_{s=1}^{i-1} \Big(\sup_{\mathcal{X}_0} \int_{\mathcal{X}_0} \big(\nu^h(M, x^{h+1}) - \nu^h(M', x^{h+1})\big)^\top \phi^h(z_s) \Big)^2 + \lambda d \\
&\quad \le \sum_{s=1}^{i-1} \mathrm{TV}(P_M^h(\cdot|z_s) \| P_{M'}^h(\cdot|z_s))^2 + \lambda d.
\end{aligned}$$

It follows that

$$\mathrm{TV}(P_M^h(\cdot|z_i) \| P_{M'}^h(\cdot|z_i)) \le \|\phi^h(z_i)\|_{\Sigma_i^{-1}} \sqrt{\sum_{s=1}^{i-1} \mathrm{TV}(P_M^h(\cdot|z_s) \| P_{M'}^h(\cdot|z_s))^2 + \lambda d}. \tag{27}$$

On the one hand, we have

$$\sum_{i=1}^{n} l_i(z_i) > n\epsilon.$$

On the other hand, we derive from (27) that

$$\begin{aligned}
\sum_{i=1}^{n} l_i(z_i) \le& \sum_{i=1}^{n} \|\phi^h(z_i)\|_{\Sigma_i^{-1}} \sqrt{\sum_{s=1}^{i-1} l_i(z_s)^2 + \lambda d} \\
\le& \sqrt{\epsilon^2 + \lambda d} \sum_{i=1}^{n} \|\phi^h(z_i)\|_{\Sigma_i^{-1}} \\
\le& (\epsilon + \sqrt{\lambda d}) \sqrt{n \sum_{i=1}^{n} \|\phi^h(z_i)\|_{\Sigma_i^{-1}}^2} \\
\le& 2\epsilon \sqrt{2nd \log(1 + n/\epsilon^2)},
\end{aligned}$$

where the last inequality holds by invoking Lemma E.5 and taking $\lambda = \epsilon^2/d$. Therefore, we obtain

$$n\epsilon \le 2\epsilon \sqrt{2nd \log(1 + n/\epsilon^2)}.$$

Then, by applying Lemma G.5 of Wang et al. (2023), we complete the proof. □

Example D.3 (Information Ratio for Linear Model). If the transition can be embedded as $P_M^h(x^{h+1}|z^h) = \nu^h(M, x^{h+1})^\top \phi^h(z^h)$, the IR defined in (1)

$$I^h(\lambda, \mathcal{M}, \mathcal{S}_t^h) \le \min\Big\{1, \|\phi^h(z_t^h)\|_{(\Sigma_t^h)^{-1}}\Big\}.$$

Proof. Let

$$\Sigma_t^h = \sum_{s=1}^{t-1} \phi^h(z_s^h) \phi^h(z_s^h)^\top.$$

We deduce from the definition of TV-norm and the Cauchy-Schwartz inequality that

$$\begin{aligned}\mathrm{TV}(P_M^h(\cdot|z_t^h)\|P_{\bar{M}_t}^h(\cdot|z_t^h)) =& \sup_{\mathcal{X}'}\left|\phi^h(z_t^h)^\top\int_{\mathcal{X}'}(\nu^h(M,x^{h+1})-\nu^h(\bar{M}_t,x^{h+1}))\mathrm{d}x^{h+1}\right| \\ \leq& \|\phi^h(z_t^h)\|_{(\Sigma_t^h)^{-1}}\cdot\sup_{\mathcal{X}'}\left\|\int_{\mathcal{X}'}(\nu^h(M,x^{h+1})-\nu^h(\bar{M}_t,x^{h+1}))\mathrm{d}x^{h+1}\right\|_{\Sigma_t^h}.\end{aligned}$$

We define

$$\mathcal{X}_t = \operatorname*{argmax}_{\mathcal{X}'\subseteq\mathcal{X}}\left\|\int_{\mathcal{X}'}(\nu^h(M,x^{h+1})-\nu^h(\bar{M}_t,x^{h+1}))\mathrm{d}x^{h+1}\right\|_{\Sigma_t^h}.$$

Additionally, we get

$$\begin{aligned}&\lambda+\sum_{s=1}^{t-1}\mathrm{TV}(P_M^h(\cdot|z_s^h)\|P_{\bar{M}_t}^h(\cdot|z_s^h))^2 \\ &=\lambda+\frac{1}{4}\sum_{s=1}^{t-1}\sup_{\mathcal{X}'}\left(\phi^h(z_s^h)^\top\int_{\mathcal{X}'}(\nu^h(M,x^{h+1})-\nu^h(\bar{M}_t,x^{h+1}))\mathrm{d}x^{h+1}\right)^2 \\ &\geq\lambda+\frac{1}{4}\sum_{s=1}^{t-1}\left(\phi^h(z_s^h)^\top\int_{\mathcal{X}_t}(\nu^h(M,x^{h+1})-\nu^h(\bar{M}_t,x^{h+1}))\mathrm{d}x^{h+1}\right)^2 \\ &=\lambda+\left(\int_{\mathcal{X}_t}(\nu^h(M,x^{h+1})-\nu^h(\bar{M}_t,x^{h+1}))\mathrm{d}x^{h+1}\right)^\top\sum_{s=1}^{t-1}\phi^h(z_s^h)\phi^h(z_s^h)^\top \\ &\quad\left(\int_{\mathcal{X}_t}(\nu^h(M,x^{h+1})-\nu^h(\bar{M}_t,x^{h+1}))\mathrm{d}x^{h+1}\right) \\ &\geq\left\|\int_{\mathcal{X}_t}(\nu^h(M,x^{h+1})-\nu^h(\bar{M}_t,x^{h+1}))\mathrm{d}x^{h+1}\right\|_{\Sigma_t^h}^2.\end{aligned}$$

Hence, we have

$$\begin{aligned}&\frac{\mathrm{TV}(P_M^h(\cdot|z_t^h)\|P_{\bar{M}_t}^h(\cdot|z_t^h))}{\sqrt{\lambda+\sum_{s=1}^{t-1}\mathrm{TV}(P_M^h(\cdot|z_s^h)\|P_{\bar{M}_t}^h(\cdot|z_s^h))^2}} \\ &\quad\leq\frac{\|\phi^h(z_t^h)\|_{(\Sigma_t^h)^{-1}}\cdot\left\|\int_{\mathcal{X}_t}(\nu^h(M,x^{h+1})-\nu^h(\bar{M}_t,x^{h+1}))\mathrm{d}x^{h+1}\right\|_{\Sigma_t^h}}{\left\|\int_{\mathcal{X}_t}(\nu^h(M,x^{h+1})-\nu^h(\bar{M}_t,x^{h+1}))\mathrm{d}x^{h+1}\right\|_{\Sigma_t^h}} \\ &\quad\leq\|\phi^h(z_t^h)\|_{(\Sigma_t^h)^{-1}},\end{aligned}$$

which conludes the proof. □

Now, we use the linear MDP to illustrate the condition in Assumption 5.3. In the following lemma, we demonstrate that the condition holds as long as the learner has excess to a well-explored dataset (28), which is a wildly-adopted assumption in the literature of offline linear MDPs (Duan et al., 2020; Wang et al., 2020; Zhong et al., 2022a).

Lemma D.4. *In the linear setting where the transition model M can be embedded into a d-dimensional vector space: $\mathcal{M}^h=\{\langle\nu^h(M,x^{h+1}),\phi^h(\cdot)\rangle:\mathcal{X}\to\mathbb{R}\}$ and $\|\phi(z)\|\leq 1$, if we assume that the data empirical covariance satisfies the following minimum eigenvalue condition: there exists an absolute constant $\bar{c}>0$ such that*

$$\sigma_{\min}\left(T^{-1}\sum_{t=1}^{T}\phi(z_t^h)\phi(z_t^h)^\top\right)=\frac{\bar{c}}{d}, \tag{28}$$

then, Assumption 5.3: for any two distinct $M,M'\in\mathcal{M}$,

$$\min_{h\in[H]}\frac{1}{T}\sum_{t=1}^{T}\mathrm{TV}(P_M^h(\cdot|z_t^h)\|P_{M'}^h(\cdot|z_t^h))^2\geq\mathrm{Cov}(\mathcal{M},\mathcal{D})\rho(M,M')^2$$

with $\mathrm{Cov}(\mathcal{M},\mathcal{D})=\bar{c}/(2d)$ will holds with probability at least $1-\delta$.

Proof. By using the linear model and the definition of TV-norm, we have for some $M, M' \in \mathcal{M}$,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathrm{TV}(P_M^h(\cdot|z_t^h)\|P_{M'}^h(\cdot|z_t^h))^2 \\
&\quad= \frac{1}{T}\sum_{t=1}^{T}\sup_{\mathcal{X}'}\Big(\phi^h(z_t^h)^\top\int_{\mathcal{X}'}(\nu^h(M,x^{h+1})-\nu^h(M',x^{h+1}))\mathrm{d}x^{h+1}\Big)^2 \\
&\quad\geq \sup_{\mathcal{X}'}\frac{1}{T}\sum_{t=1}^{T}\Big(\phi^h(z_t^h)^\top\int_{\mathcal{X}'}(\nu^h(M,x^{h+1})-\nu^h(M',x^{h+1}))\mathrm{d}x^{h+1}\Big)^2 \\
&\quad= \sup_{\mathcal{X}'}\Big(\int_{\mathcal{X}'}(\nu^h(M,x^{h+1})-\nu^h(M',x^{h+1}))\mathrm{d}x^{h+1}\Big)^\top\Lambda_T^h\Big(\int_{\mathcal{X}'}(\nu^h(M,x^{h+1})-\nu^h(M',x^{h+1}))\mathrm{d}x^{h+1}\Big) \\
&\quad\geq \frac{\bar{c}}{d}\sup_{\mathcal{X}'}\Big\|\int_{\mathcal{X}'}(\nu^h(M,x^{h+1})-\nu^h(M',x^{h+1}))\mathrm{d}x^{h+1}\Big\|^2,
\end{aligned}$$

where we define $\Lambda_T^h = T^{-1}\sum_{s=1}^{t-1}\phi^h(z_s^h)\phi^h(z_s^h)^\top$, and the last inequality uses $\lambda_{\min}(\Lambda_T^h) = \bar{c}/d$. We use the short-hand notation

$$\xi = \int_{\mathcal{X}_0}(\nu^h(M,x^{h+1})-\nu^h(M',x^{h+1}))\mathrm{d}x^{h+1},$$

where $\mathcal{X}_0$ is the maximizer of (29). Then, it follows that

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{TV}(P_M^h(\cdot|z_t^h)\|P_{M'}^h(\cdot|z_t^h))^2 \geq \frac{\bar{c}}{d}\|\xi\|^2. \tag{29}$$

Additionally, we also have

$$\begin{aligned}
\rho(M,M') =& \sup_z \mathrm{TV}(P_M^h(\cdot|z)\|P_{M'}^h(\cdot|z)) \\
\leq& \sup_z\sup_{\mathcal{X}'}\Big(\phi^h(z)^\top\int_{\mathcal{X}'}(\nu^h(M,x^{h+1})-\nu^h(M',x^{h+1}))\mathrm{d}x^{h+1}\Big)^2 \\
\leq& \sup_{\mathcal{X}'}\Big\|\int_{\mathcal{X}'}(\nu^h(M,x^{h+1})-\nu^h(M',x^{h+1}))\mathrm{d}x^{h+1}\Big\|^2 \\
=& \|\xi\|^2.
\end{aligned}$$

Therefore, we complete the proof. □

E. Technical Lemmas

Lemma E.1 (Azuma–Hoeffding inequality, Cesa-Bianchi & Lugosi 2006)**.** *Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_i\}$ satisfying $|x_i| \leq M$ for some constant M, x_i is $\mathcal{G}_{i+1}$-measurable, $\mathbb{E}[x_i|\mathcal{G}_i] = 0$. Then for any $0 < \delta < 1$, with probability at least $1-\delta$, we have*

$$\sum_{i=1}^{n} x_i \leq M\sqrt{2n\log(1/\delta)}.$$

Lemma E.2 (Theorem 13.6 of Zhang (2023))**.** *Consider a sequence of random functions $C_1(\mathcal{S}_1), \ldots, C_t(\mathcal{S}_t)$. Assume that $\xi_i \leq \mathbb{E}_{Z_i^{(y)}}\xi_i + b$ for some constant $b > 0$. Then for any $\lambda \in (0, 3/b)$, with probability at least $1-\delta$:*

$$\sum_{i=1}^{n}\xi_i \leq \sum_{i=1}^{n}\mathbb{E}_{Z_i^{(y)}}\xi_i + \frac{\lambda\sum_{i=1}^{n}\mathrm{Var}_{Z_i^{(y)}}(\xi_i)}{2(1-\lambda b/3)} + \frac{\log(1/\delta)}{\lambda}.$$

Lemma E.3. *Let $\rho = \sup_z \log(p(z)/q(z))$. Then,*

$$\int \mathrm{d}P(z) \log^2 \Big(\frac{\mathrm{d}P(z)}{\mathrm{d}Q(z)}\Big) \le 2(\rho+1)\mathrm{KL}(P\|Q).$$

Proof. It is obvious that $\rho > 0$. Now let $f_{\mathrm{KL}}(t) = t\log t - t + 1$ and $f(t) = t\log^2 t$. Define

$$\kappa = \sup_{0\le t\le \exp(\rho)} \frac{f(t)}{f_{\mathrm{KL}}(t)}.$$

By using some algebra, we know that $f(t)/f_{\mathrm{KL}}(t)$ is an increasing function of $t \in [0, +\infty)$, which implies that when $\rho \in (0, +\infty)$

$$\kappa \le \frac{f(\exp(\rho))}{f_{\mathrm{KL}}(\exp(\rho))} = \frac{\rho^2\exp(\rho)}{\rho\exp(\rho) - \exp(\rho) + 1} \le 2 + 2\rho.$$

Therefore, we have

$$\int \mathrm{d}P(z) \log^2 \Big(\frac{\mathrm{d}P(z)}{\mathrm{d}Q(z)}\Big) = \mathbb{E}_{z\sim Q} f\Big(\frac{p(z)}{q(z)}\Big) \le \kappa \mathbb{E}_{z\sim Q} f_{\mathrm{KL}}\Big(\frac{p(z)}{q(z)}\Big) = \kappa \mathrm{KL}(P\|Q),$$

which indicates the desired bounds. □

Lemma E.4 (Proposition B.11 of Zhang (2023))**.** *Let $\rho = \sup_z \log(p(z)/q(z))$. Then*

$$H(P\|Q)^2 \le \mathrm{KL}(P\|Q) \le (3+\rho)H(P\|Q)^2.$$

Lemma E.5 (Elliptical Potential Lemma (Dani et al., 2008; Rusmevichientong & Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011))**.** *Let $\{x_i\}_{i\in[T]}$ be a sequence of vectors in $\mathbb{R}^d$ with $\|x_i\|_2 \le L < \infty$ for all $t \in [T]$. Let Λ_0 be a positive-definite matrix and $\Lambda_t = \Lambda_0 + \sum_{i=1}^t x_i x_i^\top$. It holds that*

$$\log\Big(\frac{\det(\Lambda_t)}{\Lambda_0}\Big) \le \sum_{i=1}^T \|x_i\|^2_{\Lambda_{i-1}^{-1}}.$$

Further, if $\|x_i\|_2 \le L$ for all $i \in [T]$, then we have

$$\sum_{i=1}^T \min\{1, \|x_i\|^2_{\Lambda_{i-1}^{-1}}\} \le 2\log\Big(\frac{\det(\Lambda_t)}{\Lambda_0}\Big) \le 2d\log\Big(\frac{\mathrm{trace}(\Lambda_0) + nL^2}{d\det(\Lambda_0)^{1/d}}\Big).$$

Finally, if $\lambda_{\min}(\Lambda_0) \ge \max(1, L^2)$,

$$\sum_{i=1}^T \|x_i\|^2_{\Lambda_{i-1}^{-1}} \le 2\log\Big(\frac{\det(\Lambda_t)}{\Lambda_0}\Big).$$