

Overcoming the Curse of Dimensionality in Reinforcement Learning Through Approximate Factorization

Chenbei Lu^{*}, Laixi Shi^{†,1}, Zaiwei Chen[‡], Chenye Wu[§], Adam Wierman^{†,2}

^{*}*Institute for Interdisciplinary Information Sciences, Tsinghua University, lcb20@mails.tsinghua.edu.cn*

[†]*Computing & Mathematical Sciences Department, Caltech, , ¹laixis@caltech.edu, ²adamw@caltech.edu*

[‡]*Edwardson School of Industrial Engineering, Purdue University, chen5252@purdue.edu*

[§]*School of Science and Engineering, CUHK Shenzhen, chenyewu@yeah.net*

Abstract

Reinforcement Learning (RL) algorithms are known to suffer from the curse of dimensionality, which refers to the fact that large-scale problems often lead to exponentially high sample complexity. A common solution is to use deep neural networks for function approximation; however, such approaches typically lack theoretical guarantees. To provably address the curse of dimensionality, we observe that many real-world problems exhibit task-specific model structures that, when properly leveraged, can improve the sample efficiency of RL. Building on this insight, we propose overcoming the curse of dimensionality by approximately factorizing the original Markov decision processes (MDPs) into smaller, independently evolving MDPs. This factorization enables the development of sample-efficient RL algorithms in both model-based and model-free settings, with the latter involving a variant of variance-reduced Q-learning. We provide improved sample complexity guarantees for both proposed algorithms. Notably, by leveraging model structure through the approximate factorization of the MDP, the dependence of sample complexity on the size of the state-action space can be exponentially reduced. Numerically, we demonstrate the practicality of our proposed methods through experiments on both synthetic MDP tasks and a wind farm-equipped storage control problem.

1 Introduction

In recent years, reinforcement learning (RL) (Sutton and Barto, 2018) has become a popular framework for solving sequential decision-making problems in unknown environments, with applications across different domains such as robotics (Kober et al., 2013), transportation (Haydari and Yilmaz, 2020), power systems (Chen et al., 2022), and financial markets (Charpentier et al., 2021). Despite significant progress, *the curse of dimensionality* remains a major bottleneck in RL tasks (Sutton and Barto, 2018). Specifically, the sample complexity grows geometrically with the dimensionality of the state-action space of the environment, posing challenges for large-scale applications. For example, in robotic control, even adding one more degree of freedom to a single robot can significantly increase the complexity of the control problem (Spong et al., 2020).

To overcome the curse of dimensionality in sample complexity, a common approach is incorporating function approximation to approximate either the value function or the policy using a prespecified function class (e.g., neural networks) (Sutton and Barto, 2018). While this approach works in certain applications, these methods heavily rely on the design of the function approximation class, tailored parameter tuning, and other empirical insights. Moreover, they often lack theoretical guarantees. To the best of our knowledge, most existing results are limited to basic settings with linear function approximation (Tsitsiklis and Van Roy, 1996; Bhandari et al., 2018; Srikant and Ying, 2019; Chen et al., 2023). A few exceptions using

non-linear approximation (e.g., neural networks) (Fan et al., 2020; Xu and Gu, 2020) typically require additional assumptions tailored for theoretical analysis, which creates a gap with practical algorithms. As a result, achieving provable sample efficiency for large-scale sequential decision-making problems remains a significant challenge in RL (Sutton and Barto, 2018).

Formally, an RL problem is usually modeled as a Markov decision process (MDP), where the environment is represented by an unknown transition kernel and reward function. It is important to note that, in many real-world applications, the transition probabilities and reward functions exhibit inherent factored structures, which, if properly leveraged, could mitigate the challenges of the curse of dimensionality in sample complexity. As an illustration, consider model-based RL, where the curse of dimensionality translates to the challenge of estimating the transition probabilities that are potentially high-dimensional. To address this challenge, fortunately, some partial dependency structures are prevalent in applications such as sustainability efforts (Yeh et al., 2023), queuing systems (Wei et al., 2024), and networked systems (Qu et al., 2020). Specifically, the transition dependency between the current state-action pair and the next state is not fully connected, with the next state often depending on only a subset of the current state-action variables. For example, in power system control, the dynamics of power demand, while being stochastic, are independent of the power generation decisions. Similarly, in multi-agent robotic control tasks, the actions of one robot often do not affect others that are far away geographically, allowing for more localized and sparse modelings and solutions. By identifying and leveraging these structures in algorithm design, it is possible to significantly reduce sample complexity.

One notable framework that captures this idea is the factored MDP framework (Osband and Van Roy, 2014), where the original MDP is decomposed into a set of smaller and independently evolving MDPs. In this case, the overall sample complexity depends on the sum (rather than the product) of the sizes of the state-action spaces of each individual MDP. As a result, the problem size no longer scales exponentially with the dimension of the problem, successfully overcoming the curse of dimensionality. While the factored MDP framework seems to be a promising approach, it relies on the assumption that the original MDP can be perfectly factorized into a set of small MDPs. Such a structural assumption limits the applicability of the factored MDP approach. In addition, to the best of our knowledge, there are no provable model-free algorithms for the factored MDP setting, which restricts the development of more flexible approaches that do not require explicit estimation of the environment model.

1.1 Contributions

In this paper, we remove the limitations of the factored MDP framework by proposing an approximate factorization scheme that decomposes an MDP into low-dimensional components. This scheme can be viewed as an extended framework of the factored MDP, allowing for imperfectly factorized models. Building on this framework, we develop provable RL algorithms that overcome the curse of dimensionality and achieve improved sample efficiency compared to prior works in both model-based and model-free paradigms. The main contributions are summarized below.

Approximate Factorization of MDP and Efficient Sampling. To tackle the curse of dimensionality, we introduce an approximate factorization scheme that flexibly decomposes any MDP into low-dimensional components. This scheme offers more flexibility for factorization operations to find an optimal trade-off between sample complexity and optimality by identifying and exploiting the dependency structure within the environment model. Using these approximately factorized MDPs, we propose a multi-component factorized synchronous sampling approach to estimate the model, framed as a cost-optimal graph coloring problem. This method allows simultaneous sampling of multiple components from a single sample, unlike previous factored MDP methods where each sample estimates the transition of only one component. This sampling approach is key to reducing sample complexity in both our model-based and model-free algorithms.

Model-Based Algorithm with Approximate Factorization. We develop a novel model-based RL algorithm that leverages the proposed synchronous sampling method and fully exploits the low-dimensional structures identified through approximate factorization. By breaking down the large-scale kernel estimation problem into smaller, manageable components, our algorithm achieves provable problem-dependent sample complexity (cf. Theorem 5.1), outperforming existing minimax-optimal bounds for standard MDPs (Azar et al., 2012). Moreover, our approach generalizes the factored MDP framework and, when applied to this setting, improves the best-known sample complexity (Chen et al., 2020) by up to a factor of the number of components. The key technical challenge is to address the statistical correlation between samples in synchronous and factorized sampling. By controlling these correlations in the variance analysis, we reduce the sample complexity dependence from the sum of component sizes to the size of the largest component.

Model-Free Variance-reduced Q-Learning with Approximate Factorization. We propose a model-free algorithm — *variance-reduced Q-learning with approximate factorization* (VRQL-AF) — that incorporates the same synchronous sampling method. It achieves the same problem-dependent sample complexity guarantees (cf. Theorem 6.1) as our model-based approach (up to logarithmic terms), outperforming existing minimax-optimal algorithms for standard MDPs (Wainwright, 2019b). When applied to the factored MDPs (which are special cases of our setting), to the best of our knowledge, VRQL-AF is the first provable model-free algorithm that achieves near-optimal sample complexity guarantees. The improvement results from our tailored factored empirical Bellman operator for approximately factorized MDPs, integrated with a variance-reduction approach to minimize the variance of the stochastic iterative algorithm. This integration requires a refined statistical analysis to tightly control estimation errors and iteration variance across the multiple factored components.

1.2 Related Work

Our work contributes to a few key literatures within the RL community. We discuss each in turn below.

Finite-Sample Analysis for Model-Based Algorithms. Our proposed algorithm (cf. Algorithm 2) follows the model-based RL approach, where the learning process involves model estimation and planning. Model-based approaches have been extensively studied (Azar et al., 2012; Agarwal et al., 2020; Gheshlaghi Azar et al., 2013; Sidford et al., 2018; Azar et al., 2017; Jin et al., 2020), achieving minimax-optimal sample complexity of $\tilde{O}(|\mathcal{S}||\mathcal{A}|\epsilon^{-2}(1-\gamma)^{-3})$ in the generative model setting (Li et al., 2020). This minimax-optimal bound is established by considering all possible MDPs in a worst-case manner, without leveraging any additional structure in the problem. In contrast, by leveraging the structure for algorithm design, we achieve sample complexity with exponentially reduced dependency on the size of the state and action space (cf. Theorem 5.1) when the MDP is perfectly factorizable.

Finite-Sample Analysis for Model-Free Algorithms. Our proposed algorithm also aligns with model-free RL, which does not estimate the model but directly optimize the policy (Sutton and Barto, 2018). A vast body of literature focuses on Q-learning (Tsitsiklis, 1994; Jaakkola et al., 1993; Szepesvári, 1997; Kearns and Singh, 1998; Even-Dar et al., 2003; Wainwright, 2019a; Chen et al., 2024a; Li et al., 2024) with various sampling settings, demonstrating a minimax sample complexity of $\tilde{O}(|\mathcal{S}||\mathcal{A}|\epsilon^{-2}(1-\gamma)^{-4})$. With further advancements like variance reduction, Q-learning has been shown to achieve a minimax-optimal sample complexity of $\tilde{O}(|\mathcal{S}||\mathcal{A}|\epsilon^{-2}(1-\gamma)^{-3})$ in the generative model setting (Wainwright, 2019b). In contrast, our work leverages the approximate factorization structure of MDPs to further enhance sample efficiency. By designing a factored empirical Bellman operator with variance reduction, we achieve exponentially reduced sample complexity with respect to the state-action space size (cf. Theorem 6.1) with matching minimax dependence on the other parameters.

Factored MDPs. Our model generalizes the framework of factored MDPs (Boutilier et al., 1995, 1999), extending it to account for approximation errors. Most existing work on factored MDPs is set in an episodic framework and primarily analyzes regret performance (Guestrin et al., 2003; Osband and Van Roy, 2014; Xu and Tewari, 2020; Tian et al., 2020; Chen et al., 2020). In particular, the state-of-the-art results translate into a sample complexity of $\tilde{O}(\sum_{k=1}^K |\mathcal{X}_k| \epsilon^{-2} (1-\gamma)^{-3})$, where the complexity scales with the sum of the state-action space sizes $|\mathcal{X}_k|$ across all factored components. Building on this line of work, we propose a factorized synchronous sampling technique that enables simultaneous updates for multiple components using a single sample. By coupling this with refined cross-component variance analysis, we reduce the sample complexity to as low as $\tilde{O}(\max_k |\mathcal{X}_k| \epsilon^{-2} (1-\gamma)^{-3})$ in an instance-dependent manner, which only depends on the maximal component size rather than the sum. This result matches the lower bounds established in prior work (Xu and Tewari, 2020; Chen et al., 2020), up to logarithmic factors. Crucially, our approach does not require the MDP to exhibit perfect factorizability, allowing for broader applicability to general MDPs.

RL with Function Approximation. To make RL problem sample efficient for large-scale problems, a common approach is to employ function approximation (Sutton and Barto, 2018). Intuitively, the key idea is to limit the searching space of an RL problem to a predefined function class, in which each function can be specified with a parameter that is low-dimensional. This approach has achieved significant empirical success (Mnih et al., 2015; Silver et al., 2017). However, RL with function approximation is not theoretically well understood except under strong structural assumptions on the approximating function class, such as the function class being linear (Tsitsiklis and Van Roy, 1996; Bhandari et al., 2018; Srikant and Ying, 2019; Chen et al., 2023, 2024b), the Bellman completeness being satisfied (Fan et al., 2020), or others (Dai et al., 2018; Wang et al., 2020). Also, the function approximation often targets to approximate the Q-values, instead of exploiting the inherent transition kernel and reward function structures. In this work, we take a different approach by leveraging approximate factorization structures instead of implementing function approximation. It is also worth noting that our approach is highly flexible and can be further extended by incorporating function approximation techniques, providing an even broader framework for tackling large-scale RL problems.

2 Model and Background

We consider an infinite-horizon discounted MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where \mathcal{S} is the finite state space, \mathcal{A} is the finite action space, P is the transition kernel, with $P(s' | s, a)$ denoting the transition probability from state s to s' given action a , $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ ¹ is the reward function, and $\gamma \in (0, 1)$ is the discount factor. The model parameters of the underlying MDP (i.e., the transition kernel P and the reward function r) are unknown to the agent.

At each step, the agent takes action based on the current state of the environment and transitions to the next state in a stochastic manner, according to the transition probabilities, and receives a one-stage reward. This process repeats, and the goal of the agent is to find an optimal strategy (also known as policy) for choosing actions to maximize the long-term discounted cumulative reward. More formally, a policy π is a mapping from the state space \mathcal{S} to the set of probability distributions supported on the action space. Given a policy π , its Q-function $Q^\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ is defined as $Q^\pi(s, a) = \mathbb{E}_{\pi, P}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, where $\mathbb{E}_{\pi, P}[\cdot]$ means that the trajectories are generated by following the transition P and choosing actions according to the policy π , i.e., $s_{t+1} \sim P(\cdot | s_t, a_t)$, $a_{t+1} \sim \pi(\cdot | s_{t+1})$ for all $t \geq 0$. With the Q-function defined above, the goal of an agent is to find an optimal policy π^* such that its associated Q-function is maximized uniformly for all state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$. It has been shown that such an optimal policy always exists (Puterman, 2014).

¹Since we work with finite MDPs, assuming bounded reward is without loss of generality.

The heart of solving an MDP is the Bellman optimality equation, which states that the optimal Q-function, denoted by Q^* , is the unique solution to the fixed-point equation $Q = \mathcal{H}(Q)$, where $\mathcal{H} : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \rightarrow \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is the Bellman optimality operator defined as

$$[\mathcal{H}(Q)](s, a) = \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q(s', a') \right], \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Moreover, once Q^* is obtained, an optimal policy can be computed by choosing actions greedily based on Q^* . Therefore, the problem reduces to estimating the optimal Q-function Q^* .

Suppose that the model parameters of the MDP are known. Then, we can efficiently find Q^* through the value iteration method: $Q_{k+1} = \mathcal{H}(Q_k)$, where Q_0 is initialized arbitrarily. Since the operator $\mathcal{H}(\cdot)$ is a contraction mapping with respect to the ℓ_∞ -norm (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 2018; Puterman, 2014), the Banach fixed-point theorem (Banach, 1922) guarantees that the value iteration method converges geometrically to Q^* .

In RL, the model parameters of the underlying MDP are unknown to the agent. Therefore, one cannot directly perform the value iteration method to find an optimal policy. A natural way to overcome this challenge is to first estimate the model parameters through empirical sampling and then perform the value iteration method based on the estimated model. This is called the model-based method. Another approach is to solve the Bellman optimality equation using the stochastic approximation method without explicitly estimating the model, which leads to model-free algorithms such as Q-learning (Watkins and Dayan, 1992).

Throughout this paper, we work on a widely-used sampling mechanism setting — generative model (Kearns et al., 2002; Kakade, 2003). It enables us to query any state-action pair (s, a) to obtain a random sample of the next state s' and the immediate reward $r(s, a)$ according to the potentially stochastic transition kernel and reward function.

3 Approximate Factorization of Markov Decision Processes

To begin with, we introduce the concept of *approximate factorization*, which is the foundation of the contributions in this paper. Suppose that each state $s \in \mathcal{S}$ is an n -dimensional vector, and the state space can be written as $\mathcal{S} = \prod_{i=1}^n \mathcal{S}_i$, where $\mathcal{S}_i \subseteq \mathbb{R}$ for all $i \in [n] := \{1, 2, \dots, n\}$. Similarly, suppose that each action is an m -dimensional vector and $\mathcal{A} = \prod_{j=1}^m \mathcal{A}_j$, where $\mathcal{A}_j \subseteq \mathbb{R}$ for all $j \in [m]$. As n and m increase, the total number of state-action pairs increases geometrically, which is the source of the curse of dimensionality. Throughout this paper, each component $s[i] \in \mathcal{S}_i$ of a state $s \in \mathbb{R}^n$, where $i \in [n]$, is called a *substate*. A *subaction* is defined similarly. To introduce the approximate factorization scheme, we start by introducing the following useful definitions: *scope*, *scope set* and *scope variable*.

Definition 3.1. Consider a factored d -dimensional set $\mathcal{X} = \prod_{i=1}^d \mathcal{X}_i$, where $\mathcal{X}_i \subseteq \mathbb{R}$ for all $i \in [d]$. For any $Z \subseteq [d]$ (which we call a *scope*), the corresponding *scope set* $\mathcal{X}[Z]$ is defined as $\mathcal{X}[Z] = \prod_{i \in Z} \mathcal{X}_i$. Any element $x[Z] \in \mathcal{X}[Z]$ (which is a $|Z|$ -dimensional vector) is called a *scope variable*.

When Z contains only one element, i.e., $Z = \{i\}$ for some $i \in [d]$, we denote $x[i]$ as $x[\{i\}]$.

With the help of Definition 3.1, now we are ready to introduce the approximate factorization scheme of the MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$. Specifically, an approximate factorization is characterized by a tuple $\omega = \{\omega^P, \omega^R\}$, involving operations on both the transition kernel and the reward function. We begin with the transition kernel.

3.1 Approximate Factorization of the Transition Kernel

A factorization scheme ω^P of the transition kernel is characterized by a tuple $\omega^P = \{K_\omega, \{Z_k^S \subseteq [n] \mid k \in [K_\omega]\}, \{Z_k^P \subseteq [n+m] \mid k \in [K_\omega]\}\}$, where

- $K_\omega \in [n]$ represents the number of components into which we partition the transition kernel;
- $\{Z_k^S \mid k \in [K_\omega]\}$ is a set of scopes with respect to the n -dimensional state space $\mathcal{S} = \prod_{i=1}^n \mathcal{S}_i$. Moreover, $\{Z_k^S \mid k \in [K_\omega]\}$ forms a partition of $[n]$, i.e., $\cup_{k=1}^{K_\omega} Z_k^S = [n]$ and $Z_{k_1}^S \cap Z_{k_2}^S = \emptyset$ for any $k_1 \neq k_2$. For each scope $Z_k^S \subseteq [n]$, there is an associated scope set $\mathcal{S}[Z_k^S]$.
- $\{Z_k^P \mid k \in [K_\omega]\}$ is a set of scopes with respect to the joint state-action space $\mathcal{X} := \mathcal{S} \times \mathcal{A} = \prod_{i=1}^n \mathcal{S}_i \prod_{j=1}^m \mathcal{A}_j$. For example, consider an MDP with a 3-dimensional state space and a 2-dimensional action space. If $Z_k^P = \{1, 2, 4\}$, then its associated scope set $\mathcal{X}[Z_k^P] = \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{A}_1$.

For simplicity of notation, we will use $-Z_k^S$ to denote the complement of Z_k^S with respect to $[n]$, i.e., $-Z_k^S := [n] \setminus Z_k^S$. The factorization scheme ω^P is used to characterize the component-wise dependence structure within the transition kernel. To motivate our approximate factorization scheme, we first consider the special case where the transition kernel P is *perfectly factorizable* with respect to ω^P . That is,

$$\forall x = (s, a) \in \mathcal{X}, s' \in \mathcal{S}, k \in [K_\omega]: \quad P(s' \mid x) = \prod_{k=1}^{K_\omega} P_k(s'[Z_k^S] \mid x[Z_k^P]), \quad (1)$$

where $P_k(\cdot \mid x[Z_k^P])$ is a probability distribution for all $x[Z_k^P] \in \mathcal{X}[Z_k^P]$. Eq. (1) suggests that the transitions of $s'[Z_k^S]$ for different k are independent and depend only on the substate-subaction pairs in $\mathcal{X}[Z_k^P]$. An MDP that permits both perfect transition and reward decompositions (introduced later in Section 3.2) is called a factored MDP (Osband and Van Roy, 2014). Intuitively, this means that the original MDP can be broken down into K_ω smaller components with transitions and rewards depending only on subsets of variables, rather than the full state and action space, thereby reducing the complexity of finding an optimal policy for the MDP.

As illustrated in Section 1, while the factored MDP framework is a promising approach to address the curse of dimensionality challenge, requiring the underlying MDP to be perfectly factorizable is a strong assumption. To remove this limitation, in this work, we propose a more general scheme—approximate factorization—that allows the transition kernel to be imperfectly factorizable, meaning Eq. (1) does not hold exactly. Specifically, to develop an approximate factorization scheme (i.e., to get an estimate of the “marginal” transition probabilities $P(s'[Z_k^S] \mid x[Z_k^P])$ in Eq. (1)), we begin by defining the set of feasible marginal transition probabilities in the following. For any $k \in [K_\omega]$, let

$$\begin{aligned} \mathcal{P}_k &= \left\{ P_k \in \mathbb{R}^{|\mathcal{X}[Z_k^P]| \times |\mathcal{S}[Z_k^S]|} \mid \exists c \in \Sigma^{|\mathcal{X}[-Z_k^P]|} \text{ such that } P_k(s'[Z_k^S] \mid x[Z_k^P]) \right. \\ &= \left. \sum_{x'[-Z_k^P]} c[x'[-Z_k^P]] P(s'[Z_k^S] \mid x[Z_k^P], x'[-Z_k^P]), \forall s'[Z_k^S] \in \mathcal{S}[Z_k^S], x[Z_k^P] \in \mathcal{X}[Z_k^P] \right\}, \end{aligned}$$

where we use Σ^d denotes the d -dimensional probability simplex for any $d \geq 0$. Then, we choose $P_k \in \mathcal{P}_k$ arbitrarily for all $k \in [K_\omega]$. As we will see in later sections, the development of our theoretical results does not rely on the specific choice of $P_k \in \mathcal{P}_k$. To understand the definition of \mathcal{P}_k , again consider the perfectly factorizable setting, in which case $P(s'[Z_k^S] \mid x[Z_k^P], x'[-Z_k^P]) = P(s'[Z_k^S] \mid x[Z_k^P], x''[-Z_k^P])$ for any $x'[-Z_k^P], x''[-Z_k^P] \in \mathcal{X}[-Z_k^P]$ because the transition of $s'[Z_k^S]$ does not depend on the substate-subaction pairs in $\mathcal{X}[-Z_k^P]$. As a result, \mathcal{P}_k is a singleton set for all k , which is consistent with Eq. (1). Coming back to the case where the underlying MDP is not perfectly factorizable, choosing P_k from \mathcal{P}_k can be viewed as a way of approximating the “marginal” transition probabilities.

Given an approximation scheme ω^P , we define the *approximation error* with respect to the transition kernel as

$$\Delta_\omega^P = \sup_{P_1 \in \mathcal{P}_1, \dots, P_{K_\omega} \in \mathcal{P}_{K_\omega}} \max_{s' \in \mathcal{S}, x \in \mathcal{X}} \left| P(s' \mid x) - \prod_{k=1}^{K_\omega} P_k(s'[Z_k^S] \mid x[Z_k^P]) \right|. \quad (2)$$

To understand Δ_ω^P , observe that for any $s' \in \mathcal{S}$ and $x \in \mathcal{X}$, we have

$$\begin{aligned} & P(s' | x) - \prod_{k=1}^{K_\omega} P_k(s'[Z_k^S] | x[Z_k^P]) \\ &= \underbrace{P(s' | x) - \prod_{k=1}^{K_\omega} P_k(s'[Z_k^S] | x)}_{\Delta_1: \text{error from correlated transitions}} + \underbrace{\prod_{k=1}^{K_\omega} P_k(s'[Z_k^S] | x) - \prod_{k=1}^{K_\omega} P_k(s'[Z_k^S] | x[Z_k^P])}_{\Delta_2: \text{error from incomplete dependence}}. \end{aligned}$$

The first term Δ_1 on the right-hand side represents the error due to ignoring the correlation between the transitions of $s'[Z_k^S]$ for different k , and vanishes if the transitions of substates from different scope sets are independent. The second term Δ_2 on the right-hand side arises due to ignoring the dependence of the transitions of $s'[Z_k^S]$ from $x[-Z_k^P]$. Note that when the MDP is perfectly factorizable, we have $\Delta_\omega^P = 0$ (cf. Eq. (1)).

3.2 Approximate Factorization of the Reward Function

Moving to the reward function, the factorization is characterized by the tuple $\omega^R = \{\ell_\omega, \{Z_i^R \in [n+m] \mid i \in [\ell_\omega]\}\}$, where

- ℓ_ω is a positive integer representing the number of components into which we decompose the reward function.
- $\{Z_i^R \mid i \in [\ell_\omega]\}$ is a set of scopes with respect to the joint state-action space \mathcal{X} .

Under this factorization scheme, the global reward function $r(x)$ is approximated by the sum of “local” reward functions from $\{r_i : \mathcal{X}[Z_i^R] \rightarrow [0, 1]\}_{i \in [\ell_\omega]}$, where each r_i depends only on the scope variable $x[Z_i^R]$. Similarly, given an approximation scheme ω^R for the reward function, we define its approximation error as

$$\Delta_\omega^R = \max_{x \in \mathcal{X}} \left| r(x) - \sum_{i=1}^{\ell_\omega} r_i(x[Z_i^R]) \right|,$$

which is a measurement of the deviation from our approximation to the true reward function. The approximation error Δ_ω^R arises due to the fact that the reward function $r(x)$ may depend on substate-subaction pairs outside the scope set $\mathcal{X}[Z_i^R]$. In the special case where the underlying MDP can be perfectly factorized into several small MDPs and the sum of their reward functions is equal to the original reward function, we have $\Delta_\omega^R = 0$.

3.3 An Illustrative Example

To provide intuition, we present an example of applying the approximate factorization scheme in a real-world application, demonstrating that it offers additional opportunities to achieve a better balance between desired solution accuracy and sample efficiency. In particular, we consider the storage control problem in wind farms. The storage controller (the agent) aims to align the real-time wind power generation output (the state) with the desired prediction values by flexibly charging or discharging the energy storage system.

This problem can be modeled as an MDP, where the state at time t , $s_t = (w_t, p_t, c_t)$, captures the wind power generation w_t , electricity price p_t , and the state of charge (SoC) of the storage c_t . The action a_t represents the storage charging decision at time t . The transition dependence structure can be represented

as a bipartite graph, as illustrated in Figure 1(a). In the graph, nodes on the left-hand side (LHS) represent the state and action at time t , while nodes on the right-hand side (RHS) represent the state at time $t + 1$. A solid blue line between nodes indicates a strong dependence. For instance, the electricity price p_{t+1} strongly depends on the previous price p_t . While a dashed blue line represents weak dependence. For example, the wind power generation w_t can influence the next step’s electricity price p_{t+1} in the market, while its impact is weaker compared to the direct influence of the previous price p_t .

Figure 1(b) illustrates an approximate factorization scheme on this MDP, where the system’s dynamics are divided into three smaller components — the dynamics of the wind power generation, electricity price, and storage levels. Strong transition dependencies are preserved within each component, but weaker dependencies, such as the influence of wind power generation on electricity price, are disregarded. This leads to the following approximation of the transition probability:

$$\widehat{P}(s_{t+1} | s_t, a_t) = P(w_{t+1} | w_t) \cdot P(p_{t+1} | p_t) \cdot P(c_{t+1} | c_t, a_t).$$

This approximation simplifies the model by factoring the transition dynamics into smaller, more manageable components while retaining key dependencies. Note that this problem does not fall under a factored MDP, as the dynamics cannot be perfectly broken down into smaller components. However, our approximate factorization scheme offers a more flexible framework, allowing imperfectly factorizable transition dynamics to be divided into smaller components while maintaining sufficient model accuracy. This expands the factorization step’s action space, enabling a more effective search for optimal trade-offs between solution accuracy and sample efficiency, achieved through tailored to the specific dependency structures of the problems.

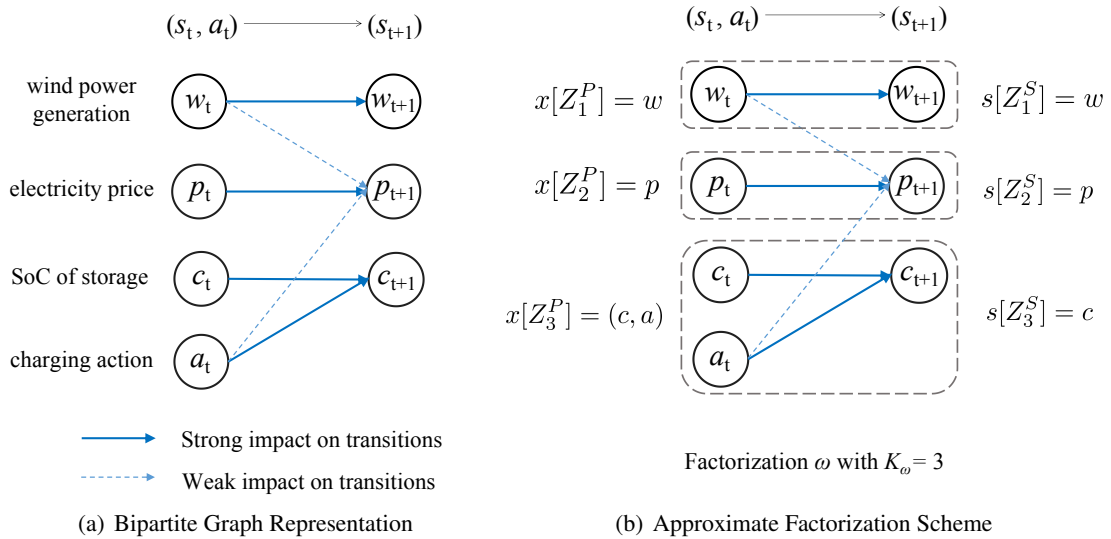


Figure 1: Bipartite Graph Representation and Approximate Factorization.

4 Sampling through Structure: Multi-Component Factorized Synchronous Sampling

Our main contributions propose tailored model-free and model-based approaches to solve approximately factored MDPs sample-efficiently. At the core of both of these approaches is an efficient sampling algorithm

that leverages the factorization structure. In this section, we introduce this core module and then apply it in both model-based and model-free algorithms in the sections that follow.

To illustrate the role of this sampling algorithm, consider classical model-based RL, which consists of two main steps: (1) estimating the transition kernel and the reward function through empirical sampling and (2) performing value iteration (or policy iteration) on the estimated model. The curse of dimensionality in terms of sample efficiency typically arises in the first step. Specifically, to obtain a relatively accurate estimate for the model, starting from each state-action pair, we need to independently sample the next state a sufficient amount of times, leading to a sample complexity proportional to $|\mathcal{S}||\mathcal{A}|$, which scales linearly with the size of the state-action space (hence exponentially with the dimension of the state-action space). By exploiting the approximate factorization structure, we develop a more efficient sampling algorithm. For ease of presentation, we focus on estimating the transition kernel here. The same approach applies to estimating the reward function.

The goal is to construct an approximation \widehat{P} of the original transition kernel P . We achieve this by constructing approximations of the low-dimensional transition probabilities $\{\widehat{P}_k \mid k \in [K_\omega]\}$ and then compute \widehat{P} according to

$$\widehat{P}(s' \mid x) = \prod_{k=1}^{K_\omega} \widehat{P}_k(s'[Z_k^S] \mid x[Z_k^P]) \quad (3)$$

for all $s' \in \mathcal{S}$ and $x = (s, a) \in \mathcal{S} \times \mathcal{A}$. In this case, instead of sampling from the global state-action space, we only need to sample from the substate-subaction spaces, which reduces the sample complexity. To estimate $\widehat{P}_k(s'[Z_k^S] \mid x[Z_k^P])$ for any $k \in [K_\omega]$, let X_k^P be defined as $X_k^P = \{x \in \mathcal{X} \mid x[-Z_k^P] = x^{\text{default}}[-Z_k^P]\}$, where $x^{\text{default}} \in \mathcal{X}$ is an arbitrary (but fixed) element from \mathcal{X} . Sampling from X_k^P allows us to estimate the transition kernel for the k -th factorized component. Importantly, while sampling from X_k , we set $x[-Z_k^P]$ to $x^{\text{default}}[-Z_k^P]$ so that we do not need to cover the rest of the state-action space. For each $x \in X_k^P$, we sample the next state N times, resulting in a set of samples $\{s_{x,i}^k\}_{i \in [N]}$. The low-dimensional empirical transition kernel \widehat{P}_k is then computed as:

$$\widehat{P}_k(s'[Z_k^S] \mid x[Z_k^P]) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(s_{x,i}^k[Z_k^S] = s'[Z_k^S]), \quad \forall s'[Z_k^S] \in \mathcal{S}[Z_k^S], x \in X_k^P. \quad (4)$$

The overall estimate \widehat{P} of the transition kernel can then be computed using Eq. (3). Using this approach, the total number of required samples is $\sum_{k=1}^{K_\omega} |\mathcal{X}[Z_k^P]|N$. Since the sum of the sizes of the substate-subaction space for each component, i.e., $\sum_{k=1}^{K_\omega} |\mathcal{X}[Z_k^P]|$, is usually much smaller (in particular, exponentially smaller) compared to the size of the overall state-action space $|\mathcal{S}||\mathcal{A}|$, the total number of samples required is substantially smaller than what would be needed in classical model-based RL.

4.1 Synchronous Sampling Properties

The sampling method described above can be further optimized to improve the sample efficiency by exploiting the structure of the factorization scheme. Previously, we used the samples $\{s_{x,i}^k\}_{i \in [N]}$ solely to construct the estimator $\widehat{P}_k(s'[Z_k^S] \mid x[Z_k^P])$ for component k . However, these samples can potentially be reused to estimate other low-dimensional transition kernels, depending on the relationships between their associated scopes. To formalize this idea, we introduce two key strategies that leverage the relationships between the scope sets of the components.

Synchronous Sampling with Inclusive Scopes. For any $k_1, k_2 \in [K_\omega]$ such that $Z_{k_1}^P \subseteq Z_{k_2}^P$, the samples used to estimate the transition kernel of component k_2 can be reused to estimate that of component k_1 . Specifically, consider the sampling set $X_{k_2}^P = \{x \in \mathcal{X} \mid x[-Z_{k_2}^P] = x^{\text{default}}[-Z_{k_2}^P]\}$ for component k_2 ,

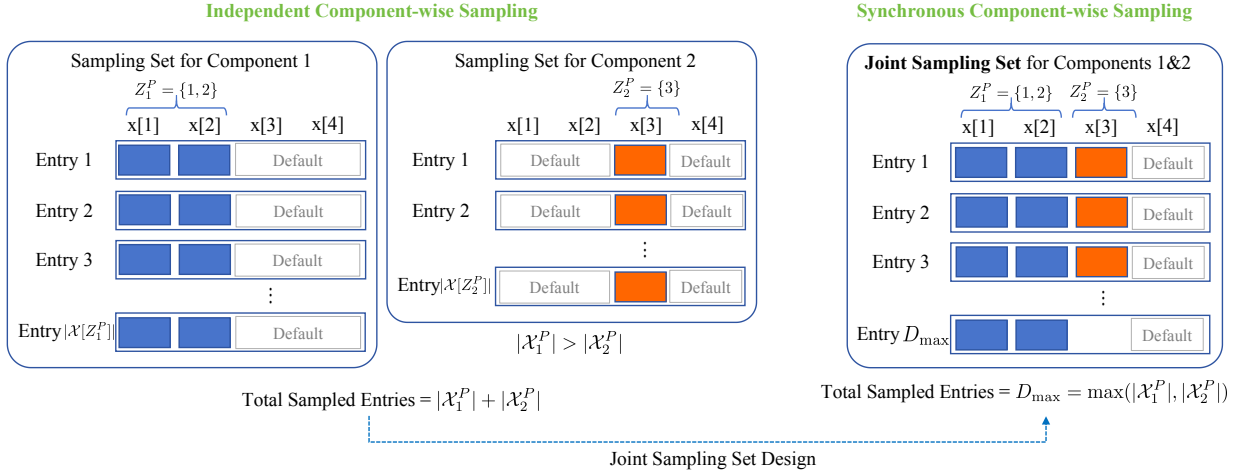


Figure 2: Synchronous Sampling with Exclusive Scopes.

where $x^{\text{default}}[-Z_{k_2}^P]$ is a fixed arbitrary element from $\mathcal{X}[-Z_{k_2}^P]$. By sampling from each $x \in X_{k_2}^P$, we obtain samples of the next state $\{s_{x,i}^{k_2}\}_{i \in [N]}$. Since $Z_{k_1}^P \subseteq Z_{k_2}^P$, these samples inherently contain information about the transitions of component k_1 . Therefore, we can estimate the transition probabilities for component k_1 according to Eq. (4) with $k = k_1$ using the same samples, where the sampling set $X_{k_1}^P$ is defined based on $X_{k_2}^P$ as

$$X_{k_1}^P = \{x \in X_{k_2}^P \mid x[Z_{k_2}^P \setminus Z_{k_1}^P] = x^{\text{default}}[Z_{k_2}^P \setminus Z_{k_1}^P]\}.$$

The reuse of samples here improves the overall sample efficiency by avoiding sampling again for component k_1 .

Synchronous Sampling with Exclusive Scopes. For any $k_1, k_2 \in [K_\omega]$ such that their associated scopes are disjoint, i.e., $Z_{k_1}^P \cap Z_{k_2}^P = \emptyset$, we can estimate the transitions for both components simultaneously using shared samples. Specifically, define the *joint sampling set* X_{k_1, k_2}^P as:

$$X_{k_1, k_2}^P = \{(x[Z_{k_1}^P]_{(i \bmod |\mathcal{X}[Z_{k_1}^P]|+1)}, x[Z_{k_2}^P]_{(i \bmod |\mathcal{X}[Z_{k_2}^P]|+1)}, x^{\text{default}}[-(Z_{k_1}^P \cup Z_{k_2}^P)]) \mid i \in [D_{\max}]\}, \quad (5)$$

where $D_{\max} = \max(|\mathcal{X}[Z_{k_1}^P]|, |\mathcal{X}[Z_{k_2}^P]|)$, $x[Z_{k_1}^P]_{(i)}$ denotes the i -th element in a fixed ordering of $\mathcal{X}[Z_{k_1}^P]$. The modulo operation ensures that we cycle through all possible values of each component's state-action space. By sampling from each $x \in X_{k_1, k_2}^P$ for N times, we obtain samples $\{s_{x,i}\}_{i=1}^N$, which are used to estimate the transition probabilities for both components according to Eq. (4). Figure 2 illustrates an example of synchronous sampling with exclusive scopes. This strategy improves sample efficiency by reducing the total number of samples needed compared to independently sampling each component.

4.2 Cost-Optimal Synchronous Sampling Strategy

Building on the aforementioned two key strategies, we next present our sampling approach. Recall that we have K_ω components associated with scope sets $Z_1^P, Z_2^P, \dots, Z_{K_\omega}^P$. For each $k \in [K_\omega]$, we define its corresponding sampling cost as the size of the scope set $\mathcal{X}[Z_k^P]$. Using the inclusive scope property, we first eliminate components whose scope sets are subsets of others. Let \mathcal{K}_ω^* denote the set of remaining components:

$$\mathcal{K}_\omega^* = \{k \in [K_\omega] \mid \nexists k_1 \neq k \text{ such that } Z_k^P \subsetneq Z_{k_1}^P\}.$$

Next, to use the exclusive scope property, we divide \mathcal{K}_ω^* into subsets $\{\mathcal{G}_i\}_{1 \leq i \leq \kappa_p}$ (where κ_p denotes the total number of subsets) such that within each subset \mathcal{G}_i , the components have disjoint scopes, i.e., $Z_{k_1}^P \cap Z_{k_2}^P = \emptyset$ for all $k_1, k_2 \in \mathcal{G}_i$. This enables us to apply the strategy of synchronous sampling with exclusive scopes within each subset. Specifically, for each subset \mathcal{G}_k of \mathcal{K}_ω^* , we construct the joint sampling sets as follows:

$$X_{\mathcal{G}_k}^P = \left\{ \left\{ x[Z_j^P]_{(i \bmod |\mathcal{X}[Z_j^P]|+1)} \right\}_{j \in \mathcal{G}_k}, x^{\text{default}}[-(\cup_{j \in \mathcal{G}_k} Z_j^P)] \mid i \in [D_{\max, k}] \right\},$$

where $D_{\max, k} = \max_{j \in \mathcal{G}_k} |\mathcal{X}[Z_j^P]|$ ensures that all scope sets associated with this subset are sampled. Therefore, the cost of sampling a joint sampling set is determined by the largest scope sets in the group:

$$c_k(\mathcal{G}_k) = D_{\max, k} = \max_{j \in \mathcal{G}_k} |\mathcal{X}[Z_j^P]|.$$

Our objective is to minimize the total sampling cost across all groups:

$$\min_{\kappa_p} \min_{\mathcal{G}_1, \dots, \mathcal{G}_{\kappa_p}} \sum_{k=1}^{\kappa_p} c_k(\mathcal{G}_k).$$

This problem is closely related to the *Graph Coloring Problem* (GCP) (Jensen and Toft, 2011; Karp, 2010). In Appendix E, we elaborate on the connection between our cost-optimal sampling problem and the GCP, and provide a proof of its computational complexity. Due to such connection, a rich array of well-established algorithms are available to tackle the problem effectively in practice. The solution yields the optimal number of groups $\kappa_p \in [K_\omega]$ and the division scheme $\{\mathcal{G}_k\}_{k \in [\kappa_p]}$. Notably, κ_p corresponds to the *chromatic number* (Erdős and Hajnal, 1966) (the minimal number of colors required) in GCP. Existing literature shows that, under mild conditions, the chromatic number is often much smaller than the number of nodes in the graph (Khot, 2001; Sopena, 1997; Goddard and Xu, 2012). This result translates to $\kappa_p \ll K_\omega$ in our context, leading to substantial sampling cost reductions. A more in-depth discussion on κ_p can refer to Section 5.2.

4.3 MDP Model Estimation based on Synchronous Sampling

Given $\{\mathcal{G}_k\}_{k \in [\kappa_p]}$ and the joint sampling sets $\{X_{\mathcal{G}_k}^P\}_{k \in [\kappa_p]}$, we are ready to sample and estimate the transition kernel. Specifically, for each state $x \in X_{\mathcal{G}_k}^P$, we perform N sampling trials to obtain next-state samples based on the generative model. Specifically, we generate N samples $\{s_{x,i}^k\}_{i=1}^N$ by sampling from $P(\cdot \mid x)$ as detailed in Algorithm 1. For each component $j \in \mathcal{G}_k$, we compute the empirical transition kernel by:

$$\hat{P}_j(s'[Z_j^S] \mid x[Z_j^P]) = \frac{1}{N} \sum_{i \in [N]} 1(s_{x,i}^k[Z_j^S] = s'[Z_j^S]), \quad \forall s'[Z_j^S] \in \mathcal{S}[Z_j^S]. \quad (6)$$

Algorithm 1 Cost-Optimal Factorized Synchronous Sampling Algorithm

- 1: **Input:** Approximate Factorization Scheme $\omega = (\omega^P, \omega^R)$.
 - 2: Solve the optimal division scheme represented by $\{\kappa_p, \{\mathcal{G}_k\}_{k \in [\kappa_p]}\}$;
 - 3: Construct the sampling sets $\{X_{\mathcal{G}_k}^P\}_{k \in [\kappa_p]}$;
 - 4: **for** $k = 1, 2, \dots, \kappa_p$ **do**
 - 5: Sample from each state-action pair $x \in X_{\mathcal{G}_k}^P$ for N times and get the samples by $\{s_{x,i}^k\}_{x \in X_{\mathcal{G}_k}^P, i \in [N]}$.
 - 6: **end for**
 - 7: **Output:** Samples $\{s_{x,i}^k\}_{k \in [\kappa_p], x \in X_{\mathcal{G}_k}^P, i \in [N]}$.
-

To estimate the reward function components, we can employ similar strategies, specifically the inclusive and exclusive scope strategies, to effectively sample the reward function. By constructing a sampling set tailored to these strategies, we can then proceed with learning the reward function. This approach helps reduce the overall sample complexity of learning the reward function components.

5 Model-Based Reinforcement Learning with Approximate Factorization

In this section, focusing on MDPs with approximate factorization and armed with the synchronous sampling approach introduced in the previous section, we propose the *model-based Q-value iteration with approximate factorization* algorithm. This algorithm offers sample complexity guarantees, improving upon prior work.

5.1 Algorithm Design

The complete algorithm is summarized in Algorithm 2, which consists of two primary steps: estimating the model parameters and performing Q-value iteration on the estimated model.

Algorithm 2 Model-Based Q-Value Iteration with Approximate Factorization

- 1: **Input:** Positive integer T and initialization $\widehat{Q}_0(s, a) = 0$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.
 - 2: Compute empirical transition kernel \widehat{P} and the reward function \widehat{r} through synchronous sampling Algorithm 1 and Eqs. (7) and (8).
 - 3: **for** $t = 1, 2, \dots, T$ **do**
 - 4: $\widehat{Q}_t(s, a) = r(s, a) + \gamma \sum_{s'} \left[\widehat{P}(s' | s, a) \max_{a'} \widehat{Q}_{t-1}(s', a') \right]$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.
 - 5: **end for**
 - 6: **Output:** Estimated Q-value function $\widehat{Q}_\omega^* = \widehat{Q}_T$.
-

The first step involves estimating the transition kernel and the reward function (see Algorithm 2, Line 2). Specifically, in Section 4, we explained how to estimate the transition probabilities \widehat{P}_k for each component $k \in [K_\omega]$. For any $s' \in \mathcal{S}$ and $x \in \mathcal{X}$, the overall transition probability $\widehat{P}(s' | x)$ is computed by combining the individual component estimates as follows:

$$\widehat{P}(s' | x) = \prod_{k=1}^{K_\omega} \widehat{P}_k(s' | Z_k^S | x[Z_k^P]). \quad (7)$$

Similarly, for the reward function, we aggregate the estimated rewards \widehat{r}_i for each component $i \in [\ell_\omega]$ to obtain the overall reward function:

$$\widehat{r}(x) = \sum_{i=1}^{\ell_\omega} \widehat{r}_i(x[Z_i^R]). \quad (8)$$

The second step is to apply the value iteration method using these estimated model parameters (see Algorithm 2, Lines 3–5). In particular, Line 4 of Algorithm 2 represents the empirical version of the Bellman iteration. Through value iteration, we can compute the desired optimal-Q-value function and derive the corresponding greedy policy as the final solution to the RL problem.

5.2 Sample Complexity Guarantees

We now present the sample complexity guarantees of Algorithm 2; the proof can be found in Appendix B. Before proceeding, without loss of generality, we assume that $|\mathcal{X}[Z_1^P]| \geq |\mathcal{X}[Z_2^P]| \geq \dots \geq |\mathcal{X}[Z_{K_\omega}^P]|$ and $|\mathcal{X}[Z_1^R]| \geq |\mathcal{X}[Z_2^R]| \geq \dots \geq |\mathcal{X}[Z_{\ell_\omega}^R]|$, i.e., the component scope sets are ordered in descending order based on their cardinality.

Theorem 5.1. *Given any approximate factorization scheme ω , let $\mathcal{E}_\omega = \gamma(1 - \gamma)^{-2} \Delta_\omega^P + (1 - \gamma)^{-1} \Delta_\omega^R$. For any confidence level $\delta > 0$ and the desired accuracy level $\epsilon \in (0, 1)$, with probability at least $1 - \delta$, the output Q-function \widehat{Q}_ω^* from Algorithm 2 after $T \geq \bar{c}_2 \log(\epsilon^{-1}(1 - \gamma)^{-1})$ iterations satisfies*

$$\|\widehat{Q}_\omega^* - Q^*\|_\infty \leq \epsilon + \mathcal{E}_\omega, \quad (9)$$

provided that the total number of samples, denoted by D_ω , satisfies:

$$D_\omega \geq \frac{\bar{c}_0 \left(\sum_{k \in [\kappa_p]} |\mathcal{X}[Z_k^P]| \right) \log(\bar{c}_1 |\mathcal{X}[\cup_{k \in [K_\omega]} Z_k^P]| \delta^{-1})}{\epsilon^2 (1 - \gamma)^3} + \sum_{i \in [\kappa_r]} |\mathcal{X}[Z_i^R]|, \quad (10)$$

where $\kappa_p \in [0, K_\omega]$ and $\kappa_r \in [0, \ell_\omega]$ are problem-dependent parameters, and $\bar{c}_0, \bar{c}_1, \bar{c}_2$ are absolute constants.

Now we are in position to discuss further implications of our results.

Model Misspecification Bias. The parameter \mathcal{E}_ω is called the model misspecification bias, which consists of two terms: $\gamma(1 - \gamma)^{-2} \Delta_\omega^P$ and $(1 - \gamma)^{-1} \Delta_\omega^R$, both of which are linearly dependent on the approximation errors of the transition kernel and reward function, respectively. This bias arises from the inaccuracies introduced by the factorization of the transition kernel and reward function. If the factorization were accurate, these approximation errors would vanish, and \mathcal{E}_ω would reduce to zero, as in the factored MDP case, which is discussed below.

Sample Complexity Implications and Comparison to Prior Works For a fair comparison, we consider our results using a factorization scheme ω that is perfect with the bias $\mathcal{E}_\omega = 0$. In this case, our sample complexity required to achieve an ϵ -optimal policy is of the following order:

$$\tilde{O} \left(\frac{\sum_{k \in [\kappa_p]} |\mathcal{X}[Z_k^P]|}{\epsilon^2 (1 - \gamma)^3} + \sum_{i \in [\kappa_r]} |\mathcal{X}[Z_i^R]| \right). \quad (11)$$

Compared with the minimax-optimal sample complexity (Azar et al., 2012) for solving standard MDPs:

$$\tilde{O} \left(\frac{|\mathcal{S}||\mathcal{A}|}{\epsilon^2 (1 - \gamma)^3} \right), \quad (12)$$

we have identical dependence on δ, ϵ , and the effective horizon $1/(1 - \gamma)$. The key improvement lies in the dependence on the size of the state-action space, where we improve the minimax $\tilde{O}(|\mathcal{S}||\mathcal{A}|)$ dependence (Azar et al., 2012) to $\tilde{O}(\sum_{k \in [\kappa_p]} |\mathcal{X}[Z_k^P]|) \leq \tilde{O}(K_\omega \max_k |\mathcal{X}[Z_k^P]|)$. Notably, $\tilde{O}(K_\omega \max_k |\mathcal{X}[Z_k^P]|)$ is a problem-dependent sample complexity — almost proportional to the sample complexity of solving the largest individual component among the factorized parts of the entire transition kernel, which is exponentially smaller than $|\mathcal{S}||\mathcal{A}|$ due to reduced dimensionality. For example, consider an approximate factorization scheme that decomposes an MDP into 10 disjoint components with identical cardinalities. The corresponding sample complexity becomes $\tilde{O}(|\mathcal{S}||\mathcal{A}|^{\frac{1}{10}})$, achieving an exponential reduction in sample complexity relative to the number of factorized components.

In addition, when the bias $\mathcal{E}_\omega = 0$, the MDPs with approximate factorization reduce to the well-known setting of factored MDPs. Compared to the best-known sample complexity upper bound $\tilde{O} \left(\sum_{k=1}^{K_\omega} |\mathcal{X}[Z_k^P]| \epsilon^{-2} (1 - \gamma)^{-3} + \sum_{i=1}^{\ell_\omega} |\mathcal{X}[Z_i^R]| \right)$ for factored MDPs from (Chen et al., 2020)², our result $\tilde{O} \left(\sum_{k=1}^{\kappa_p} |\mathcal{X}[Z_k^P]| \epsilon^{-2} (1 - \gamma)^{-3} + \sum_{i=1}^{\kappa_r} |\mathcal{X}[Z_i^R]| \right)$ offers strictly better sample complexity, where $\kappa_p \in [K_\omega]$ and $\kappa_r \in [\ell_\omega]$ are instance-dependent parameters. Indeed, κ_p reflects the sparsity of the dependence structure in the MDP. In many real-world applications such as UAV swarm control (Campion et al., 2018) and power system economic dispatch (Chen et al., 2022), the dynamics of different UAVs or different power generators are highly independent, then $\kappa_p \ll K_\omega$ and can be small constant if the dependence

²We translate their results to our setting for easy understanding and clear comparison.

structure is sparse. Also, by connecting it to the *Graph Coloring Problem*, $\kappa_p \ll K_\omega$ is provable under mild conditions (can refer to the discussion in Section 4.2). As a result, it indicates our sample complexity can improve the prior arts in factored MDPs by a factor of up to $\mathcal{O}(K_\omega)$. Similarly, κ_r can be significantly smaller than ℓ_ω if the reward function components have such sparse dependence.

The sample complexity lower bound for factored MDPs is $\tilde{\mathcal{O}}(\max_k |\mathcal{X}[Z_k^P]| \epsilon^{-2} (1 - \gamma)^{-3} + \max_i |\mathcal{X}[Z_i^R]|)$, which is established in (Xu and Tewari, 2020; Chen et al., 2020). It worth noting that, our algorithm is the first to match this lower bound in an instance-dependent manner when $\kappa_p, \kappa_r = \mathcal{O}(1)$, demonstrating that we not only improve upon existing upper bounds but also achieve the theoretical minimum sample complexity under certain conditions.

Trade-Off Between Sample Complexity and Model Misspecification Bias. There exists a trade-off between the sample complexity D_ω and the model misspecification bias \mathcal{E}_ω . The approximate factorization scheme ω can be viewed as a tunable hyperparameter. In general, when the number of components K_ω increases, we get a finer decomposition of P and the size of each component decays exponentially with K_ω , which significantly reduces the sample complexity D_ω . However, the drawback of increasing K_ω arbitrarily is that it may result in a larger misspecification bias \mathcal{E}_ω due to model mismatch. To illustrate this trade-off, consider the example depicted in Figure 1, where we decompose an MDP into three components by disregarding certain weak dependencies. If the induced error \mathcal{E}_ω is smaller than the desired accuracy level ϵ , this imperfect factorization is sufficient to achieve the target solution. More importantly, such approximate factorization improves sample efficiency by reducing the problem’s dimensionality. While when we require higher accuracy (smaller ϵ), a more careful factorization scheme with generally fewer components should be chosen. For instance, grouping w_t and p_t within a single component will reduce the bias \mathcal{E}_ω , though at the cost of increased sample requirements. This trade-off will be revisited again in our numerical simulations in Section 7.1. As a final note, when $K_\omega = 1$, this theorem recovers the classical sample complexity result (without decomposition) established in (Azar et al., 2012).

6 Model-Free Reinforcement Learning with Approximate Factorization

In this section, we focus on designing model-free RL algorithms to solve MDPs with approximate factorization, offering a more flexible paradigm than model-based RL that eliminates the need for explicitly estimating the environment model (i.e., transitions and reward functions). Towards this, we introduce the *Variance-Reduced Q-Learning with Approximate Factorization* (VRQL-AF) algorithm, the first provably near-optimal model-free approach for both our problem and factored MDPs to the best of knowledge.

6.1 Warn-Up: Classical Q-Learning

Our proposed algorithm is a variant of Q-learning (Watkins and Dayan, 1992), one of the most well-known model-free algorithms. With access to a generative model, classical Q-learning (also known as synchronous Q-learning (Watkins and Dayan, 1992; Bertsekas and Tsitsiklis, 1996)) learns by approximating the optimal Q-function through stochastic approximation. Specifically, at each iteration t , given the current estimate \hat{Q}_t , Q-learning samples each state-action pair (s, a) once, obtaining the corresponding reward $r(s, a)$ and a random next state s' according to the unknown transition dynamics. Using these samples, the empirical Bellman operator at time step t is constructed as

$$[\hat{\mathcal{H}}_t(Q)](s, a) = r(s, a) + \gamma \max_{a'} Q(s', a'), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}, Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}.$$

The Q-function estimate \hat{Q}_t is then updated using a small-stepsize version of the empirical Q-value iteration:

$$\hat{Q}_{t+1} = \hat{Q}_t + \eta_t (\hat{\mathcal{H}}_t(\hat{Q}_t) - \hat{Q}_t),$$

where $\eta_t \in (0, 1)$ is the learning rate. Such an algorithm can also be viewed as a stochastic approximation algorithm for solving the Bellman optimality equation (Bertsekas and Tsitsiklis, 1996; Tsitsiklis, 1994).

This algorithm updates the Q-value $\widehat{Q}_t(s, a)$ by incorporating information from the empirical Bellman operator $\widehat{H}_t(\cdot)$, which estimates the expected return for each action based on sampled next states and rewards. Over time, with appropriately chosen learning rates and a sufficient number of iterations, the Q-function converges asymptotically to the optimal Q-function Q^* (Tsitsiklis, 1994).

6.2 Algorithm Design

To address the challenge of low sample efficiency, we propose *Variance-Reduced Q-Learning with Approximate Factorization* (VRQL-AF), which incorporates two key modifications over classical Q-learning: (1) a factored empirical Bellman operator design and (2) a variance-reduced Q-iteration.

The factored Bellman operator is designed to replace the standard empirical Bellman operator in order to enhance sampling efficiency by leveraging the structure. Its construction is outlined in Algorithm 3. Specifically, in vanilla Q-learning, we need to sample all state-action pairs to generate a single empirical Bellman operator, which requires $|\mathcal{S}||\mathcal{A}|$ samples and is highly inefficient. Our approach aims to reduce the number of required samples by exploiting the structure of the problem through a factorization scheme, thus enabling the repeated use of the same samples.

Algorithm 3 Empirical Factored Bellman Operator Generation

- 1: **Input:** Factorization Scheme $\omega = (\omega^P, \omega^R)$.
 - 2: Solve the optimal Synchronous sampling scheme represented by $\{\kappa_p, \{\mathcal{G}_k\}_{k \in [\kappa_p]}\}$;
 - 3: Construct the sampling sets $\{X_{\mathcal{G}_k}^P\}_{k \in [\kappa_p]}$;
 - 4: **for** $k = 1, 2, \dots, \kappa_p$ **do**
 - 5: Sample the transition from each state-action pair $x \in X_{\mathcal{G}_k}^P$ once, and obtain the next state by $\{s_x^k\}_{k \in [\kappa_p]}$.
 - 6: **end for**
 - 7: Get empirical Bellman operator $\widehat{\mathcal{H}}$ (for any Q) following Eq. (13);
 - 8: **Output:** Empirical Bellman operator $\widehat{\mathcal{H}}_i$;
-

Specifically, we adopt the synchronous sampling scheme $\{\kappa_p, \{\mathcal{G}_k\}_{k \in [\kappa_p]}\}$ discussed in Section 4, and then obtain the corresponding κ_p sampling sets $\{X_{\mathcal{G}_k}^P\}_{k \in [\kappa_p]}$. For each $k \in [\kappa_p]$, we sample over each entry x from the set $\mathcal{X}_{\mathcal{G}_k}^P$ once, obtaining the next state s_x^k randomly generated based on transition and corresponding reward. This gives us the samples $\{s_x^k\}_{x \in \mathcal{X}_{\mathcal{G}_k}^P, k \in [\kappa_p]}$. These samples are sufficient to construct the factored empirical Bellman operator $\widehat{\mathcal{H}}(Q)$, which is defined as follows:

$$[\widehat{\mathcal{H}}(Q)](\tilde{s}, \tilde{a}) = r(\tilde{s}, \tilde{a}) + \gamma \max_{a'} Q(s'_{\tilde{x},1}[Z_1^S], s'_{\tilde{x},2}[Z_2^S], \dots, s'_{\tilde{x},K_\omega}[Z_{K_\omega}^S], a') \quad (13)$$

for all $\tilde{x} := (\tilde{s}, \tilde{a}) \in \mathcal{S} \times \mathcal{A}$, where for each $j \in [K_\omega]$, $s'_{\tilde{x},j}$ is the sample satisfying:

$$s'_{\tilde{x},j} \in \{s_x^k \mid x \in \mathcal{X}_{\mathcal{G}_k}^P, x[Z_j^P] = \tilde{x}[Z_j^P]\}.$$

This ensures that the samples from different components are used to compute the transitions for all dimensions of the Bellman operator. By using the synchronous sampling scheme, we cover the transitions for all components, allowing us to estimate the entire factored Bellman operator efficiently.

The variance reduction technique aims to reduce the variance in the update process, thereby decreasing the number of iterations required for convergence (Sidford et al., 2018, 2023; Wainwright, 2019b). Armed

with both schemes, our proposed VRQL-AF is summarized in Algorithm 4. Specifically, the algorithm employs a two-loop structure. In each outer epoch τ , a new reference is first constructed, followed by an inner loop that iteratively updates the Q-value estimate using both the reference and real-time empirical Bellman operators.

Algorithm 4 Variance-Reduced Q-Learning with Approximate Factorization (VRQL-AF)

Input: Number of epochs T ; Epoch length M ; Reference sample size N_τ ($\tau \leq T$); Learning rate η_t .

Output: ϵ -accurate Q Function estimation \bar{Q}_M with $1 - \delta$ probability;

- 1: Initialize $\bar{Q}_0(s, a) = 0$ for all s and a ;
- 2: **for** epoch $\tau = 1, 2, \dots, T$ **do**
- 3: Generate N_τ factored empirical Bellman operators $\{\hat{\mathcal{B}}_i\}_{i=1,2,\dots,N_\tau}$ though Algorithm 3;
- 4: Calculate the reference Bellman operator:

$$\bar{\mathcal{H}}_\tau(\bar{Q}_{\tau-1}) = \frac{1}{N_\tau} \sum_{i=1}^{N_\tau} \hat{\mathcal{B}}_i(\bar{Q}_{\tau-1}).$$

- 5: Initialize $Q_0 = \bar{Q}_{\tau-1}$;
- 6: **for** $t = 0, \dots, M - 1$ **do**
- 7: Generate factored empirical Bellman operator $\hat{\mathcal{H}}_t$ though Algorithm 3;
- 8: Compute the variance-reduced update:

$$Q_{t+1} = Q_t + \eta_t \left(\hat{\mathcal{H}}_t(Q_t) - \hat{\mathcal{H}}_t(\bar{Q}_{\tau-1}) + \bar{\mathcal{H}}_\tau(\bar{Q}_{\tau-1}) - Q_t \right); \quad (14)$$

- 9: **end for**
 - 10: $\bar{Q}_\tau = Q_M$;
 - 11: **end for**
 - 12: **return** \bar{Q}_T ;
-

In particular, in each outer loop τ , we first compute a *reference Bellman operator* $\bar{\mathcal{H}}(\bar{Q}_{\tau-1})$, which is an average of N_τ factored empirical Bellman operators $\hat{\mathcal{B}}_i$ generated through a sampling process applied to the fixed Q-value function $\bar{Q}_{\tau-1}$ (obtained from previous steps). This reference Bellman operator provides a low-variance estimate of the Q-value function.

Next, the inner loop performs variance-reduced Q-iterations. In each iteration t , we generate a new factored empirical Bellman operator $\hat{\mathcal{H}}_{t-1}$ using new samples. In Line 14, the update of the Q-function uses a linear combination of both the high-variance unbiased stochastic factored empirical Bellman operator $\hat{\mathcal{H}}_t(Q_t)$ and a term to reduce the variance of the stochastic process using the constructed reference. This combination of reference and real-time operators allows the algorithm to reduce the number of iterations required to converge to an accurate Q-value function. Meanwhile, there are four parameters to be designed: the number of epochs T in the outer loop, the number of iteration M in the inner loop, the number of samples N_τ to generate the reference Bellman operators in each epoch, and the learning rate η_t at each iteration, which will be specified in the sample complexity guarantee part.

6.3 Sample Complexity Guarantees

We now provide a bound on the sample complexity of Algorithm 4; the proof is provided in Appendix C:

Theorem 6.1. *Given any approximate factorization scheme ω , a confidence level $\delta > 0$, and the desired accuracy level $\epsilon \in (0, 1)$. Let $\mathcal{E}_\omega = \gamma(1 - \gamma)^{-2}\Delta_\omega^P + (1 - \gamma)^{-1}\Delta_\omega^R$, $T = c_1 \log((1 - \gamma)^{-1}\epsilon^{-1})$, $M =$*

$c_2 \log \left(6T |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| (1-\gamma)^{-1} \delta^{-1} \right) (1-\gamma)^{-3}$, $N_\tau = c_3 4^\tau \log \left(6T |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right) (1-\gamma)^{-2}$, and $\eta_t = (1 + (1-\gamma)(t+1))^{-1}$. With probability at least $1-\delta$, the output Q-function \widehat{Q}_ω^* from Algorithm VRQL-AF satisfies

$$\|Q^* - \widehat{Q}_\omega^*\|_\infty \leq \epsilon + \mathcal{E}_\omega,$$

provided that the total number of samples, denoted by D_ω , satisfies the following lower bound:

$$D_\omega \geq \frac{c_0 (\sum_{k \in [\kappa_p]} |\mathcal{X}[Z_k^P]|) \log \left(c_1 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| (1-\gamma)^{-1} \delta^{-1} \right) \log (c_2 (1-\gamma)^{-1} \epsilon^{-1})}{\epsilon^2 (1-\gamma)^3} + \sum_{i \in [\kappa_r]} |\mathcal{X}[Z_i^R]|,$$

where $c_0, c_1, c_2 > 0$ are universal constants.

In evaluating the performance of VRQL-AF, we note that the estimation bias \mathcal{E}_ω is the same as in our model-based algorithm in Theorem 5.1.

Therefore, we focus our discussion on the sample complexity, in the case of perfect factorization $\mathcal{E}_\omega = 0$ for fair comparisons. In this case, our sample complexity required to achieve an ϵ -optimal policy is of the following order:

$$\tilde{O} \left(\frac{\sum_{k \in [\kappa_p]} |\mathcal{X}[Z_k^P]|}{\epsilon^2 (1-\gamma)^3} + \sum_{i \in [\kappa_r]} |\mathcal{X}[Z_i^R]| \right), \quad (15)$$

which achieves the same order as our model-based algorithm in (11). This indicates that our proposed model-free VRQL-AF can also overcome the persistent curse of dimensionality challenges in standard RL, as already discussed for our model-based algorithm. For more details, please refer to Section 5.2.

New Frontiers in Model-Free Paradigm for Factorizable MDPs. VRQL-AF stands out as the first near-optimal model-free algorithm for our approximate factorization framework and also the well-known factored MDPs, which breaks the curse of dimensionality by effectively leveraging problem structure. The technical foundation of these results comes from two aspects: 1) We propose a tailored factored empirical Bellman operator based on the factorization structure and a synchronous sampling scheme. This operator updates the Q-value estimation in a sample-efficient manner by effectively leveraging the scope structures inherent across multiple sub-components of the transition kernel; 2) we integrate advanced variance-reduction techniques with the factored Bellman operator to minimize the oscillation of the stochastic process and accelerate convergence, which requires meticulous design to control the complex cross-dimensional correlations introduced by the factored empirical Bellman operator, ensuring that the variance during the variance-reduced Q-iteration remains manageable.

7 Numerical Experiments

We now present numerical experiments to evaluate our proposed model-based and model-free algorithms based on approximate factorization with comparisons to state-of-the-art RL algorithms. We focus on two types of tasks: (i) synthetic MDP tasks and (ii) an electricity storage control problem in power system operation.

7.1 Synthetic MDP Tasks

The two synthetic MDP tasks that we consider are distinguished by their transition kernels: one with a perfectly factorizable transition kernel and another with an imperfectly factorizable one. For the former task, we consider a three-dimensional state space, where each state is represented as $s = (s[1], s[2], s[3])$, along with a one-dimensional action a . The transitions for $s[1]$ and $s[2]$ are independent of other components, while $s[3]$ is influenced by both its own state and the action a . Each substate and the action space consists of 5 discrete elements. Both the transition kernel and the reward function are randomly generated. We conduct 50 trials to compare our model-based and model-free algorithms (VRQL-AF) with approximate factorization against traditional model-based RL and a variance-reduced Q-learning method without factorization. Figures. 3(a) and 3(b) illustrate the l_∞ -norm of the Q -function error for both settings, respectively, as a function of the number of samples. As shown in the two figures, our approximate factorization methods (depicted in red) consistently exhibit lower Q -function error across varying sample sizes compared to the vanilla RL approaches (depicted in blue). This demonstrates that our approach significantly reduces sample complexity, showcasing superior sample efficiency. These findings underline the effectiveness of our factorization-based framework in enhancing learning efficiency for multi-dimensional MDPs.

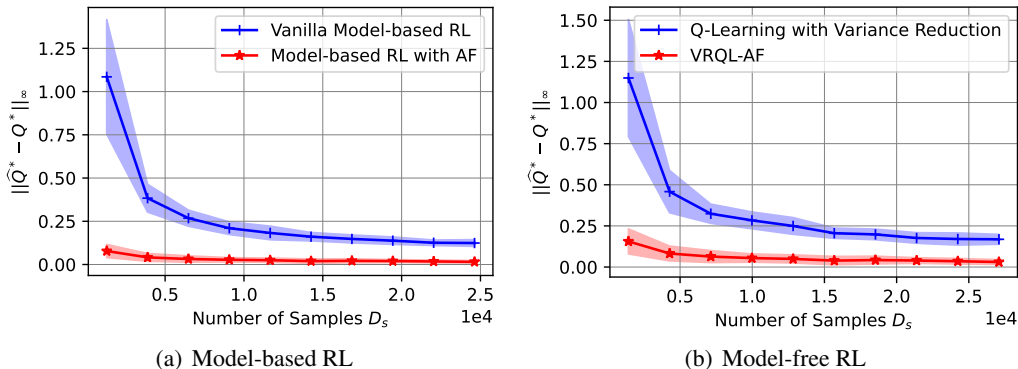


Figure 3: Performance on Perfectly Factorizable MDPs

The second task involves MDPs with imperfectly factorizable transition kernels. Here, the state consists of 4 substates, each taking one of 5 possible values. To investigate the trade-off between sample complexity and estimation error, we evaluate three factorization schemes: (1) full factorization with $K_\omega = 4$, (2) partial factorization with $K_\omega = 2$, and (3) no factorization with ($K_\omega = 1$), corresponding to vanilla RL. We conduct 200 trials to compare their performance with different sample amounts. As shown in Figure 4, our model-based and model-free RL algorithms exhibit distinct convergence behaviors under these different factorization levels. The results indicate two key breaking points (D_1, \mathcal{E}_1) and (D_2, \mathcal{E}_2) in performance trade-offs (illustrated in Figure 4(b)), which inform the optimal factorization choice depending on the sample availability. Specifically, in the small sample range with $D_s \leq D_1$, the full factorization scheme ($K_\omega = 4$) converges quickly but has a higher asymptotic error, making it suitable when the samples are limited or the required accuracy is not high. In the intermediate sample range with $D_1 < D_s \leq D_2$, the partial factorization scheme ($K_\omega = 2$) provides a balanced trade-off, achieving moderate convergence speed and a lower asymptotic error. This scheme is advantageous when a compromise between sample efficiency and accuracy is needed. In the large sample range with $D_s > D_2$, the no factorization scheme ($K_\omega = 1$) results in the smallest asymptotic error but converges slowly. This approach is best suited when very high precision is required, and a large sample size is available. Therefore, it highlights the benefits of selecting an appropriate factorization level to match the sample size and desired accuracy. The flexibility in choosing among different factorization strategies enables the optimization of performance based on specific

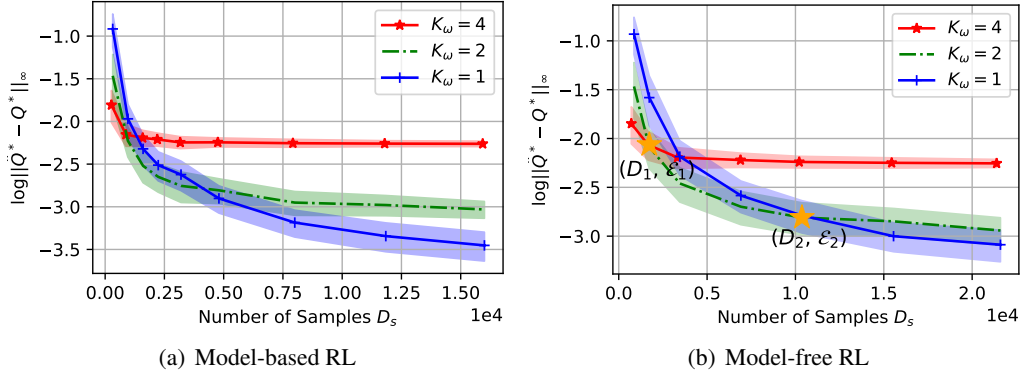


Figure 4: Performance on Imperfectly Factorizable MDPs

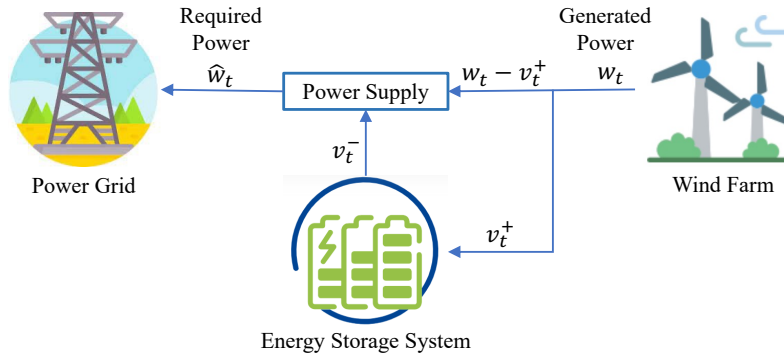


Figure 5: Wind Farm-equipped Storage Control.

requirements, making our approach adaptable to various settings.

7.2 Wind Farm Storage Control Problem

We next evaluate the performance of our proposed approach on the wind farm storage control problem introduced in Section 3.3. As depicted in Figure 5, this problem involves managing energy storage systems to mitigate mismatches between variable wind power generation and demand, thereby minimizing penalty costs.

The control actions consist of charging and discharging decisions influenced by real-time wind generation and unit penalty prices. Constraints include storage capacity limits, prohibiting simultaneous charging and discharging, and lossy energy dynamics. For a detailed model description, please refer to Appendix F. The total state-action space in this problem is 10^4 .

Figure 6 illustrates the performance of our approach, where we decompose the MDP into three components as described earlier. Specifically, Figure 6(a) demonstrates that our approximate factorization-based algorithm converges significantly faster than the vanilla model-based approach by leveraging the model’s structural properties. Additionally, our method exhibits considerably lower variance, indicating more stable performance for different sample conditions. Figure 6(b) presents the economic benefits of our algorithm using a sample size of 60,000 data points (typically taking the wind farm a year to collect), evaluating cumulative system penalty costs over 3,000 time slots. The results show that our approach reduces costs by 19.3% compared to the vanilla model-based method, highlighting its enhanced effectiveness in managing complex problems with large-scale state spaces.

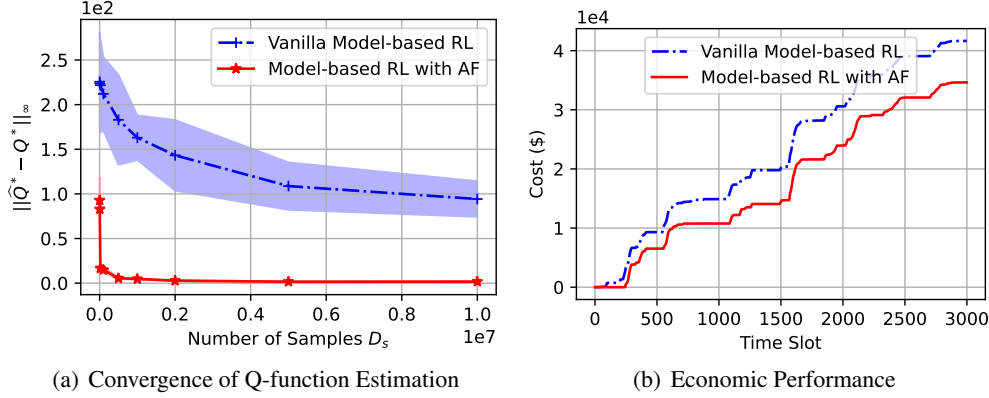


Figure 6: Performance on Storage Control

8 Conclusion

Our work introduces a comprehensive framework for approximate factorization in MDPs, effectively addressing the curse of dimensionality that typically hinders scalability in large-scale RL problems. By decomposing MDPs into smaller and manageable components, our approach significantly enhances learning efficiency. We develop a synchronous sampling method to optimize the sampling process across multiple components and propose a corresponding model-based RL algorithm that fully leverages the factorization structure. Additionally, we design variance-reduced Q-learning algorithms incorporating a factored empirical Bellman operator, which facilitates efficient learning in an online setting.

Our theoretical findings demonstrate significant improvements in sample complexity, surpassing existing bounds for both vanilla and factored MDPs. Numerical simulations on synthetic and real-world high-dimensional problems confirm the practical advantages of our factorization-based approach, indicating its potential to transform scalability in RL. This framework not only provides a versatile tool for large-scale RL but also opens up promising directions for future work, including exploring other forms of factorization with provable guarantees and adapting the approach to varied RL environments.

References

- Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. (2019). Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep.*, 32.
- Agarwal, A., Kakade, S., and Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR.
- Azar, M. G., Munos, R., and Kappen, H. J. (2012). On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1707–1714.
- Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR.
- Banach, S. (1922). Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fund. math*, 3(1):133–181.
- Bertsekas, D. and Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

- Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR.
- Boutilier, C., Dean, T., and Hanks, S. (1999). Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94.
- Boutilier, C., Dearden, R., Goldszmidt, M., et al. (1995). Exploiting structure in policy construction. In *IJCAI*, volume 14, pages 1104–1113.
- California ISO (2021). Electricity Price Data. <https://www.energyonline.com/Data/>. Online; accessed on 20 December 2022.
- Campion, M., Ranganathan, P., and Faruque, S. (2018). Uav swarm communication and control architectures: a review. *Journal of Unmanned Vehicle Systems*, 7(2):93–106.
- Charpentier, A., Elie, R., and Remlinger, C. (2021). Reinforcement learning in economics and finance. *Computational Economics*, pages 1–38.
- Chen, X., Hu, J., Li, L., and Wang, L. (2020). Efficient reinforcement learning in factored MDPs with application to constrained RL. In *International Conference on Learning Representations*.
- Chen, X., Qu, G., Tang, Y., Low, S., and Li, N. (2022). Reinforcement learning for selective key applications in power systems: Recent advances and future challenges. *IEEE Transactions on Smart Grid*, 13(4):2935–2958.
- Chen, Z., Clarke, J.-P., and Maguluri, S. T. (2023). Target network and truncation overcome the deadly triad in Q -learning. *SIAM Journal on Mathematics of Data Science*, 5(4):1078–1101.
- Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2024a). A Lyapunov theory for finite-sample guarantees of Markovian stochastic approximation. *Operations Research*, 72(4):1352–1367.
- Chen, Z., Zhang, K., Mazumdar, E., Ozdaglar, A., and Wierman, A. (2024b). Two-timescale Q -learning with function approximation in zero-sum stochastic games. *The 25th ACM Conference on Economics and Computation*.
- Dai, B., Shaw, A., Li, L., Xiao, L., He, N., Liu, Z., Chen, J., and Song, L. (2018). Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134. PMLR.
- Erdős, P. and Hajnal, A. (1966). On chromatic number of graphs and set-systems. *Acta Math. Acad. Sci. Hungar.*, 17(61-99):1.
- Even-Dar, E., Mansour, Y., and Bartlett, P. (2003). Learning rates for q -learning. *Journal of machine learning Research*, 5(1).
- Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). A theoretical analysis of deep Q -learning. In *Learning for dynamics and control*, pages 486–489. PMLR.
- Gary, M. R. and Johnson, D. S. (1979). Computers and intractability: A guide to the theory of np -completeness.
- Gheshlaghi Azar, M., Munos, R., and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91:325–349.

- Goddard, W. and Xu, H. (2012). The s-packing chromatic number of a graph. *Discussiones Mathematicae Graph Theory*, 32(4):795–806.
- Guestrin, C., Koller, D., Parr, R., and Venkataraman, S. (2003). Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research*, 19:399–468.
- Haydari, A. and Yilmaz, Y. (2020). Deep reinforcement learning for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):11–32.
- Jaakkola, T., Jordan, M., and Singh, S. (1993). Convergence of stochastic iterative dynamic programming algorithms. *Advances in Neural Information Processing Systems*, 6.
- Jensen, T. R. and Toft, B. (2011). *Graph coloring problems*. John Wiley & Sons.
- Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.
- Kakade, S. M. (2003). *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom).
- Karp, R. M. (2010). *Reducibility among combinatorial problems*. Springer.
- Kearns, M., Mansour, Y., and Ng, A. Y. (2002). A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine learning*, 49:193–208.
- Kearns, M. and Singh, S. (1998). Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in Neural Information Processing Systems*, 11.
- Kellerer, H., Pferschy, U., Pisinger, D., Kellerer, H., Pferschy, U., and Pisinger, D. (2004). *Multidimensional knapsack problems*. Springer.
- Khot, S. (2001). Improved inapproximability results for maxclique, chromatic number and approximate graph coloring. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 600–609. IEEE.
- Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274.
- Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2024). Is q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 72(1):222–236.
- Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 33:12861–12872.
- Lin, S. (1965). Computer solutions of the traveling salesman problem. *Bell System Technical Journal*, 44(10):2245–2269.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.
- Osband, I. and Van Roy, B. (2014). Near-optimal reinforcement learning in factored MDPs. *Advances in Neural Information Processing Systems*, 27.

- Puterman, M. L. (2014). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons.
- Qu, G., Wierman, A., and Li, N. (2020). Scalable reinforcement learning of localized policies for multi-agent networked systems. In *Learning for Dynamics and Control*, pages 256–266. PMLR.
- Schaefer, T. J. (1978). The complexity of satisfiability problems. In *Proceedings of the tenth annual ACM symposium on Theory of computing*, pages 216–226.
- Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. (2024). The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 36.
- Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. (2018). Near-optimal time and sample complexities for solving Markov decision processes with a generative model. *Advances in Neural Information Processing Systems*, 31.
- Sidford, A., Wang, M., Wu, X., and Ye, Y. (2023). Variance reduced value iteration and faster algorithms for solving markov decision processes. *Naval Research Logistics (NRL)*, 70(5):423–442.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359.
- Sopena, E. (1997). The chromatic number of oriented graphs. *Journal of Graph Theory*, 25(3):191–205.
- Spong, M. W., Hutchinson, S., and Vidyasagar, M. (2020). *Robot modeling and control*. John Wiley & Sons.
- Srikant, R. and Ying, L. (2019). Finite-time error bounds for linear stochastic approximation and TD-learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR.
- Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.
- Szepesvári, C. (1997). The asymptotic convergence-rate of q-learning. *Advances in Neural Information Processing Systems*, 10.
- Tian, Y., Qian, J., and Sra, S. (2020). Towards minimax optimal reinforcement learning in factored markov decision processes. *Advances in Neural Information Processing Systems*, 33:19896–19907.
- Tsitsiklis, J. and Van Roy, B. (1996). Analysis of temporal-difference learning with function approximation. *Advances in Neural Information Processing Systems*, 9.
- Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and q-learning. *Machine learning*, 16:185–202.
- Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.
- Wainwright, M. J. (2019a). Stochastic approximation with cone-contractive operators: Sharp l-infty bounds for q-learning. *arXiv preprint arXiv:1905.06265*.
- Wainwright, M. J. (2019b). Variance-reduced q-learning is minimax optimal. *Preprint arXiv:1906.04697*.

- Wang, R., Salakhutdinov, R. R., and Yang, L. (2020). Reinforcement learning with general value function approximation: Provably efficient approach via bounded Eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135.
- Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8:279–292.
- Wei, H., Liu, X., Wang, W., and Ying, L. (2024). Sample efficient reinforcement learning in mixed systems through augmented samples and its applications to queueing networks. *Advances in Neural Information Processing Systems*, 36.
- Xu, P. and Gu, Q. (2020). A finite-time analysis of Q-learning with neural network function approximation. In *International Conference on Machine Learning*, pages 10555–10565. PMLR.
- Xu, Z. and Tewari, A. (2020). Reinforcement learning in factored MDPs: Oracle-efficient algorithms and tighter regret bounds for the non-episodic setting. *Advances in Neural Information Processing Systems*, 33:18226–18236.
- Yeh, C., Li, V., Datta, R., Arroyo, J., Christianson, N., Zhang, C., Chen, Y., Hosseini, M. M., Golmohammadi, A., Shi, Y., et al. (2023). Sustaingym: Reinforcement learning environments for sustainable energy systems. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Appendices

A Preliminary Definitions and Lemmas

In the following, we provide all necessary definition, preliminary lemmas. For self-contained, we provide the proofs to the lemmas according to the notations of this work.

Definition A.1. Let $\text{Var}_P(V) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be the value function variance with transition kernel P , which is given by:

$$\text{Var}_P(V) = P(V)^2 - (PV)^2,$$

where $P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$ is the transition kernel, and $(V)^2 \in \mathbb{R}^{|\mathcal{S}|}$ is the element-wise product of V , i.e., $(V)^2 = V \circ V$.

Definition A.2. We define $\Sigma_M^\pi(s, a)$ as the variance of discounted reward under policy π and MDP M given state s and action a , i.e.,

$$\Sigma_M^\pi(s, a) = \mathbb{E} \left[\left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) - Q_M^\pi(s_0, a_0) \right)^2 \middle| s_0 = s, a_0 = a \right],$$

where Q_M^π denotes the Q-function induced by policy π under MDP M .

Lemma A.1. (Lemma 2.2 in (Agarwal et al., 2019)) For any policy π , it holds that:

$$Q^\pi - \widehat{Q}^\pi = \gamma(I - \gamma\widehat{P}^\pi)^{-1}(P - \widehat{P})V^\pi,$$

where Q^π and \widehat{Q}^π are Q-functions induced by the policy π under transition kernels P and \widehat{P} , respectively, and $\widehat{P}^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ represents the transition matrix of the Markov chain $\{(s_t, a_t)\}$ induced by π with transition kernel \widehat{P} .

Proof. Recall that Q^π is the unique solution of the Bellman equation:

$$Q^\pi = r + \gamma P^\pi Q^\pi.$$

Since $I - \gamma P^\pi$ is invertible when $0 < \gamma < 1$, we have

$$Q^\pi = (I - \gamma P^\pi)^{-1} r. \tag{16}$$

It follows that

$$\begin{aligned} Q^\pi - \widehat{Q}^\pi &= (I - \gamma P^\pi)^{-1} r - (I - \gamma \widehat{P}^\pi)^{-1} r \\ &= (I - \gamma \widehat{P}^\pi)^{-1} ((I - \gamma \widehat{P}^\pi) - (I - \gamma P^\pi)) (I - \gamma P^\pi)^{-1} r \\ &= \gamma (I - \gamma \widehat{P}^\pi)^{-1} (P^\pi - \widehat{P}^\pi) (I - \gamma P^\pi)^{-1} r \\ &= \gamma (I - \gamma \widehat{P}^\pi)^{-1} (P^\pi - \widehat{P}^\pi) Q^\pi \\ &= \gamma (I - \gamma \widehat{P}^\pi)^{-1} (P - \widehat{P}) V^\pi, \end{aligned}$$

where the fourth equality is due to Eq. (16). The last equality is because $P^\pi Q^\pi = PV^\pi$. \square

Lemma A.2. (Lemma 2.5 in (Agarwal et al., 2019)) For any two optimal Q functions Q^* and \widehat{Q}^* , which are induced by the same reward function r , but different transition kernels P and \widehat{P} . It holds that:

$$\begin{aligned} Q^* - \widehat{Q}^* &\leq \gamma(I - \gamma\widehat{P}^{\pi^*})^{-1}(P - \widehat{P})V^*, \\ Q^* - \widehat{Q}^* &\geq \gamma(I - \gamma\widehat{P}^{\widehat{\pi}^*})^{-1}(P - \widehat{P})V^*, \end{aligned}$$

where π^* and $\widehat{\pi}^*$ are the optimal policies induced by Q -value functions Q^* and \widehat{Q}^* . The matrix $\widehat{P}^{\pi^*} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ represents the transition matrix of the Markov chain $\{(s_t, a_t)\}$ induced by the policy π^* under the transition kernel \widehat{P} . Similarly, $\widehat{P}^{\widehat{\pi}^*} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ represents the transition matrix of the Markov chain $\{(s_t, a_t)\}$ induced by the policy $\widehat{\pi}^*$ under the transition kernel \widehat{P} . The optimal value function V^* is induced by Q^* , which satisfies $V^*(s) = \max_a Q^*(s, a)$ for all states $s \in \mathcal{S}$.

Proof. The two conditions can be proved using Lemma A.1. Specifically, the first inequality can be proved as follows:

$$Q^* - \widehat{Q}^* = Q^{\pi^*} - \widehat{Q}^{\widehat{\pi}^*} \leq Q^{\pi^*} - \widehat{Q}^{\pi^*} = \gamma(I - \gamma\widehat{P}^{\pi^*})^{-1}(P - \widehat{P})V^*,$$

where the inequality is because $\widehat{Q}^{\widehat{\pi}^*} \geq \widehat{Q}^{\pi^*}$ for any policy π . The last equality comes from Lemma A.1.

For the second condition, we have:

$$\begin{aligned} Q^* - \widehat{Q}^* &= Q^{\pi^*} - \widehat{Q}^{\widehat{\pi}^*} \\ &= Q^{\pi^*} - (I - \gamma\widehat{P}^{\widehat{\pi}^*})^{-1}r \\ &= (I - \gamma\widehat{P}^{\widehat{\pi}^*})^{-1}(I - \gamma\widehat{P}^{\widehat{\pi}^*})Q^{\pi^*} - (I - \gamma\widehat{P}^{\widehat{\pi}^*})^{-1}(I - \gamma P^{\pi^*})Q^{\pi^*} \\ &= (I - \gamma\widehat{P}^{\widehat{\pi}^*})^{-1}((I - \gamma\widehat{P}^{\widehat{\pi}^*}) - (I - \gamma P^{\pi^*}))Q^{\pi^*} \\ &= \gamma(I - \gamma\widehat{P}^{\widehat{\pi}^*})^{-1}(P^{\pi^*} - \widehat{P}^{\widehat{\pi}^*})Q^{\pi^*} \\ &\geq \gamma(I - \gamma\widehat{P}^{\widehat{\pi}^*})^{-1}(P^{\pi^*} - \widehat{P}^{\pi^*})Q^{\pi^*} \\ &= \gamma(I - \gamma\widehat{P}^{\widehat{\pi}^*})^{-1}(P - \widehat{P})V^*. \end{aligned}$$

The second equality is because $Q^{\pi^*} = (I - \gamma P^{\pi^*})^{-1}r$, and the inequality is because $\widehat{P}^{\widehat{\pi}^*} Q^{\pi^*} \leq \widehat{P}^{\pi^*} Q^{\pi^*}$ due to the optimality of policy π^* regarding Q -function Q^{π^*} . \square

Lemma A.3. (Adapted from Lemma 5 in (Azar et al., 2012)) Given a MDP M with transition kernel P and reward function r , the following identity holds for all policy π :

$$\Sigma_M^\pi = \gamma^2(1 - \gamma^2 P^\pi)^{-1} \text{Var}_P(V_M^\pi)$$

Proof. We start with the definition of $\Sigma_M^\pi(s, a)$:

$$\begin{aligned} \Sigma_M^\pi(s, a) &= \mathbb{E}_{P, \pi} \left[\left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) - Q_M^\pi(s_0, a_0) \right)^2 \middle| s_0 = s, a_0 = a \right] \\ &= \mathbb{E}_{P, \pi} \left[\left(\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) - \gamma Q_M^\pi(s_1, a_1) - (Q_M^\pi(s_0, a_0) - r(s_0, a_0) - \gamma Q_M^\pi(s_1, a_1)) \right)^2 \middle| s_0 = s, a_0 = a \right] \\ &= \mathbb{E}_{P, \pi} \left[\left(\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) - \gamma Q_M^\pi(s_1, a_1) \right)^2 \middle| s_0 = s, a_0 = a \right] \end{aligned}$$

$$\begin{aligned}
& - 2\mathbb{E}_{P,\pi} \left[\left(\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) - \gamma Q_M^\pi(s_1, a_1) \right) (Q_M^\pi(s_0, a_0) - r(s_0, a_0) - \gamma Q_M^\pi(s_1, a_1)) \middle| s_0 = s, a_0 = a \right] \\
& + \mathbb{E}_{P,\pi} \left[(Q_M^\pi(s_0, a_0) - r(s_0, a_0) - \gamma Q_M^\pi(s_1, a_1))^2 \middle| s_0 = s, a_0 = a \right] \\
& = \gamma^2 \mathbb{E}_{P,\pi} \left[\left(\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) - Q_M^\pi(s_1, a_1) \right)^2 \middle| s_0 = s, a_0 = a \right] \\
& - 2\mathbb{E}_{P,\pi} \left[\mathbb{E} \left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) - \gamma Q_M^\pi(s_1, a_1) \middle| s_1, a_1 \right] (Q_M^\pi(s_0, a_0) - r(s_0, a_0) - \gamma Q_M^\pi(s_1, a_1)) \middle| s_0 = s, a_0 = a \right] \\
& + \mathbb{E}_{P,\pi} \left[(Q_M^\pi(s_0, a_0) - r(s_0, a_0) - \gamma Q_M^\pi(s_1, a_1))^2 \middle| s_0 = s, a_0 = a \right] \\
& = \gamma^2 \mathbb{E}_{P,\pi} \left[\left(\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) - Q_M^\pi(s_1, a_1) \right)^2 \middle| s_0 = s, a_0 = a \right] \\
& + \mathbb{E}_{P,\pi} \left[(Q_M^\pi(s_0, a_0) - r(s_0, a_0) - \gamma Q_M^\pi(s_1, a_1))^2 \middle| s_0 = s, a_0 = a \right] \\
& = \gamma^2 \mathbb{E}_{P,\pi} \left[\left(\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) - Q_M^\pi(s_1, a_1) \right)^2 \middle| s_0 = s, a_0 = a \right] \\
& + \gamma^2 \mathbb{E}_{P,\pi} \left[(\mathbb{E}_{s_1, a_1 \sim P(\cdot | s_0, a_0)} [Q_M^\pi(s_1, a_1)] - Q_M^\pi(s_1, a_1))^2 \middle| s_0 = s, a_0 = a \right] \\
& = \gamma^2 \sum_{s_1, a_1} P^\pi(s_1, a_1 | s, a) \Sigma_M^\pi(s_1, a_1) + \gamma^2 \text{Var}_P(V_M^\pi)(s, a),
\end{aligned}$$

where the third equality is obtained by dividing the quadratic term; the fourth equality is derived by the law of total expectation; the fifth equality holds due to $\mathbb{E} \left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) - \gamma Q_M^\pi(s_1, a_1) \middle| s_1, a_1 \right] = 0$. The last equality is derived based on the definitions of $\Sigma_M^\pi(s, a)$ and $\text{Var}_P(V_M^\pi)$. \square

Lemma A.4. (Lemma 6 in (Wainwright, 2019b)) Given two Q -functions Q^* and \tilde{Q}^* , which are the induced by the same transition kernel P , but different reward functions r and $\tilde{r} = r + \Delta r$. It holds that,

$$|Q^* - \tilde{Q}^*| \leq \max \left\{ (I - \gamma P^{\pi^*})^{-1} |\Delta r|, (I - \gamma P^{\tilde{\pi}^*})^{-1} |\Delta r| \right\},$$

where π^* and $\tilde{\pi}^*$ denote the optimal policy induced by Q^* and \tilde{Q}^* , respectively.

Proof. This lemma can be proved by showing the following two conditions:

$$\max(Q^* - \tilde{Q}^*, \mathbf{0}) \leq (I - \gamma P^{\pi^*}) |\Delta r|, \quad (17)$$

$$\max(\tilde{Q}^* - Q^*, \mathbf{0}) \leq (I - \gamma P^{\tilde{\pi}^*}) |\Delta r|. \quad (18)$$

For condition (17), we can prove the following:

$$\begin{aligned}
Q^* - \tilde{Q}^* & = r + \gamma P^{\pi^*} Q^* - (r + \Delta r + \gamma P^{\tilde{\pi}^*} \tilde{Q}^*) \\
& \leq |\Delta r| + \gamma P^{\pi^*} (Q^* - \tilde{Q}^*) \\
& \leq |\Delta r| + \gamma P^{\pi^*} \max(Q^* - \tilde{Q}^*, \mathbf{0}),
\end{aligned}$$

where the second inequality comes from $P^{\pi^*} \tilde{Q}^* \leq P^{\tilde{\pi}^*} \tilde{Q}^*$.

Since the right-hand-side term is positive in all entries. Thus, we have:

$$\max(Q^* - \tilde{Q}^*, \mathbf{0}) \leq |\Delta r| + \gamma P^{\pi^*} \max(Q^* - \tilde{Q}^*, \mathbf{0}).$$

Rearranging the inequality yields Eq. (17).

For proving Eq. (18), we follow a similar routine and get:

$$\begin{aligned} \tilde{Q}^* - Q^* &= (r + \Delta r + \gamma P^{\tilde{\pi}^*} \tilde{Q}^*) - (r + \gamma P^{\pi^*} Q^*) \\ &\leq |\Delta r| + \gamma P^{\tilde{\pi}^*} (\tilde{Q}^* - Q^*) \\ &\leq |\Delta r| + \gamma P^{\tilde{\pi}^*} \max(\tilde{Q}^* - Q^*, \mathbf{0}). \end{aligned}$$

Hence, we have:

$$\max(\tilde{Q}^* - Q^*, \mathbf{0}) \leq |\Delta r| + \gamma P^{\pi^*} \max(\tilde{Q}^* - Q^*, \mathbf{0}).$$

Rearranging the inequality yields Eq. (18). Using $|Q^* - \tilde{Q}^*| = \max(\max(Q^* - \tilde{Q}^*, \mathbf{0}), \max(\tilde{Q}^* - Q^*, \mathbf{0}))$ and combining Eq. (17)-(18) yields the desired result. \square

B Proof of Theorem 5.1

We aim to ensure the total number of samples D_ω satisfies the following lower bound:

$$D_\omega \geq N_{\text{entry}} \cdot N + D_r, \quad (19)$$

where N_{entry} denotes the number of unique state-action pairs that must be sampled, N is the sampling frequency for each pair, and D_r is the sample complexity required to estimate the exact reward function.

The term N_{entry} is directly specified in Algorithm 1. It is defined as the total number of state-action pairs across all components to be sampled. Mathematically:

$$N_{\text{entry}} = \sum_{i \in [\kappa_p]} \max_{k \in \mathcal{G}_i} |\mathcal{X}[Z_k^P]| \leq \sum_{k \in [\kappa_p]} |\mathcal{X}[Z_k^P]|,$$

where κ_p is the total number of sampling sets, \mathcal{G}_i is the set of component indices associated with the i -th sampling set, $|\mathcal{X}[Z_k^P]|$ denotes the size of the state-action space for component k .

For the sample complexity associated with estimating the reward function, D_r , it is sufficient to sample all necessary state-action pairs once for each reward component r_i to obtain the exact reward values. Similar to the transition kernel sampling, the sample complexity of the reward function is bounded by:

$$D_r \leq \max_{i \in [\kappa_r]} |\mathcal{X}[Z_i^R]|$$

where κ_r is the minimal number of sampling sets required to estimate the reward function, and $|\mathcal{X}[Z_i^R]|$ is the size of the state-action space for reward component r_i .

Therefore, we only need to determine the required sampling frequency N . Recall that, the error $\|Q^* - \hat{Q}_\omega^*\|_\infty$ comes from two aspects: 1) the computation error in Algorithm 2 due to finite value function iterations, and 2) finite sample error due to inaccurate estimation of \hat{P} . Since the value iteration algorithm converges exponentially fast, according to Section 5.2 in (Shi et al., 2024), $T = \bar{c}_0 \log(\frac{1}{(1-\gamma)\epsilon})$ is enough to guarantee the computation error $\leq \mathcal{O}(\epsilon)$. Instead, we focus on the estimation error due to finite samples. To do so, we decompose the estimation error of the Q -value function into two terms, the bias \mathcal{E}_ω and the finite sample error α_N . By applying the triangle inequality to the error, we get:

$$\|Q^* - \hat{Q}_\omega^*\|_\infty \leq \underbrace{\|Q^* - Q_\omega^*\|_\infty}_{\text{Bias } \mathcal{E}_\omega} + \underbrace{\|Q_\omega^* - \hat{Q}_\omega^*\|_\infty}_{\text{Finite Sample Error } \alpha_N}, \quad (20)$$

where

- Q^* is the optimal Q -value function induced by the actual optimal policy π^* , transition kernel P , and reward function r .
- \widehat{Q}_ω^* is the estimated Q -value function based on the estimated policy $\widehat{\pi}_\omega^*$, estimated transition kernel \widehat{P}_ω , and estimated reward function r_ω , where

$$\widehat{P}_\omega(s'|x) = \prod_{k=1}^{K_\omega} \widehat{P}_k(s'[Z_k^S] | x[Z_k^P]), \forall s' \in \mathcal{S}, \forall x \in \mathcal{X},$$

$$r_\omega(x) = \sum_{i=1}^{\ell_\omega} r_i(x[Z_i^R]), \forall x \in \mathcal{X}.$$

- Q_ω^* is the optimal Q -value function based on the factorization scheme ω , which assumes the utilization of infinite samples. It is induced by the policy π_ω , transition kernel P_ω , and reward function r_ω , where $P_\omega(s'|x)$ is defined as:

$$P_\omega(s'|x) := \lim_{N \rightarrow \infty} \widehat{P}_\omega(s'|x).$$

The limits exists due to the fixed sampling algorithm and the law of large numbers. DUE to the deterministic mapping between P and Q , we have $\lim_{N \rightarrow \infty} \widehat{Q}_\omega^* = Q_\omega^*$ with probability 1.

Thus, the finite sample error term $\alpha_N := \|Q_\omega^* - \widehat{Q}_\omega^*\|_\infty$ in the decomposition can be made arbitrarily small as N increases. Instead, the bias \mathcal{E}_ω doesn't vanish due to the approximation errors from factorization.

For the bias term \mathcal{E}_ω , it can be bounded as follows:

Lemma B.1 (Proof in Appendix D.1). *Given any factorization scheme ω , the following condition holds:*

$$\mathcal{E}_\omega := \|Q^* - Q_\omega^*\|_\infty \leq \frac{\Delta_\omega^R}{1-\gamma} + \frac{\gamma \Delta_\omega^P}{(1-\gamma)^2}. \quad (21)$$

Then, we focus on the finite sample error term α_N . Using the result in Lemma A.2, we have:

$$Q_\omega^* - \widehat{Q}_\omega^* \leq \underbrace{\gamma(I - \gamma \widehat{P}_\omega^{\pi_\omega^*})^{-1} (P_\omega - \widehat{P}_\omega) V_\omega^*}_{:=\Delta_1}, \quad (22)$$

$$Q_\omega^* - \widehat{Q}_\omega^* \geq \underbrace{\gamma(I - \gamma \widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} (P_\omega - \widehat{P}_\omega) V_\omega^*}_{:=\Delta_2}, \quad (23)$$

where $\widehat{P}_\omega^{\pi_\omega^*} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ represents the transition matrix of Markov chain $\{(s_t, a_t)\}$ induced by policy π_ω^* under transition kernel \widehat{P}_ω , and $\widehat{P}_\omega^{\widehat{\pi}_\omega^*} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ is induced by policy $\widehat{\pi}_\omega^*$ under transition kernel \widehat{P}_ω .

Based on Eq. (22) and (23), the error term α_N satisfies:

$$\alpha_N := \|Q_\omega^* - \widehat{Q}_\omega^*\|_\infty \leq \max\{\|\Delta_1\|_\infty, \|\Delta_2\|_\infty\}, \quad (24)$$

Now we only need to control $\|\Delta_1\|_\infty$ and $\|\Delta_2\|_\infty$. We first control the absolute value of the common term $(P_\omega - \widehat{P}_\omega)V_\omega^*$ for both Δ_1 and Δ_2 in Lemma B.2.

Lemma B.2 (Proof in Appendix D.2). *Given sample size N , then with probability at least $1-\delta$, the following condition holds:*

$$|(P_\omega - \widehat{P}_\omega)V_\omega^*| \leq \frac{2 \log(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P])}{3N(1-\gamma)} \cdot \mathbf{1} + \sqrt{\frac{2 \log(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]) \text{Var}_{P_\omega}(V_\omega^*)}{N}},$$

where $|\cdot|$ denotes the absolute value function, $\text{Var}_{P_\omega}(V_\omega^*)$ denotes the value function variance with transition kernel P_ω , as defined in Definition A.1.

Using the naive bound $\|\gamma(I - \gamma\widehat{P}_\omega^\pi)^{-1}\|_\infty \leq \frac{\gamma}{1-\gamma}$ for any π , and combining Lemma B.2 with Eq. (22) and (23), we can show that

$$|\Delta_1| \leq \frac{2\gamma \log(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)}{3N(1-\gamma)^2} \cdot \mathbf{1} + \gamma \sqrt{\frac{2 \log(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)}{N}} (I - \gamma\widehat{P}_\omega^{\pi^*})^{-1} \sqrt{\text{Var}_{P_\omega}(V_\omega^*)}, \quad (25)$$

$$|\Delta_2| \leq \frac{2\gamma \log(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)}{3N(1-\gamma)^2} \cdot \mathbf{1} + \gamma \sqrt{\frac{2 \log(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)}{N}} (I - \gamma\widehat{P}_\omega^{\pi^*})^{-1} \sqrt{\text{Var}_{P_\omega}(V_\omega^*)}. \quad (26)$$

The remaining challenge is to control $(I - \gamma\widehat{P}_\omega^{\pi^*})^{-1} \sqrt{\text{Var}_{P_\omega}(V_\omega^*)}$ and $(I - \gamma\widehat{P}_\omega^{\pi^*})^{-1} \sqrt{\text{Var}_{P_\omega}(V_\omega^*)}$ for Δ_1 and Δ_2 , respectively. We first bound the term $(I - \gamma\widehat{P}_\omega^{\pi^*})^{-1} \sqrt{\text{Var}_{P_\omega}(V_\omega^*)}$ for Δ_1 , which is slightly more complex, and bounding another term is analogous. Specifically,

$$\begin{aligned} & |(I - \gamma\widehat{P}_\omega^{\pi^*})^{-1} \sqrt{\text{Var}_{P_\omega}(V_\omega^*)}| \\ &= (I - \gamma\widehat{P}_\omega^{\pi^*})^{-1} \sqrt{\text{Var}_{P_\omega}(V_\omega^*) - \text{Var}_{\widehat{P}_\omega}(V_\omega^*) + \text{Var}_{\widehat{P}_\omega}(V_\omega^*)} \\ &= (I - \gamma\widehat{P}_\omega^{\pi^*})^{-1} \sqrt{P_\omega(V_\omega^*)^2 - (P_\omega V_\omega^*)^2 - \widehat{P}_\omega(V_\omega^*)^2 + (\widehat{P}_\omega V_\omega^*)^2 + \text{Var}_{\widehat{P}_\omega}(V_\omega^*)} \\ &= (I - \gamma\widehat{P}_\omega^{\pi^*})^{-1} \sqrt{(P_\omega - \widehat{P}_\omega)(V_\omega^*)^2 - ((P_\omega V_\omega^*)^2 - (\widehat{P}_\omega V_\omega^*)^2) + \text{Var}_{\widehat{P}_\omega}(V_\omega^*)} \\ &\leq (I - \gamma\widehat{P}_\omega^{\pi^*})^{-1} \left(\sqrt{|(P_\omega - \widehat{P}_\omega)(V_\omega^*)^2|} + \sqrt{|(P_\omega V_\omega^*)^2 - (\widehat{P}_\omega V_\omega^*)^2|} + \sqrt{\text{Var}_{\widehat{P}_\omega}(V_\omega^*)} \right) \\ &= \underbrace{(I - \gamma\widehat{P}_\omega^{\pi^*})^{-1} \sqrt{|(P_\omega - \widehat{P}_\omega)(V_\omega^*)^2|}}_{T_1} + \underbrace{(I - \gamma\widehat{P}_\omega^{\pi^*})^{-1} \sqrt{|(P_\omega V_\omega^*)^2 - (\widehat{P}_\omega V_\omega^*)^2|}}_{T_2} \\ &\quad + \underbrace{(I - \gamma\widehat{P}_\omega^{\pi^*})^{-1} \sqrt{\text{Var}_{\widehat{P}_\omega}(V_\omega^*)}}_{T_3}. \end{aligned}$$

This splition allows us to focus on T_1 , T_2 and T_3 , which can be controlled by the following lemmas:

Lemma B.3 (Proof in Appendix D.3). *With a probability at least $1 - \delta$, T_1 satisfies:*

$$T_1 \leq \frac{1}{(1-\gamma)^2} \sqrt[4]{\frac{2 \log(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)}{N}} \cdot \mathbf{1}.$$

Lemma B.4 (Proof in Appendix D.4). *With a probability at least $1 - \delta$, T_2 satisfies:*

$$T_2 \leq \frac{\sqrt{2}}{(1-\gamma)^2} \sqrt[4]{\frac{2 \log(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)}{N}} \cdot \mathbf{1}.$$

Lemma B.5 (Proof in Appendix D.5). *The term T_3 can be bounded as follows:*

$$T_3 \leq \left(\frac{4\gamma}{(1-\gamma)^3} \sqrt{\frac{2 \log(4|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)}{N}} + \sqrt{\frac{1+\gamma}{(1-\gamma)^3}} \right) \cdot \mathbf{1}.$$

Combining Lemmas B.2-B.5 yields the bound of Δ_2 :

Lemma B.6 (Proof in Appendix D.6). *With probability at least $1 - \delta$, the estimated Q -function \widehat{Q}^* satisfies:*

$$|\Delta_2| \leq 18 \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N(1-\gamma)^3} \right) + 6 \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N(1-\gamma)^3} \right)^{\frac{1}{2}}.$$

Analogously, we can show that with probability at least $1 - \delta$, $|\Delta_1|$ also satisfies:

$$|\Delta_1| \leq 18 \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N(1-\gamma)^3} \right) + 6 \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N(1-\gamma)^3} \right)^{\frac{1}{2}}.$$

Combining the bounds on $|\Delta_1|$ and $|\Delta_2|$, we have that with probability at least $1 - \delta$,

$$\begin{aligned} \|Q_\omega^* - \widehat{Q}_\omega^*\|_\infty &\leq \max(|\Delta_1|, |\Delta_2|) \\ &\leq 18 \left(\frac{\log \left(24 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N(1-\gamma)^3} \right) + 6 \left(\frac{\log \left(24 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N(1-\gamma)^3} \right)^{\frac{1}{2}}. \end{aligned}$$

Taking any $\epsilon = \|Q_\omega^* - \widehat{Q}_\omega^*\|_\infty \leq 1$, we can first verify that $\frac{\log(24 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)}{N(1-\gamma)^3} \leq 1$. Therefore,

$$\epsilon \leq (18 + 6) \left(\frac{\log \left(24 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N(1-\gamma)^3} \right)^{\frac{1}{2}}.$$

It directly yields the bound for N with probability at least $1 - \delta$:

$$N \geq \frac{576 \log \left(24 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{(1-\gamma)^3 \epsilon^2}. \quad (27)$$

Substituting Eq. (27) into Eq. (19) yields:

$$D_\omega \geq \frac{576 (\kappa_p \max_{k \in [K_\omega]} |\mathcal{X}[Z_k^P]|) \log \left(24 |\mathcal{X}[\cup_{k \in [K_\omega]} Z_k^P]| \delta^{-1} \right)}{\epsilon^2 (1-\gamma)^3} + \kappa_r \max_{i \in [\ell_\omega]} |\mathcal{X}[Z_i^R]|.$$

Letting $\bar{c}_0 = 576$ and $\bar{c}_1 = 24$ leads to our result.

C Proof of Theorem 6.1

Proof. Similar to the model-based case, we show the total amount of samples D_ω should satisfy:

$$D_\omega \geq N_{\text{entry}} \cdot N + D_r, \quad (28)$$

where N_{entry} denotes the number of sampled state-action pairs for generating a single empirical Bellman operator in Algorithm 3, and N denotes the number of generated empirical Bellman operators. Notation D_r denotes the sample complexity of finding the exact reward function.

The term N_{entry} can be directly obtained in Algorithm 3, which equals the sum of numbers of all state-action pairs in all decomposed components, satisfying:

$$N_{\text{entry}} = \sum_{i \in [\kappa_p]} \max_{k \in \mathcal{G}_i} |\mathcal{X}[Z_k^P]| \leq \sum_{k \in [\kappa_p]} |\mathcal{X}[Z_k^P]|,$$

where κ_p is the total number of sampling sets, \mathcal{G}_i is the set of component indices associated with the i -th sampling set, $|\mathcal{X}[Z_k^P]|$ denotes the size of the state-action space for component k .

For the sample complexity associated with estimating the reward function, D_r , it is sufficient to sample all necessary state-action pairs once for each reward component r_i to obtain the exact reward values. Similar to the transition kernel sampling, the sample complexity of the reward function is bounded by:

$$D_r \leq \sum_{i \in [\kappa_r]} |\mathcal{X}[Z_i^R]|$$

where κ_r is the minimal number of sampling sets required to estimate the reward function, and $|\mathcal{X}[Z_i^R]|$ is the size of the state-action space for reward component r_i .

For the sampling times N , according to Algorithm 4, we can conclude that:

$$N = \sum_{\tau=1}^T (N_\tau + M) = \sum_{\tau=1}^T N_\tau + MT, \quad (29)$$

where T is the number of epochs, N_τ is the sampling frequency for estimating the reference Bellman operator in the τ -th epoch, M is the number of variance-reduced updates in a single epoch.

The key of the proof is to show the estimation error decays exponentially when the number of epochs increases, i.e.,

Lemma C.1 (Proof in Appendix D.7). *For any $\delta \geq 0$, with probability at least $1 - \delta/T$, the Q -function estimate \bar{Q}_τ after τ epochs satisfies:*

$$\|\bar{Q}_\tau - Q_\omega^*\|_\infty \leq \frac{1}{(1-\gamma)2^\tau}, \quad \forall \tau = 1, \dots, T, \quad (30)$$

provided that the number of samples M and the number of iterations N_τ satisfy the following conditions:

$$M = c_2 \frac{\log \left(\frac{6T |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{(1-\gamma)^\delta} \right)}{(1-\gamma)^3}, \quad (31)$$

$$N_\tau = c_3 4^\tau \frac{\log \left(6T |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{(1-\gamma)^2}, \quad (32)$$

where $\eta_t = \frac{1}{1+(1-\gamma)(t+1)}$, $c_2, c_3 > 0$ are sufficiently large constants.

We first ensure the error decreases to a relatively small value $\frac{1}{\sqrt{1-\gamma}}$ after T_1 epochs. Specifically, we set:

$$T_1 = \left\lceil \log_2 \left(\frac{1}{\sqrt{1-\gamma}} \right) \right\rceil. \quad (33)$$

Using the union bound over all $\tau \leq T_1$ in Eq. (30), we obtain that with probability at least $1 - \delta$:

$$\|\bar{Q}_{T_1} - Q^*\|_\infty \leq \frac{1}{\sqrt{1-\gamma}}.$$

Substituting N_τ , T_1 , and M into Eq. (29), the total number of samples for this phase, denoted $N_{[1]}$, is bounded by:

$$N_{[1]} \leq c_4 \cdot \frac{\log \left(\frac{6T_1 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{(1-\gamma)^\delta} \right) \cdot \log \left(\frac{1}{1-\gamma} \right)}{(1-\gamma)^3}, \quad (34)$$

where c_4 is a constant depending on c_2 and c_3 .

We now show that from initial point $\bar{Q}_0 = \bar{Q}_{T_1}$, an additional $T_2 = \left\lceil c \log_2 \left(\frac{1}{\sqrt{1-\gamma}\epsilon} \right) \right\rceil$ epochs are sufficient to reduce the error to ϵ , where $c > 0$ is a constant.

We utilize the following lemma:

Lemma C.2 (Proof in Appendix D.15). *Given $\|\bar{Q}_0 - Q^*\|_\infty \leq \frac{1}{\sqrt{1-\gamma}}$, and setting:*

$$M = c_2 \cdot \frac{\log \left(\frac{6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{(1-\gamma)\delta} \right)}{(1-\gamma)^3}, \quad (35)$$

$$N_\tau = c_3 \cdot 4^\tau \cdot \frac{\log \left(6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{(1-\gamma)^2}, \quad (36)$$

then, with probability at least $1 - \delta$:

$$\|\bar{Q}_\tau - Q_\omega^*\|_\infty \leq \frac{1}{2^\tau \sqrt{1-\gamma}}, \quad \forall \tau \geq 0.$$

The proof of Lemma C.2 follows a similar routine to that of Lemma C.1 but leverages the better initial estimate \bar{Q}_0 . This lemma indicates that starting from an initial error of $\frac{1}{\sqrt{1-\gamma}}$, the variance-reduced iteration halves the error at each step.

After $T_2 = \left\lceil c' \log_2 \left(\frac{1}{\sqrt{1-\gamma}\epsilon} \right) \right\rceil$ epochs (with $c' > 0$), the error reduces to ϵ . The total number of samples for this phase, denoted $N_{[2]}$, is bounded by:

$$N_{[2]} \leq c_5 \cdot \frac{\log \left(\frac{6T_1|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{(1-\gamma)\delta} \right) \cdot \log \left(\frac{1}{1-\gamma} \right)}{(1-\gamma)^3 \epsilon^2}, \quad (37)$$

where c_5 is a sufficiently large constant.

Combining the samples from both phases, the total number of samples required is:

$$N = N_{[1]} + N_{[2]}.$$

Substituting back into Eq. (28), we conclude that the total sample complexity D_ω satisfies the desired bound:

$$D_\omega \geq N_{\text{entry}} \cdot (N_{[1]} + N_{[2]}) + D_r.$$

This completes the proof. □

D Proof for Auxiliary Lemmas

D.1 Proof for Lemma B.1

Proof. We first write the Bellman equation of Q^* and Q_ω^* as follows:

$$Q^*(s, a) = r(s, a) + \gamma \sum_{s'} P(s'|s, a) \max_{a'} Q^*(s', a'), \quad (38)$$

$$Q_\omega^*(s, a) = r_\omega(s, a) + \gamma \sum_{s'} P_\omega(s'|s, a) \max_{a'} Q_\omega^*(s', a'). \quad (39)$$

Subtracting Eq. (38) with (39) yields:

$$\begin{aligned}
Q^*(s, a) - Q_\omega^*(s, a) &= r(s, a) - r_\omega(s, a) + \gamma \sum_{s'} (P(s'|s, a) \max_{a'} Q^*(s', a') - P_\omega(s'|s, a) \max_{a'} Q_\omega^*(s', a')) \\
&= r(s, a) - r_\omega(s, a) + \gamma \sum_{s'} (P(s'|s, a) (\max_{a'} Q^*(s', a') - \max_{a'} Q_\omega^*(s', a')) \\
&\quad + (P(s'|s, a) - P_\omega(s'|s, a)) \max_{a'} Q_\omega^*(s', a'))
\end{aligned}$$

Taking absolute value on both side yields:

$$\begin{aligned}
|Q^*(s, a) - Q_\omega^*(s, a)| &\leq \|r - r_\omega\|_\infty + \gamma \sum_{s'} (P(s'|s, a) (\max_{a'} Q^*(s', a') - \max_{a'} Q_\omega^*(s', a')) \\
&\quad + \gamma \|P - P_\omega\|_\infty \max_{s', a'} Q_\omega^*(s', a')) \\
&\leq \|r - r_\omega\|_\infty + \gamma \|Q^* - Q_\omega^*\|_\infty + \gamma \|P - P_\omega\|_\infty \max_{s', a'} Q_\omega^*(s', a'), \quad (40)
\end{aligned}$$

where the last inequality comes from $|\max_a Q^*(s, a) - \max_a Q_\omega^*(s, a)| \leq \max_a |Q^*(s, a) - Q_\omega^*(s, a)| \leq \|Q^* - Q_\omega^*\|_\infty$. Due to $Q_\omega^*(s', a') \leq \frac{1}{1-\gamma}$ for any state-action pair (s', a') , we have:

$$\begin{aligned}
\|Q^* - Q_\omega^*\|_\infty &= \max_{s, a} |Q^*(s, a) - Q_\omega^*(s, a)| \\
&\leq \|r - r_\omega\|_\infty + \gamma \|Q^* - Q_\omega^*\|_\infty + \frac{\gamma}{1-\gamma} \|P - P_\omega\|_\infty. \quad (41)
\end{aligned}$$

Standard mathematical manipulation on Eq. (41) yields:

$$\|Q^* - Q_\omega^*\|_\infty \leq \frac{\|r - r_\omega\|_\infty}{1-\gamma} + \frac{\gamma \|P - P_\omega\|_\infty}{(1-\gamma)^2}. \quad (42)$$

Applying the definitions of approximation errors, we have:

$$\|Q^* - Q_\omega^*\|_\infty \leq \frac{\Delta_\omega^R}{1-\gamma} + \frac{\gamma \Delta_\omega^P}{(1-\gamma)^2}. \quad (43)$$

This concludes our proof. \square

D.2 Proof for Lemma B.2

Proof. We leverage the structure of the factorized transition kernel and show that the vector $(P_\omega - \hat{P}_\omega)V_\omega^*$ contains multiple identical entries. By identifying the distinct entries, we can focus our analysis on a subset of state-action pairs.

Recall that $(P_\omega - \hat{P}_\omega)V_\omega^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times 1}$ represents the difference between the actual and estimated reward vector. Due to the factorized form of the transition kernel P_ω , the estimated transition probability $\hat{P}(s' | x)$ is given by:

$$\hat{P}(s' | x) = \prod_{k=1}^{K_\omega} \hat{P}_k(s'[Z_k^S] | x[Z_k^P]),$$

where each $\hat{P}_k(s'[Z_k^S] | x[Z_k^P])$ depends only on the subset Z_k^P of the state-action pair x . As a result, only the state-action components $x[Z_k^P]$ for $k \in [K_\omega]$ determine the transition probabilities. Thus, there are at most $|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|$ distinct rows in the matrix $(P_\omega - \hat{P}_\omega)$.

We use \mathcal{X}^* to denote a set of state-action pairs indicating distinct entries, satisfying:

$$\mathcal{X}^* = \left\{ x \in \mathcal{X} \mid \forall x', x'' \in \mathcal{X}^*, x' \neq x'' \Rightarrow \exists k \in [K_\omega] \text{ such that } x'[Z_k^P] \neq x''[Z_k^P] \right\}. \quad (44)$$

For any $x \in \mathcal{X}^*$, we can easily verify that the estimation $\widehat{P}(s' | x)$ is unbiased:

$$\begin{aligned} & \mathbb{E}[\widehat{P}(s' | x) - P(s' | x)] \\ &= \mathbb{E} \left[\prod_{i \in [\kappa_p]} \left(\prod_{k \in \mathcal{G}_i} \widehat{P}_k(s'[Z_k^S] | x[Z_k^P]) \right) - \prod_{i \in [\kappa_p]} \left(\prod_{k \in \mathcal{G}_i} P_k(s'[Z_k^S] | x[Z_k^P]) \right) \right] \\ &= \mathbb{E} \left[\prod_{i \in [\kappa_p]} \widehat{P}(s'[\cup_{k \in \mathcal{G}_i} Z_k^S] | x[\cup_{k \in \mathcal{G}_i} Z_k^P]) - \prod_{i \in [\kappa_p]} P(s'[\cup_{k \in \mathcal{G}_i} Z_k^S] | x[\cup_{k \in \mathcal{G}_i} Z_k^P]) \right] \\ &= \prod_{i \in [\kappa_p]} \mathbb{E}[\widehat{P}(s'[\cup_{k \in \mathcal{G}_i} Z_k^S] | x[\cup_{k \in \mathcal{G}_i} Z_k^P])] - \prod_{i \in [\kappa_p]} P(s'[\cup_{k \in \mathcal{G}_i} Z_k^S] | x[\cup_{k \in \mathcal{G}_i} Z_k^P]) \\ &= 0. \end{aligned}$$

The third equality is due to the independent sampling of different sampling sets, and the last equality is due to $\mathbb{E}[\widehat{P}(s'[\cup_{k \in \mathcal{G}_i} Z_k^S] | x[\cup_{k \in \mathcal{G}_i} Z_k^P])] = P(s'[\cup_{k \in \mathcal{G}_i} Z_k^S] | x[\cup_{k \in \mathcal{G}_i} Z_k^P])$ within each sampling set due to the law of large numbers.

Also, since $\mathbf{0} \leq V_\omega^* \leq \frac{1}{1-\gamma} \cdot \mathbf{1}$, we have $\|(P_\omega - \widehat{P}_\omega)V_\omega^*\|_\infty \leq \frac{1}{1-\gamma}$. Combining with the Bernstein's inequality (Vershynin, 2018) with sample size N yields that:

$$P \left(|(P_\omega(\cdot|x) - \widehat{P}_\omega(\cdot|x))V_\omega^*| \geq t \right) \leq 2 \exp \left(- \frac{\frac{N^2}{2} t^2}{N \text{Var}_{P_\omega}(V_\omega^*)_{(x)} + N \frac{t}{3(1-\gamma)}} \right), \quad (45)$$

where $\text{Var}_{P_\omega}(V_\omega^*)_{(x)}$ denotes the entry of $\text{Var}_{P_\omega}(V_\omega^*)$ corresponding to x .

Letting the right-hand-side of Eq. (45) equal $\frac{\delta}{|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}$ leads to the following inequality:

$$\frac{\delta}{|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|} = 2 \exp \left(- \frac{\frac{N^2}{2} t^2}{N \text{Var}_{P_\omega}(V_\omega^*)_{(x)} + N \frac{t}{3(1-\gamma)}} \right).$$

Rearranging this term yields that

$$t \leq \frac{2 \log(|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)}{3N(1-\gamma)} + \sqrt{\frac{2 \log(|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|) \text{Var}_{P_\omega}(V_\omega^*)_{(x)}}{N}}.$$

Taking the union bound across all state-action pairs $x \in \mathcal{X}^*$, and using the identical entry property yields that, with probability at least $1 - \delta$:

$$\left| (P_\omega - \widehat{P}_\omega)V_\omega^* \right| \leq \sqrt{\frac{2 \log(|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|) \text{Var}_{P_\omega}(V_\omega^*)_{(x)}}{N}} + \frac{2 \log(|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)}{3N(1-\gamma)} \cdot \mathbf{1}.$$

This concludes our proof. \square

D.3 Proof for Lemma B.3

Proof. We have that

$$\|(P_\omega - \widehat{P}_\omega)(V_\omega^*)^2\|_\infty = \left\| \left(\prod_{k=1}^{K_\omega} P_k(s'[Z_k^S] | x[Z_k^P]) - \prod_{k=1}^{K_\omega} \widehat{P}_k(s'[Z_k^S] | x[Z_k^P]) \right) (V_\omega^*)^2 \right\|_\infty.$$

Following the same routine in the proof of Lemma B.2, we know $(P_\omega - \widehat{P}_\omega)(V_\omega^*)^2$ contains at most $|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|$ distinct entries.

Meanwhile, for each single entry of $(P_\omega - \widehat{P}_\omega)(V_\omega^*)^2$, denoted by $(P_\omega(\cdot|x) - \widehat{P}_\omega(\cdot|x))(V_\omega^*)^2$, we can show the following conditions:

$$\begin{aligned}\mathbb{E}((P_\omega(\cdot|x) - \widehat{P}_\omega(\cdot|x))(V_\omega^*)^2) &= 0, \forall x \in \mathcal{X}, \\ \|(P_\omega - \widehat{P}_\omega)(V_\omega^*)^2\|_\infty &\leq \|\widehat{P}_\omega(V_\omega^*)^2\|_\infty + \|P_\omega(V_\omega^*)^2\|_\infty \leq \frac{2}{(1-\gamma)^2}.\end{aligned}$$

With N *i.i.d.* samples for estimating $(P_\omega(\cdot|x) - \widehat{P}_\omega(\cdot|x))(V_\omega^*)^2 \in \mathbb{R}$, we apply the standard Hoeffding's inequality to $(P_\omega(\cdot|x) - \widehat{P}_\omega(\cdot|x))(V_\omega^*)^2$ as follows:

$$P\left(|(P_\omega(\cdot|x) - \widehat{P}_\omega(\cdot|x))(V_\omega^*)^2| \geq \epsilon\right) \leq 2 \exp\left(-\frac{2N\epsilon^2}{\left(\frac{2}{(1-\gamma)^2}\right)^2 \cdot N}\right) = 2 \exp\left(-\frac{(1-\gamma)^4 N \epsilon^2}{2}\right), \forall x \in \mathcal{X},$$

Letting the right-hand-side term be $\frac{\delta}{|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}$ and applying the union bound across all distinct state-action pair $x \in \mathcal{X}^*$ (with definition in Eq. (44)) yield:

$$\|(P_\omega - \widehat{P}_\omega)(V_\omega^*)^2\|_\infty \leq \frac{1}{(1-\gamma)^2} \cdot \sqrt{\frac{2 \log\left(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|\right)}{N}}.$$

Thus, T_1 satisfies:

$$\begin{aligned}T_1 &= (I - \gamma \widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \sqrt{|(P_\omega - \widehat{P}_\omega)(V_\omega^*)^2|} \\ &\leq \|(I - \gamma \widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1}\|_\infty \sqrt{\|(P_\omega - \widehat{P}_\omega)(V_\omega^*)^2\|_\infty} \cdot \mathbf{1} \\ &\leq \frac{1}{(1-\gamma)^2} \sqrt[4]{\frac{2 \log\left(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|\right)}{N}} \cdot \mathbf{1}.\end{aligned}$$

This concludes our proof. □

D.4 Proof for Lemma B.4

Proof. We first bound $\|(P_\omega V_\omega^*)^2 - (\widehat{P}_\omega V_\omega^*)^2\|_\infty$ as follows:

$$\begin{aligned}\|(P_\omega V_\omega^*)^2 - (\widehat{P}_\omega V_\omega^*)^2\|_\infty &= \|(P_\omega V_\omega^* + \widehat{P}_\omega V_\omega^*)(P_\omega V_\omega^* - \widehat{P}_\omega V_\omega^*)\|_\infty \\ &\leq \|P_\omega V_\omega^* + \widehat{P}_\omega V_\omega^*\|_\infty \|P_\omega V_\omega^* - \widehat{P}_\omega V_\omega^*\|_\infty \\ &\leq 2\|V_\omega^*\|_\infty \|(P_\omega - \widehat{P}_\omega)V_\omega^*\|_\infty.\end{aligned}$$

Applying the Hoeffding's inequality to $\|(P_\omega - \widehat{P}_\omega)V_\omega^*\|_\infty$ yields that, with probability at least $1 - \delta$:

$$\|(P_\omega - \widehat{P}_\omega)V_\omega^*\|_\infty \leq \frac{1}{1-\gamma} \sqrt{\frac{2 \log\left(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|\right)}{N}}. \quad (46)$$

Therefore, T_2 satisfies:

$$T_2 = (I - \gamma \widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \sqrt{|(P_\omega V_\omega^*)^2 - (\widehat{P}_\omega V_\omega^*)^2|}$$

$$\begin{aligned}
&\leq \|(I - \gamma \widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1}\|_\infty \sqrt{\|(P_\omega V_\omega^*)^2 - (\widehat{P}_\omega V_\omega^*)^2\|_\infty} \cdot \mathbf{1} \\
&\leq \frac{1}{1-\gamma} \sqrt{2\|V_\omega^*\|_\infty \|(P_\omega - \widehat{P}_\omega)V_\omega^*\|_\infty} \cdot \mathbf{1} \\
&\leq \frac{\sqrt{2}}{(1-\gamma)^2} \sqrt[4]{\frac{2 \log \left(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N}} \cdot \mathbf{1}.
\end{aligned}$$

□

D.5 Proof for Lemma B.5

Proof. We decompose T_3 as follows:

$$\begin{aligned}
T_3 &= (I - \gamma \widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \sqrt{\text{Var}_{\widehat{P}_\omega}(V_\omega^*)} \\
&= (I - \gamma \widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \sqrt{\text{Var}_{\widehat{P}_\omega}(V_\omega^* - \widehat{V}_\omega^{\pi_\omega^*} + \widehat{V}_\omega^{\pi_\omega^*} - \widehat{V}_\omega^{\widehat{\pi}_\omega^*} + \widehat{V}_\omega^{\widehat{\pi}_\omega^*})} \\
&\leq (I - \gamma \widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \sqrt{2\text{Var}_{\widehat{P}_\omega}(V_\omega^* - \widehat{V}_\omega^{\pi_\omega^*}) + 2\text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\pi_\omega^*}) + 2\text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*} - \widehat{V}_\omega^{\pi_\omega^*})} \\
&\leq \underbrace{(I - \gamma \widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \sqrt{2\text{Var}_{\widehat{P}_\omega}(V_\omega^* - \widehat{V}_\omega^{\pi_\omega^*})}}_{T_{31}} + \underbrace{(I - \gamma \widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \sqrt{2\text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\pi_\omega^*})}}_{T_{32}} \\
&\quad + \underbrace{(I - \gamma \widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \sqrt{2\text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*} - \widehat{V}_\omega^{\pi_\omega^*})}}_{T_{33}}.
\end{aligned}$$

Then, we bound T_{31} , T_{32} and T_{33} separately.

Step 1. Bounding T_{31} :

For T_{31} , the following condition holds:

$$\begin{aligned}
T_{31} &= (I - \gamma \widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \sqrt{2\text{Var}_{\widehat{P}_\omega}(V_\omega^* - \widehat{V}_\omega^{\pi_\omega^*})} \\
&\leq \|(I - \gamma \widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1}\|_\infty \sqrt{2\|\text{Var}_{\widehat{P}_\omega}(V_\omega^* - \widehat{V}_\omega^{\pi_\omega^*})\|_\infty} \cdot \mathbf{1} \\
&\leq \frac{\sqrt{2}}{1-\gamma} \sqrt{\|V_\omega^* - \widehat{V}_\omega^{\pi_\omega^*}\|_\infty^2} \cdot \mathbf{1} \\
&\leq \frac{\sqrt{2}}{1-\gamma} \sqrt{\|Q_\omega^* - \widehat{Q}_\omega^{\pi_\omega^*}\|_\infty^2} \cdot \mathbf{1}.
\end{aligned}$$

Applying Lemma A.1 yields:

$$\begin{aligned}
\|Q_\omega^* - \widehat{Q}_\omega^{\pi_\omega^*}\|_\infty^2 &= \|\gamma(I - \gamma \widehat{P}_\omega^{\pi_\omega^*})^{-1}(P_\omega - \widehat{P}_\omega)V_\omega^*\|_\infty^2 \\
&\leq \frac{\gamma^2}{(1-\gamma)^2} \|(P_\omega - \widehat{P}_\omega)V_\omega^*\|_\infty^2.
\end{aligned}$$

Applying Eq. (46), we have that, with probability at least $1 - \delta$, the term $\|Q_\omega^* - \widehat{Q}_\omega^{\pi_\omega^*}\|_\infty^2$ satisfies:

$$\|Q_\omega^* - \widehat{Q}_\omega^{\pi_\omega^*}\|_\infty^2 \leq \frac{\gamma^2}{(1-\gamma)^4} \left(\sqrt{\frac{2 \log \left(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N}} \right)^2$$

$$= \frac{2\gamma^2 \log \left(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N(1-\gamma)^4}. \quad (47)$$

Therefore, with probability at least $1 - \delta$, T_{31} satisfies that:

$$T_{31} \leq \frac{2\gamma}{(1-\gamma)^3} \sqrt{\frac{2 \log \left(2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N}} \cdot \mathbf{1}.$$

□

Step 2. Bounding T_{32} :

Note that, $(1-\gamma)(I - \gamma\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1}$ is a matrix of probability with each row being a probability distribution. For a positive vector \mathbf{v} and distribution ν , Jensen's inequality implies that $\nu\sqrt{\mathbf{v}} \leq \sqrt{\nu \cdot \mathbf{v}}$ (inequality of expectation). This implies:

$$\begin{aligned} \|T_{32}\|_\infty &= \left\| (I - \gamma\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \sqrt{\text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*})} \right\|_\infty \\ &= \frac{1}{1-\gamma} \left\| (1-\gamma)(I - \gamma\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \sqrt{\text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*})} \right\|_\infty \\ &\leq \sqrt{\left\| \frac{1}{1-\gamma} (I - \gamma\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*}) \right\|_\infty}. \end{aligned} \quad (48)$$

Also, we can reformulate $\|(I - \gamma\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*})\|_\infty$ as follows:

$$\begin{aligned} \|(I - \gamma\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*})\|_\infty &= \|(I - \gamma\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} (I - \gamma^2\widehat{P}_\omega^{\widehat{\pi}_\omega^*}) (I - \gamma^2\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*})\|_\infty \\ &= \|(I - \gamma\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} (I - \gamma\widehat{P}_\omega^{\widehat{\pi}_\omega^*}) (I + \gamma\widehat{P}_\omega^{\widehat{\pi}_\omega^*}) (I - \gamma^2\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*})\|_\infty \\ &= \|(I + \gamma\widehat{P}_\omega^{\widehat{\pi}_\omega^*}) (I - \gamma^2\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*})\|_\infty \\ &\leq \|(I + \gamma\widehat{P}_\omega^{\widehat{\pi}_\omega^*})\|_\infty \|(I - \gamma^2\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*})\|_\infty \\ &\leq (1+\gamma) \|(I - \gamma^2\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*})\|_\infty. \end{aligned}$$

Thus, we have

$$\left\| (I - \gamma\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \sqrt{\text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*})} \right\|_\infty \leq \sqrt{\left\| \frac{1+\gamma}{1-\gamma} (I - \gamma^2\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*}) \right\|_\infty}.$$

We now connect this result with the definition of Σ_M^π in Lemma A.3:

$$\Sigma_M^{\widehat{\pi}_\omega^*} = \gamma^2 (1 - \gamma^2\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*}),$$

where the transition kernel of MDP M is \widehat{P}_ω . Therefore,

$$(I - \gamma^2\widehat{P}_\omega^{\widehat{\pi}_\omega^*})^{-1} \text{Var}_{\widehat{P}_\omega}(\widehat{V}_\omega^{\widehat{\pi}_\omega^*}) = \frac{\Sigma_M^{\widehat{\pi}_\omega^*}}{\gamma^2} \leq \frac{1}{(1-\gamma)^2} \cdot \mathbf{1},$$

where the inequality comes from $\|\Sigma_M^\pi\|_\infty \leq \frac{\gamma^2}{(1-\gamma)^2}$ for any policy π , which can be easily verified according to Definition A.2:

$$\|\Sigma_M^\pi\|_\infty = \max_{(s,a)} \left\| \mathbb{E} \left[\left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) - Q_M^\pi(s_0, a_0) \right)^2 \middle| s_0 = s, a_0 = a \right] \right\|_\infty$$

$$\begin{aligned} &\leq \max_{(s,a)} \mathbb{E} \left[\left\| \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) - Q_M^\pi(s_0, a_0) \right\|_{\infty}^2 \middle| s_0 = s, a_0 = a \right] \\ &\leq \frac{\gamma^2}{(1-\gamma)^2}. \end{aligned}$$

Substituting this condition into Eq. (48) yields:

$$T_{32} \leq \sqrt{\frac{1+\gamma}{(1-\gamma)^3}} \cdot \mathbf{1}.$$

Step 3. Bounding T_{33} : The term T_{33} satisfies:

$$\begin{aligned} T_{33} &= (I - \gamma \widehat{P}_{\omega}^{\widehat{\pi}_{\omega}^*})^{-1} \sqrt{2 \text{Var}_{\widehat{P}_{\omega}}(\widehat{V}_{\omega}^{\widehat{\pi}_{\omega}^*} - \widehat{V}_{\omega}^{\pi_{\omega}^*})} \\ &\leq \|(I - \gamma \widehat{P}_{\omega}^{\widehat{\pi}_{\omega}^*})^{-1}\|_{\infty} \sqrt{2 \|\text{Var}_{\widehat{P}_{\omega}}(\widehat{V}_{\omega}^{\widehat{\pi}_{\omega}^*} - \widehat{V}_{\omega}^{\pi_{\omega}^*})\|_{\infty}} \cdot \mathbf{1} \\ &\leq \frac{\sqrt{2}}{1-\gamma} \sqrt{\|\widehat{V}_{\omega}^{\widehat{\pi}_{\omega}^*} - \widehat{V}_{\omega}^{\pi_{\omega}^*}\|_{\infty}^2} \cdot \mathbf{1} \\ &\leq \frac{\sqrt{2}}{1-\gamma} \sqrt{\|\widehat{Q}_{\omega}^* - \widehat{Q}_{\omega}^{\pi_{\omega}^*}\|_{\infty}^2} \cdot \mathbf{1}. \end{aligned}$$

Similar to the bound of T_{31} in Eq. (47), we have that with probability at least $1 - \delta$,

$$\|\widehat{Q}_{\omega}^* - \widehat{Q}_{\omega}^{\pi_{\omega}^*}\|_{\infty}^2 \leq \frac{2\gamma^2 \log\left(2|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|\right)}{N(1-\gamma)^4}.$$

Thus, with probability at least $1 - \delta$, T_{33} satisfies:

$$T_{33} \leq \frac{2\gamma}{(1-\gamma)^3} \sqrt{\frac{2 \log\left(2|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|\right)}{N}} \cdot \mathbf{1}.$$

Step 4. Combining the Results:

Combining the upper bounds for T_{31} , T_{32} and T_{33} , we can bound T_3 that, with probability at least $1 - \delta$,

$$T_3 \leq \left(\frac{4\gamma}{(1-\gamma)^3} \sqrt{\frac{2 \log\left(4|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|\right)}{N}} + \sqrt{\frac{1+\gamma}{(1-\gamma)^3}} \right) \cdot \mathbf{1}.$$

D.6 Proof for Lemma B.6

Proof. Taking δ to be $\frac{\delta}{3}$, and applying Lemma B.2-B.5 yields:

$$\begin{aligned} |\Delta_2| &\leq \frac{3}{(1-\gamma)^3} \left(\frac{\log\left(12|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|\right)}{N} \right) + \frac{7}{(1-\gamma)^2} \left(\frac{\log\left(12|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|\right)}{N} \right)^{\frac{3}{4}} \\ &\quad + \sqrt{\frac{2}{(1-\gamma)^3}} \left(\frac{\log\left(12|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|\right)}{N} \right)^{1/2}. \end{aligned} \tag{49}$$

Applying Cauchy–Schwarz inequality yields:

$$\begin{aligned}
& \frac{3}{(1-\gamma)^3} \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N} \right) + \sqrt{\frac{2}{(1-\gamma)^3}} \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N} \right)^{1/2} \\
& \geq \sqrt{\frac{3}{(1-\gamma)^3} \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N} \right) \cdot \frac{2}{(1-\gamma)^3} \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N} \right)^{1/2}} \\
& = \sqrt{3} (1-\gamma)^{-\frac{9}{4}} \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N} \right)^{\frac{3}{4}} \\
& \geq \frac{\sqrt{3}}{(1-\gamma)^2} \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N} \right)^{\frac{3}{4}}.
\end{aligned}$$

Applying this condition to Eq. (49) yields:

$$\begin{aligned}
\|Q_\omega^* - \hat{Q}_\omega^*\|_\infty & \leq \frac{3(1 + \frac{7}{\sqrt{3}})}{(1-\gamma)^3} \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N} \right) + \frac{\sqrt{2}(1 + \frac{7}{\sqrt{3}})}{\sqrt{(1-\gamma)^3}} \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N} \right)^{1/2} \\
& \leq 18 \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N(1-\gamma)^3} \right) + 6 \left(\frac{\log \left(12 |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \right)}{N(1-\gamma)^3} \right)^{\frac{1}{2}}.
\end{aligned}$$

This concludes our proof. □

D.7 Proof for Lemma C.1

We prove this lemma by induction. We first show that the base case ($\tau = 1$) satisfies Eq. (30), and then prove the inductive condition when $\tau \geq 2$.

Step 1. Showing Base Case with $\tau = 1$:

Due to the initialization, we have $\bar{Q}_{\tau-1}(s, a) = \bar{Q}_0(s, a) = 0$ for any state-action pair (s, a) . Therefore, for $\tau = 1$, both the empirical and reference Bellman operators equal to the immediate reward:

$$\hat{\mathcal{H}}_t(\bar{Q}_{\tau-1})_{s,a} = r(s, a) \quad \text{and} \quad \bar{\mathcal{H}}_\tau(\bar{Q}_{\tau-1})_{s,a} = r(s, a), \quad \forall (s, a).$$

Given this, the variance-reduced update in Eq. (14) simplifies to the standard Q-learning update:

$$Q_t = (1 - \eta_{t-1})Q_{t-1} + \eta_{t-1}\hat{\mathcal{H}}_{t-1}(Q_{t-1}), \quad \forall t = 1, \dots, M.$$

Let $\Delta_t = Q_t - Q_\omega^*$ denote the estimation error at iteration t , where Q_ω^* is the unique fixed-point solution of the Bellman equation $\mathcal{H}(Q) = Q$. Then, we have:

$$\begin{aligned}
\Delta_t & = (1 - \eta_{t-1})\Delta_{t-1} + \eta_{t-1}(\hat{\mathcal{H}}_{t-1}(Q_\omega^* + \Delta_{t-1}) - \mathcal{H}(Q_\omega^*)) \\
& = (1 - \eta_{t-1})\Delta_{t-1} + \eta_{t-1}(\hat{\mathcal{H}}_{t-1}(Q_\omega^* + \Delta_{t-1}) - \hat{\mathcal{H}}_{t-1}(Q_\omega^*) + \hat{\mathcal{H}}_{t-1}(Q_\omega^*) - \mathcal{H}(Q_\omega^*))
\end{aligned}$$

$$= (1 - \eta_{t-1})\Delta_{t-1} + \eta_{t-1} \underbrace{(\widehat{\mathcal{H}}_{t-1}(Q_\omega^* + \Delta_{t-1}) - \widehat{\mathcal{H}}_{t-1}(Q_\omega^*))}_{\text{Contractive Error: } \mathcal{W}_{t-1}(\Delta_{t-1})} + \eta_{t-1} \underbrace{(\widehat{\mathcal{H}}_{t-1}(Q_\omega^*) - \mathcal{H}(Q_\omega^*))}_{\text{Random Error: } \mathcal{E}_{t-1}}. \quad (50)$$

Observe that the error iteration Δ_t consists of three components: the decaying term $(1 - \eta_{t-1})\Delta_{t-1}$, the contractive error term $\mathcal{W}_{t-1}(\Delta_{t-1})$, and the random error \mathcal{E}_{t-1} .

The contractive error $\mathcal{W}_{t-1}(\Delta_{t-1})$ depends on both Δ_{t-1} and the empirical Bellman operator $\widehat{\mathcal{H}}_{t-1}$, and it is bounded due to the γ -contractiveness of $\widehat{\mathcal{H}}_{t-1}$:

$$\|\mathcal{W}_{t-1}(\Delta_{t-1})\|_\infty \leq \gamma \|\Delta_{t-1}\|_\infty. \quad (51)$$

The random error \mathcal{E}_{t-1} is *i.i.d.* for different values of t . By applying the iteration in Eq. (50) and using the contraction condition in Eq. (51), we can express Δ_t fully as follows:

Lemma D.1 (Proof in Appendix D.8). *For any $t \geq 1$, the estimation error Δ_t is bounded above and below by:*

$$\Delta_t \leq \prod_{k=0}^{t-1} (1 - (1 - \gamma)\eta_k) \|\Delta_0\|_\infty \mathbf{1} + \gamma \eta_{t-1} \|P_{t-1}\|_\infty \mathbf{1} + \gamma \sum_{i=1}^{t-2} \left(\left(\prod_{j=i+1}^{t-1} (1 - (1 - \gamma)\eta_j) \right) \eta_i \|P_i\|_\infty \right) \mathbf{1} + P_t, \quad (52)$$

$$\Delta_t \geq - \prod_{k=0}^{t-1} (1 - (1 - \gamma)\eta_k) \|\Delta_0\|_\infty \mathbf{1} - \gamma \eta_{t-1} \|P_{t-1}\|_\infty \mathbf{1} - \gamma \sum_{i=1}^{t-2} \left(\left(\prod_{j=i+1}^{t-1} (1 - (1 - \gamma)\eta_j) \right) \eta_i \|P_i\|_\infty \right) \mathbf{1} + P_t, \quad (53)$$

where P_t represents a discounted sum of the random error \mathcal{E}_t , defined as:

$$P_t = \begin{cases} \mathbf{0} & \text{if } t = 0, \\ \sum_{k=0}^{t-1} \left(\left(\prod_{j=k+1}^{t-1} (1 - \eta_j) \right) \eta_k \mathcal{E}_k \right) & \text{if } t \geq 1. \end{cases} \quad (54)$$

To effectively manage the product term $\prod_{k=0}^{t-1} (1 - (1 - \gamma)\eta_k)$, we apply a step size defined as $\eta_k = \frac{1}{1 + (1 - \gamma)(k + 1)}$. Under this choice of step size, we can demonstrate that the expression $1 - (1 - \gamma)\eta_k$ satisfies:

$$\begin{aligned} 1 - (1 - \gamma)\eta_k &= 1 - \frac{1 - \gamma}{1 + (1 - \gamma)(k + 1)} \\ &= \frac{1 + (1 - \gamma)k}{1 + (1 - \gamma)(k + 1)} \\ &= \frac{\eta_k}{\eta_{k-1}}, \quad \forall k \geq 1. \end{aligned}$$

Using this result, we can simplify the product-form coefficients in Eq. (52) as follows:

$$\prod_{k=0}^{t-1} (1 - (1 - \gamma)\eta_k) = (1 - (1 - \gamma)\eta_0) \prod_{k=1}^{t-1} \frac{\eta_k}{\eta_{k-1}} = \eta_{t-1}, \quad \forall t, \quad (55)$$

$$\left(\prod_{j=i+1}^{t-1} (1 - (1 - \gamma)\eta_j) \right) \eta_i = \left(\prod_{j=i+1}^{t-1} \frac{\eta_j}{\eta_{j-1}} \right) \eta_i = \eta_{t-1}, \quad \forall t. \quad (56)$$

Substituting Eq. (55) and (56) into the bounds in Eq. (52) and (53), we obtain:

$$\|\Delta_t\|_\infty \leq \eta_{t-1} \|\Delta_0\|_\infty + \gamma \eta_{t-1} \sum_{k=0}^{t-1} \|P_k\|_\infty + \|P_t\|_\infty. \quad (57)$$

To complete the proof, we need to control $\|P_k\|_\infty$ for each k . The following lemma provides the necessary bound:

Lemma D.2 (Proof in Appendix D.9). *Given the step size $\eta_k = \frac{1}{1+(1-\gamma)k}$, for any $k \geq 1$, with probability at least $1 - \delta$, the weighted error sum P_k satisfies:*

$$\|P_k\|_\infty \leq \frac{2}{3(1-\gamma)(1+(1-\gamma)k)} \log \left(\frac{2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right) + \frac{2\sqrt{2\|\sigma_\xi^2\|_\infty \log \left(\frac{2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}}{\sqrt{1+(1-\gamma)k}}, \quad (58)$$

where σ_ξ^2 is the variance of each random error \mathcal{E}_t , satisfying $\sigma_\xi^2(s, a) = \text{Var} \left(\widehat{\mathcal{H}}(Q_\omega^*)(s, a) \right)$.

This lemma shows that $\|P_k\|_\infty$ is decreasing in the iteration number k . Substituting the bound on $\|P_k\|_\infty$ into Eq. (57) yields the following results:

Lemma D.3 (Proof in Appendix D.10). *With probability at least $1 - \delta$, the estimation error Δ_M after M iterations in epoch $\tau = 1$ satisfies:*

$$\|\Delta_M\|_\infty \leq \frac{3 + 2 \log \left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}{3(1-\gamma)^2 M} + \frac{2 \log \left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right) \log(1+(1-\gamma)M)}{3(1-\gamma)^3 M} + \frac{6\sqrt{2\|\sigma_\xi^2\|_\infty \log \left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}}{(1-\gamma)^{\frac{3}{2}} M^{\frac{1}{2}}},$$

Let $M = c_2 \frac{\log \left(\frac{6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{(1-\gamma)\delta} \right)}{(1-\gamma)^3}$ with a sufficiently large c_2 , we have:

$$\|\bar{Q}_1 - Q_\omega^*\|_\infty = \|\Delta_M\|_\infty \leq \frac{\sqrt{\|\sigma_\xi^2\|_\infty + 1}}{2} \leq \frac{1}{2(1-\gamma)} \quad (59)$$

holds with probability at least $1 - \frac{\delta}{T}$, where the last inequality is due to $\|\sigma_\xi^2\|_\infty \leq \frac{\gamma^2}{(1-\gamma)^2}$. This finishes the proof of the basic case with $\tau = 1$.

Step 2. Showing the Inductive Case with $\tau \geq 2$:

We assume that the input \bar{Q}_τ in epoch τ satisfies the bound

$$\|\bar{Q}_\tau - Q_\omega^*\|_\infty \leq \frac{\sqrt{\|\sigma_\xi^2\|_\infty + 1}}{2^\tau} := b_\tau,$$

and our goal is to show that $\|\bar{Q}_{\tau+1} - Q_\omega^*\|_\infty \leq \frac{b_\tau}{2}$ with probability at least $1 - \frac{\delta}{T}$.

Specifically, $\bar{Q}_{\tau+1}$ is equivalent to the output Q_M of running M rounds of variance-reduced Q-learning from the initialization $Q_0 = \bar{Q}_\tau$. The reference Bellman operator $\bar{\mathcal{H}}_\tau$ is obtained using N_τ empirical samples.

The variance-reduced update can be rewritten into:

$$Q_{t+1} - Q^* = (1 - \eta_t)(Q_t - Q^*) + \underbrace{\eta_t (\hat{\mathcal{H}}_t(Q_t) - \hat{\mathcal{H}}_t(Q^*))}_{\text{Contractive Error: } \mathcal{W}_t(\Delta_t)} + \underbrace{\eta_t (\hat{\mathcal{H}}_t(Q^*) - \mathcal{H}(Q_\omega^*) - \hat{\mathcal{H}}_t(\bar{Q}_\tau) + \bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau))}_{\text{Variance-reduced Error: } \bar{\mathcal{E}}_t}.$$

Observe that, the update form contains three components: the decaying term $(1 - \eta_t)(Q_t - Q^*)$, the contractive error term $\mathcal{W}_t(\Delta_t)$ and the variance reduced error term $\bar{\mathcal{E}}_t$. The only difference between the variance-reduced iteration and the vanilla iteration comes from the variance-reduced error $\bar{\mathcal{E}}_t$, which can be further split into the following form:

$$\begin{aligned} & \hat{\mathcal{H}}_t(Q^*) - \mathcal{H}(Q_\omega^*) - \hat{\mathcal{H}}_t(\bar{Q}_\tau) + \bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau) \\ = & \hat{\mathcal{H}}_t(Q^*) - \hat{\mathcal{H}}_t(\bar{Q}_\tau) + \bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau) - \bar{\mathcal{H}}_{\tau+1}(Q_\omega^*) + \bar{\mathcal{H}}_{\tau+1}(Q_\omega^*) - \mathcal{H}(Q_\omega^*) \\ = & \underbrace{\mathcal{H}(\bar{Q}_\tau) - \mathcal{H}(Q_\omega^*) - \hat{\mathcal{H}}_t(\bar{Q}_\tau) + \hat{\mathcal{H}}_t(Q_\omega^*)}_{\bar{\mathcal{E}}_t^a} + \underbrace{\bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau) - \bar{\mathcal{H}}_{\tau+1}(Q_\omega^*) - \mathcal{H}(\bar{Q}_\tau) + \mathcal{H}(Q_\omega^*)}_{\bar{\mathcal{E}}^b} + \underbrace{\bar{\mathcal{H}}_{\tau+1}(Q_\omega^*) - \mathcal{H}(Q_\omega^*)}_{\bar{\mathcal{E}}^c} \end{aligned}$$

An important observation is that, $\bar{\mathcal{E}}_t^a$ depends on the empirical Bellman operator sampled in each iteration t , while $\bar{\mathcal{E}}^b$ and $\bar{\mathcal{E}}^c$ are independent of each iteration, but only dependent on the reference Bellman operator \bar{T}_τ sampled in the beginning of iteration.

We apply the result in Lemma D.1 for the bound on the error accumulation in iterative updates, and get the following bound on $\Delta_t = Q_t - Q^*$:

$$\|\Delta_t\|_\infty \leq \eta_{t-1}\|\Delta_0\|_\infty + \gamma\eta_{t-1} \sum_{k=0}^{t-1} (\|P_k^a\|_\infty + \|P_k^b\|_\infty + \|P_k^c\|_\infty) + \|P_t^a\|_\infty + \|P_t^b\|_\infty + \|P_t^c\|_\infty, \quad (60)$$

where P_t^a , P_t^b and P_t^c are the discounted sums of $\bar{\mathcal{E}}_t^a$, $\bar{\mathcal{E}}^b$ and $\bar{\mathcal{E}}^c$ satisfying:

$$P_t^a = \begin{cases} \mathbf{0} & \text{if } t = 0, \\ \sum_{k=0}^{t-1} \left(\left(\prod_{j=k+1}^{t-1} (1 - \eta_j) \right) \eta_k \bar{\mathcal{E}}_k^a \right) & \text{if } t \geq 1. \end{cases} \quad (61)$$

$$P_t^b = \begin{cases} \mathbf{0} & \text{if } t = 0, \\ \sum_{k=0}^{t-1} \left(\left(\prod_{j=k+1}^{t-1} (1 - \eta_j) \right) \eta_k \bar{\mathcal{E}}^b \right) & \text{if } t \geq 1. \end{cases} \quad (62)$$

$$P_t^c = \begin{cases} \mathbf{0} & \text{if } t = 0, \\ \sum_{k=0}^{t-1} \left(\left(\prod_{j=k+1}^{t-1} (1 - \eta_j) \right) \eta_k \bar{\mathcal{E}}^c \right) & \text{if } t \geq 1. \end{cases} \quad (63)$$

To simplify, we have that $\|P_t^b\|_\infty \leq \|\bar{\mathcal{E}}^b\|_\infty$ and $\|P_t^c\|_\infty \leq \|\bar{\mathcal{E}}^c\|_\infty$. Since $\|\bar{\mathcal{E}}^b\|_\infty$ and $\|\bar{\mathcal{E}}^c\|_\infty$ are independent of t , by substituting $\eta_{t-1} = \frac{1}{1+(1-\gamma)t}$, we have

$$\begin{aligned} & \gamma\eta_{t-1} \sum_{k=0}^{t-1} (\|\bar{\mathcal{E}}^b\|_\infty + \|\bar{\mathcal{E}}^c\|_\infty) + \|\bar{\mathcal{E}}^b\|_\infty + \|\bar{\mathcal{E}}^c\|_\infty \\ = & \left(1 + \frac{\gamma t}{1 + (1 - \gamma)t} \right) (\|\bar{\mathcal{E}}^b\|_\infty + \|\bar{\mathcal{E}}^c\|_\infty) \\ \leq & \frac{2}{1 - \gamma} (\|\bar{\mathcal{E}}^b\|_\infty + \|\bar{\mathcal{E}}^c\|_\infty). \end{aligned}$$

Hence, Δ_t satisfies:

$$\|\Delta_t\|_\infty \leq \eta_{t-1}\|\Delta_0\|_\infty + \gamma\eta_{t-1} \sum_{k=0}^{t-1} \|P_k^a\|_\infty + \|P_t^a\|_\infty + \frac{2}{1 - \gamma} (\|\bar{\mathcal{E}}^b\|_\infty + \|\bar{\mathcal{E}}^c\|_\infty). \quad (64)$$

The remaining challenge is to bound the following three terms: $\|\bar{\mathcal{E}}^b\|$, $\|\bar{\mathcal{E}}^c\|$ and $\|P_k^a\|$.

Lemma D.4 (Proof in Appendix D.11). In τ -th iteration, with probability at least $1 - \frac{\delta}{3T}$, $\bar{\mathcal{E}}^b$ satisfies:

$$\|\bar{\mathcal{E}}^b\|_\infty \leq b_\tau \sqrt{\frac{2 \log \left(\frac{6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}{N_\tau}}.$$

Lemma D.5 (Proof in Appendix D.12). In τ -th iteration, with probability at least $1 - \frac{\delta}{3T}$, $\bar{\mathcal{E}}^c$ satisfies:

$$\|\bar{\mathcal{E}}^c\|_\infty \leq \bar{c}_1 \sqrt{\frac{\log \left(\frac{2T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}{N_\tau}} \left(\sqrt{\|\sigma_{\mathcal{E}}^2\|_\infty} + 1 \right),$$

where \bar{c}_1 is an absolute constant.

Lemma D.6 (Proof in Appendix D.13). In the τ -th iteration, with probability at least $1 - \frac{\delta}{3T}$, it satisfies:

$$\gamma \eta_{t-1} \sum_{k=0}^{t-1} \|P_k^a\|_\infty + \|P_t^a\|_\infty \leq \bar{c}_2 b_\tau \left(\frac{\log \left(\frac{6TM|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right) \log(1 + (1 - \gamma)t)}{(1 - \gamma)^2 t} + \frac{\sqrt{\log \left(\frac{6TM|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}}{(1 - \gamma)^{\frac{3}{2}} t^{\frac{1}{2}}} \right),$$

where \bar{c}_2 is an absolute constant.

Combining Lemmas D.4, D.5 and D.6 yields:

$$\begin{aligned} \|\Delta_M\|_\infty &\leq \frac{b_\tau}{1 + (1 - \gamma)M} + \frac{\bar{c}_1 \sqrt{\frac{\log \left(\frac{2T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}{N_\tau}} \left(\sqrt{\|\sigma_{\mathcal{E}}^2\|_\infty} + 1 \right) + b_\tau \sqrt{\frac{\log \left(\frac{2T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}{N_\tau}}}{1 - \gamma} \\ &\quad + \bar{c}_2 b_\tau \left(\frac{\log \left(\frac{6TM|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right) \log(1 + (1 - \gamma)M)}{(1 - \gamma)^2 M} + \frac{\sqrt{\log \left(\frac{6TM|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}}{(1 - \gamma)^{\frac{3}{2}} M^{\frac{1}{2}}} \right). \end{aligned}$$

Substituting $M = c_2 \frac{\log \left(\frac{6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{(1 - \gamma)\delta} \right)}{(1 - \gamma)^3}$ and $N_\tau = c_3 4^\tau \frac{\log(6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)}{(1 - \gamma)^2}$ with large enough c_2 and c_3 yields that:

$$\|\Delta_M\|_\infty \leq \frac{\sqrt{\|\sigma_{\mathcal{E}}^2\|_\infty} + 1}{2^{\tau+1}} \leq \frac{1}{(1 - \gamma)2^{\tau+1}}.$$

This concludes our proof.

D.8 Proof to Lemma D.1

Proof. We will prove the result by induction. First, by applying Eq. (50), we have that Δ_1 satisfies:

$$\Delta_1 = (1 - \eta_0)\Delta_0 + \eta_0(\widehat{\mathcal{H}}_0(Q_\omega^* + \Delta_0) - \widehat{\mathcal{H}}_0(Q_\omega^*)) + \eta_0(\widehat{\mathcal{H}}_0(Q_\omega^*) - \mathcal{H}(Q_\omega^*))$$

$$\begin{aligned}
&\leq (1 - \eta_0)\|\Delta_0\|_\infty \mathbf{1} + \eta_0\gamma\|\Delta_0\|_\infty \mathbf{1} + \eta_0\mathcal{E}_0 \\
&= (1 - (1 - \gamma)\eta_0)\|\Delta_0\|_\infty \mathbf{1} + \eta_0\mathcal{E}_0 \\
&= (1 - (1 - \gamma)\eta_0)\|\Delta_0\|_\infty \mathbf{1} + P_1,
\end{aligned} \tag{65}$$

where the inequality follows from the contraction property $\widehat{\mathcal{H}}_0(Q_\omega^* + \Delta_0) - \widehat{\mathcal{H}}_0(Q_\omega^*) \leq \gamma\|\Delta_0\|_\infty \mathbf{1}$.

Similarly, Δ_1 can be lower bounded as follows:

$$\begin{aligned}
\Delta_1 &= (1 - \eta_0)\Delta_0 + \eta_0(\widehat{\mathcal{H}}_0(Q_\omega^* + \Delta_0) - \widehat{\mathcal{H}}_0(Q_\omega^*)) + \eta_0(\widehat{\mathcal{H}}_0(Q_\omega^*) - \mathcal{H}(Q_\omega^*)) \\
&\geq -(1 - \eta_0)\|\Delta_0\|_\infty \mathbf{1} - \eta_0\gamma\|\Delta_0\|_\infty \mathbf{1} + \eta_0\mathcal{E}_0 \\
&= -(1 - (1 - \gamma)\eta_0)\|\Delta_0\|_\infty \mathbf{1} + \eta_0\mathcal{E}_0 \\
&= -(1 - (1 - \gamma)\eta_0)\|\Delta_0\|_\infty \mathbf{1} + P_1.
\end{aligned} \tag{66}$$

The expressions in Eq. (65) and (66) are consistent with the upper and lower bounds provided by Eq. (52) and (53) in Lemma D.1.

Now, suppose that the conditions in Eq. (52) and (53) hold for Δ_t , that is:

$$\Delta_t \leq a_t\|\Delta_0\|_\infty \mathbf{1} + b_t \mathbf{1} + P_t, \tag{67}$$

$$\Delta_t \geq -a_t\|\Delta_0\|_\infty \mathbf{1} - b_t \mathbf{1} + P_t, \tag{68}$$

where

$$a_t = \prod_{k=0}^{t-1} (1 - (1 - \gamma)\eta_k), \quad b_t = \gamma\eta_{t-1}\|P_{t-1}\|_\infty + \gamma \sum_{i=1}^{t-2} \left(\left(\prod_{j=i+1}^{t-1} (1 - (1 - \gamma)\eta_j) \right) \eta_i \|P_i\|_\infty \right).$$

Applying Eq. (50), Δ_{t+1} satisfies:

$$\begin{aligned}
\Delta_{t+1} &= (1 - \eta_t)\Delta_t + \eta_t\mathcal{W}_t(\Delta_t) + \eta_t\mathcal{E}_t \\
&\leq (1 - \eta_t)(a_t\|\Delta_0\|_\infty \mathbf{1} + b_t \mathbf{1} + P_t) + \eta_t\gamma\|\Delta_t\|_\infty \mathbf{1} + \eta_t\mathcal{E}_t
\end{aligned} \tag{69}$$

$$\leq (1 - \eta_t)(a_t\|\Delta_0\|_\infty \mathbf{1} + b_t \mathbf{1} + P_t) + \eta_t\gamma(a_t\|\Delta_0\|_\infty + b_t + \|P_t\|_\infty) \mathbf{1} + \eta_t\mathcal{E}_t \tag{70}$$

$$= \underbrace{(1 - (1 - \gamma)\eta_t)a_t}_{a_{t+1}} \|\Delta_0\|_\infty \mathbf{1} + \underbrace{(\gamma\eta_t\|P_t\|_\infty + (1 - (1 - \gamma)\eta_t)b_t)}_{b_{t+1}} \mathbf{1} + \underbrace{(1 - \eta_t)P_t + \eta_t\mathcal{E}_t}_{P_{t+1}}, \tag{71}$$

where inequality (69) is obtained by applying the upper bound condition (67) and the contraction property $\mathcal{W}_t(\Delta_t) \leq \gamma\|\Delta_t\|_\infty \mathbf{1}$; Inequality (70) is obtained by using conditions (67); Equality (71) is achieved by rearranging the terms.

Following the same steps, we can establish the corresponding lower bound for Δ_{t+1} in Eq. (53). This completes the proof. \square

D.9 Proof to Lemma D.2

Proof. Recall the definition in Eq. (54). For any $t \geq 2$, P_t can be expressed as a weighted sum:

$$P_t = \sum_{k=0}^{t-1} a_{t,k} \mathcal{E}_k,$$

where $a_{t,k} = \left(\prod_{j=k+1}^{t-1} (1 - \eta_j) \right) \eta_k$.

Denote the (s, a) -th entry of P_t as $P_t(s, a)$. Since all \mathcal{E}_k are independent and identically distributed, we can apply Bernstein's inequality (Vershynin, 2018) to $P_t(s, a)$. This gives us the following probability bound:

$$\mathbb{P}(|P_t(s, a)| \geq \epsilon_t) \leq 2 \exp\left(-\frac{\epsilon_t^2}{2 \sum_{k=0}^{t-1} a_{t,k}^2 \sigma_{\mathcal{E}}^2(s, a) + \frac{2}{3}(\max_k |a_{t,k}| M_{\mathcal{E}(s,a)}) \epsilon_t}\right), \quad \forall (s, a),$$

where $\sigma_{\mathcal{E}}^2(s, a)$ is the variance of $\mathcal{E}(s, a)$, and $M_{\mathcal{E}(s,a)}$ is the maximum absolute value of $\mathcal{E}(s, a)$.

Recall that P_t is the weighted sum of different \mathcal{E}_k 's, and all \mathcal{E}_k are *i.i.d.* random variables. Each \mathcal{E}_k has at most $|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|$ distinct entries due to the sampling scheme, where $\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]$ represents the set of possible state-action pairs involved in the components' scopes. We denote these distinct state-action pairs by \mathcal{X}^* defined in Eq. (44).

Taking the union bound over all state-action pairs $x := (s, a) \in \mathcal{X}^*$, we have:

$$\mathbb{P}(\|P_t\|_\infty \geq \epsilon_t) \leq 2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \exp\left(-\frac{\epsilon_t^2}{2 \sum_{k=0}^{t-1} a_{t,k}^2 \sigma_{\mathcal{E}}^2(s, a) + \frac{2}{3}(\max_k |a_{t,k}| M_{\mathcal{E}(s,a)}) \epsilon_t}\right).$$

To set this probability to δ , let $\mathbb{P}(\|P_t\|_\infty \geq \epsilon_t) = \delta$. Then, we can conclude:

$$\delta \leq 2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]| \exp\left(-\frac{\epsilon_t^2}{2 \sum_{k=0}^{t-1} a_{t,k}^2 \sigma_{\mathcal{E}}^2(s, a) + \frac{2}{3}(\max_k |a_{t,k}| M_{\mathcal{E}(s,a)}) \epsilon_t}\right).$$

Taking the natural logarithm on both sides and rearranging terms gives:

$$\epsilon_t^2 - \log\left(\frac{2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right) \left(2 \sum_{k=0}^{t-1} a_{t,k}^2 \sigma_{\mathcal{E}}^2(s, a) + \frac{2}{3} \left(\max_k |a_{t,k}| M_{\mathcal{E}(s,a)}\right) \epsilon_t\right) \leq 0.$$

To solve for ϵ_t , we treat this as a quadratic inequality:

$$\epsilon_t^2 - \left(\frac{2}{3} \max_k |a_{t,k}| M_{\mathcal{E}(s,a)}\right) \log\left(\frac{2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right) \epsilon_t - 2 \log\left(\frac{2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right) \sum_{k=0}^{t-1} a_{t,k}^2 \sigma_{\mathcal{E}}^2(s, a) \leq 0.$$

To simplify further:

$$\epsilon_t \leq \frac{2}{3} \log\left(\frac{2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right) \max_k |a_{t,k}| M_{\mathcal{E}(s,a)} + 2 \sqrt{\log\left(\frac{2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right) \sum_{k=0}^{t-1} a_{t,k}^2 \sigma_{\mathcal{E}}^2(s, a)}. \quad (72)$$

To complete the analysis, we need to bound $\max_k |a_{t,k}|$, $M_{\mathcal{E}(s,a)}$, and $\sum_{k=0}^{t-1} a_{t,k}^2$.

Step 1: Bounding $M_{\mathcal{E}(s,a)}$

For any k , since the empirical Bellman operator $\hat{\mathcal{H}}_k(Q^*)$ and the true Bellman operator $\mathcal{H}(Q_\omega^*)$ are both bounded between 0 and $\frac{1}{1-\gamma}$, we have:

$$M_{\mathcal{E}(s,a)} = \sup \|\mathcal{E}_k\|_\infty = \sup \|\hat{\mathcal{H}}_k(Q^*) - \mathcal{H}(Q_\omega^*)\|_\infty \leq \frac{1}{1-\gamma}.$$

Step 2: Bounding $\max_k |a_{t,k}|$

Next, we analyze the ratio $\frac{a_{t,k}}{a_{t,k-1}}$ to find $\max_k |a_{t,k}|$. Recall the definition of $a_{t,k}$:

$$a_{t,k} = \left(\prod_{j=k+1}^{t-1} (1 - \eta_j)\right) \eta_k.$$

To compute $\frac{a_{t,k}}{a_{t,k-1}}$, we have:

$$\frac{a_{t,k}}{a_{t,k-1}} = \frac{\eta_k}{\eta_{k-1}(1-\eta_k)}.$$

Using the step size $\eta_k = \frac{1}{1+(1-\gamma)(k+1)}$, we can express:

$$\frac{a_{t,k}}{a_{t,k-1}} = \frac{1 - (1-\gamma)\eta_k}{1 - \eta_k}.$$

Given that $0 < \gamma < 1$, we have $\eta_k \leq \frac{1}{1+(1-\gamma)k} < 1$, and thus:

$$\frac{a_{t,k}}{a_{t,k-1}} \geq 1, \quad \forall k.$$

This means that $a_{t,k}$ is non-decreasing with respect to k . Hence, the maximum value $\max_k |a_{t,k}|$ occurs at $k = t - 1$:

$$\max_k |a_{t,k}| = a_{t,t-1} = \eta_{t-1} = \frac{1}{1 + (1-\gamma)t}.$$

Step 3: Bounding $\sum_{k=0}^{t-1} a_{t,k}^2$

We also need to bound $\sum_{k=0}^{t-1} a_{t,k}^2$. The following lemma provides this bound:

Lemma D.7 (Proof in Appendix D.14). *For any $t \geq 1$, the sum $\sum_{k=0}^{t-1} a_{t,k}^2$ satisfies:*

$$\sum_{k=0}^{t-1} a_{t,k}^2 \leq \frac{2}{1 + (1-\gamma)t}.$$

Combining the results from Steps 1, 2, and 3, we can now derive the bound for ϵ_t . Substituting these bounds into Eq. (72), we get that with probability at least $1 - \delta$:

$$\|P_t\|_\infty \leq \epsilon_t \leq \frac{2}{3(1-\gamma)(1+(1-\gamma)t)} \log \left(\frac{2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right) + \frac{2\sqrt{2\|\sigma_\xi^2\|_\infty \log \left(\frac{2|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}}{\sqrt{1+(1-\gamma)t}}.$$

This concludes our proof. □

D.10 Proof of Lemma D.3

Proof. Letting the upper bound of each $\|P_k\|_\infty$ (in Eq. (58)) with $k = 1, \dots, M$ holds with probability $1 - \frac{\delta}{M}$, we have:

$$\begin{aligned} \|\Delta_M\|_\infty &\leq \frac{\|\Delta_0\|_\infty}{1+(1-\gamma)M} + \frac{2 \log \left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}{3(1-\gamma)^2 M} + \frac{2\sqrt{2\|\sigma_\xi^2\|_\infty \log \left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}}{\sqrt{1+(1-\gamma)M}} \\ &\quad + \frac{2 \log \left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}{3(1-\gamma)^2 M} \sum_{i=1}^M \frac{1}{1+(1-\gamma)i} \end{aligned}$$

$$+ \frac{2\sqrt{2\|\sigma_{\mathcal{E}}^2\|_{\infty} \log\left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|}{\delta}\right)}}{1 + (1 - \gamma)M} \sum_{i=1}^M \frac{1}{\sqrt{1 + (1 - \gamma)i}}.$$

Note that,

$$\sum_{i=1}^M \frac{1}{1 + (1 - \gamma)i} \leq \int_0^M \frac{1}{1 + (1 - \gamma)x} dx = \frac{\log(1 + (1 - \gamma)M)}{1 - \gamma},$$

$$\sum_{i=1}^M \frac{1}{\sqrt{1 + (1 - \gamma)i}} \leq \int_0^M \frac{1}{\sqrt{1 + (1 - \gamma)x}} dx = \frac{2(\sqrt{1 + (1 - \gamma)M} - 1)}{1 - \gamma} \leq \frac{2\sqrt{1 + (1 - \gamma)M}}{1 - \gamma}.$$

Hence, Δ_M satisfies:

$$\begin{aligned} \|\Delta_M\|_{\infty} &\leq \frac{\|\Delta_0\|_{\infty}}{1 + (1 - \gamma)M} + \frac{2 \log\left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|}{\delta}\right)}{3(1 - \gamma)^2 M} + \frac{2\sqrt{2\|\sigma_{\mathcal{E}}^2\|_{\infty} \log\left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|}{\delta}\right)}}{\sqrt{1 + (1 - \gamma)M}} \\ &\quad + \frac{2 \log\left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|}{\delta}\right) \log(1 + (1 - \gamma)M)}{3(1 - \gamma)^3 M} + \frac{4\sqrt{2\|\sigma_{\mathcal{E}}^2\|_{\infty} \log\left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|}{\delta}\right)}}{(1 - \gamma)^{\frac{3}{2}} M^{\frac{1}{2}}} \\ &\leq \frac{3 + 2 \log\left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|}{\delta}\right)}{3(1 - \gamma)^2 M} + \frac{2 \log\left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|}{\delta}\right) \log(1 + (1 - \gamma)M)}{3(1 - \gamma)^3 M} \\ &\quad + \frac{6\sqrt{2\|\sigma_{\mathcal{E}}^2\|_{\infty} \log\left(\frac{2M|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|}{\delta}\right)}}{(1 - \gamma)^{\frac{3}{2}} M^{\frac{1}{2}}}. \end{aligned}$$

This concludes our proof. □

D.11 Proof of Lemma D.4

Proof. Recall the definition of $\bar{\mathcal{E}}^b$, which is given by:

$$\bar{\mathcal{E}}^b = \bar{\mathcal{H}}_{\tau+1}(\bar{Q}_{\tau}) - \bar{\mathcal{H}}_{\tau+1}(Q_{\omega}^*) - \mathcal{H}(\bar{Q}_{\tau}) + \mathcal{H}(Q_{\omega}^*).$$

We first note that the expectation of the difference between the empirical Bellman operators is:

$$\mathbb{E}_{\bar{\mathcal{H}}_{\tau+1}} [\bar{\mathcal{H}}_{\tau+1}(\bar{Q}_{\tau}) - \bar{\mathcal{H}}_{\tau+1}(Q_{\omega}^*)] = \mathcal{H}(\bar{Q}_{\tau}) - \mathcal{H}(Q_{\omega}^*).$$

Due to the γ -contraction property of both the empirical Bellman operator $\bar{\mathcal{H}}_{\tau+1}$ and the true Bellman operator \mathcal{H} , we have:

$$\|\bar{\mathcal{H}}_{\tau+1}(\bar{Q}_{\tau}) - \bar{\mathcal{H}}_{\tau+1}(Q_{\omega}^*) - \mathcal{H}(\bar{Q}_{\tau}) + \mathcal{H}(Q_{\omega}^*)\|_{\infty} \leq \|\bar{\mathcal{H}}_{\tau+1}(\bar{Q}_{\tau}) - \bar{\mathcal{H}}_{\tau+1}(Q_{\omega}^*)\|_{\infty} + \|\mathcal{H}(\bar{Q}_{\tau}) - \mathcal{H}(Q_{\omega}^*)\|_{\infty}.$$

Since both operators are γ -contractions, it follows that:

$$\|\bar{\mathcal{H}}_{\tau+1}(\bar{Q}_{\tau}) - \bar{\mathcal{H}}_{\tau+1}(Q_{\omega}^*)\|_{\infty} \leq \gamma \|\bar{Q}_{\tau} - Q_{\omega}^*\|_{\infty} \quad \text{and} \quad \|\mathcal{H}(\bar{Q}_{\tau}) - \mathcal{H}(Q_{\omega}^*)\|_{\infty} \leq \gamma \|\bar{Q}_{\tau} - Q_{\omega}^*\|_{\infty}.$$

Combining these, we get:

$$\|\bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau) - \bar{\mathcal{H}}_{\tau+1}(Q_\omega^*) - \mathcal{H}(\bar{Q}_\tau) + \mathcal{H}(Q_\omega^*)\|_\infty \leq 2\gamma\|\bar{Q}_\tau - Q_\omega^*\|_\infty \leq 2\gamma b_\tau.$$

Applying Hoeffding's inequality to the (s, a) -th entry of $\bar{\mathcal{E}}^b$, denoted as $\bar{\mathcal{E}}^b(s, a)$, we obtain:

$$\mathbb{P}\left(|\bar{\mathcal{E}}^b(s, a)| \geq \epsilon\right) \leq 2 \exp\left(-\frac{2N_\tau\epsilon^2}{4b_\tau^2}\right) = 2 \exp\left(-\frac{N_\tau\epsilon^2}{2b_\tau^2}\right).$$

Letting $2 \exp\left(-\frac{N_\tau\epsilon^2}{2b_\tau^2}\right) = \frac{\delta}{3T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}$, and taking union bound for $x \in \mathcal{X}^*$, we have that:

$$\|\bar{\mathcal{E}}^b\|_\infty \leq b_\tau \sqrt{\frac{2 \log\left(\frac{6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right)}{N_\tau}}.$$

This completes the proof. \square

D.12 Proof of Lemma D.5

Proof. Recall that $\bar{E}^c = \bar{T}(Q^*) - \mathcal{H}(Q_\omega^*)$ represents the estimation error using N_τ empirical Bellman operators. We know that $\mathbb{E}[\bar{T}(Q^*)] = \mathcal{H}(Q_\omega^*)$.

Applying Bernstein's inequality (Vershynin, 2018) to the (s, a) -th entry of \bar{E}^c , denoted as $\bar{\mathcal{E}}^c_{(s,a)}$, we obtain:

$$\mathbb{P}\left(|\bar{\mathcal{E}}^c_{(s,a)}| \geq \epsilon\right) \leq 2 \exp\left(-\frac{\frac{1}{2}N_\tau\epsilon^2}{\|\sigma_{\bar{\mathcal{E}}^c}^2\|_\infty + \frac{\epsilon}{3(1-\gamma)}}\right).$$

We set the right-hand side probability equal to $\frac{\delta}{3T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}$, which gives us:

$$2 \exp\left(-\frac{\frac{1}{2}N_\tau\epsilon^2}{\|\sigma_{\bar{\mathcal{E}}^c}^2\|_\infty + \frac{\epsilon}{3(1-\gamma)}}\right) = \frac{\delta}{3T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}.$$

Rearranging this equation for ϵ , we get:

$$\frac{1}{2}N_\tau\epsilon^2 - \frac{\epsilon \log(6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{3(1-\gamma)} - \|\sigma_{\bar{\mathcal{E}}^c}^2\|_\infty \log(6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta) = 0.$$

Solving this quadratic equation for ϵ , we have:

$$\epsilon \leq \frac{\frac{2 \log(6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{3(1-\gamma)} + \sqrt{\frac{4 \log^2(6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{9(1-\gamma)^2} + 2N_\tau\|\sigma_{\bar{\mathcal{E}}^c}^2\|_\infty \log(6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}}{N_\tau}.$$

Simplifying further, we find:

$$\epsilon \leq \frac{2 \log(6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{3N_\tau(1-\gamma)} + \sqrt{\frac{2\|\sigma_{\bar{\mathcal{E}}^c}^2\|_\infty \log(6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{N_\tau}}.$$

Applying the union bound over all $x := (s, a) \in \mathcal{X}^*$, we conclude that with probability at least $1 - \frac{\delta}{3T}$, the infinity norm $\|\bar{\mathcal{E}}^c\|_\infty$ satisfies:

$$\|\bar{\mathcal{E}}^c\|_\infty \leq \frac{2 \log(6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{3N_\tau(1-\gamma)} + \sqrt{\frac{2\|\sigma_{\bar{\mathcal{E}}^c}^2\|_\infty \log(6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{N_\tau}}. \quad (73)$$

Given that $N_\tau \geq c \cdot 4^\tau \frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{(1-\gamma)^2}$, it follows that $\frac{\log(2T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{N_\tau(1-\gamma)} \leq 1$ for a large enough c . Thus, $\|\bar{\mathcal{E}}^c\|_\infty$ is further bounded by:

$$\begin{aligned} \|\bar{\mathcal{E}}^c\|_\infty &\leq \bar{c}_1 \left(\sqrt{\frac{\|\sigma_{\bar{\mathcal{E}}^c}^2\|_\infty \log\left(\frac{6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right)}{N_\tau}} + 1 \right) \\ &\leq \bar{c}_1 \sqrt{\frac{\log\left(\frac{6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right)}{N_\tau}} \left(\sqrt{\|\sigma_{\bar{\mathcal{E}}^c}^2\|_\infty} + 1 \right), \end{aligned}$$

where \bar{c}_1 is an absolute constant. This completes the proof. \square

D.13 Proof of Lemma D.6

Proof. The term P_k^a evolves in an exactly the same way as P_k discussed before. Hence, we apply the result in Lemma D.2 and get that, with probability at least $1 - \frac{\delta}{3TM}$:

$$\|P_t^a\|_\infty \leq \frac{2M_{\mathcal{E},a}}{3(1+(1-\gamma)t)} \log\left(\frac{6TM|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right) + \frac{2\sqrt{2\|\sigma_{\bar{\mathcal{E}},a}^2\|_\infty \log\left(\frac{6TM|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right)}}{\sqrt{1+(1-\gamma)t}}, \forall t.$$

Here $M_{\mathcal{E},a}$ and $\sigma_{\bar{\mathcal{E}},a}^2$ denote the maximal absolute value and variance of error $\bar{\mathcal{E}}_t^a$.

First, we observe that $\bar{\mathcal{E}}_t^a$ satisfies:

$$\begin{aligned} \|\bar{\mathcal{E}}_t^a\|_\infty &= \|\mathcal{H}(\bar{Q}_\tau) - \mathcal{H}(Q_\omega^*) - \hat{\mathcal{H}}_t(\bar{Q}_\tau) + \hat{\mathcal{H}}_t(Q_\omega^*)\|_\infty \\ &\leq \|\mathcal{H}(\bar{Q}_\tau) - \mathcal{H}(Q_\omega^*)\|_\infty + \|\hat{\mathcal{H}}_t(\bar{Q}_\tau) - \hat{\mathcal{H}}_t(Q_\omega^*)\|_\infty \\ &\leq 2\gamma\|\bar{Q}_\tau - Q_\omega^*\|_\infty \\ &\leq 2b_\tau. \end{aligned}$$

Thus, we have $M_{\mathcal{E},a} \leq 2b_\tau$ and $\sqrt{\|\sigma_{\bar{\mathcal{E}},a}^2\|_\infty} \leq 2b_\tau$. And $\|P_t^a\|_\infty$ then satisfies:

$$\|P_t^a\|_\infty \leq b_\tau \left(\frac{4 \log\left(\frac{6TM|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right)}{3(1+(1-\gamma)t)} + \sqrt{\frac{8 \log\left(\frac{6TM|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right)}{1+(1-\gamma)t}} \right)$$

Thus, we can control the whole term by:

$$\gamma\eta_{t-1} \sum_{k=0}^{t-1} \|P_k^a\|_\infty + \|P_t^a\|_\infty \leq \gamma\eta_{t-1} \sum_{i=0}^{t-1} b_\tau \left(\frac{4 \log\left(\frac{6TM|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right)}{3(1+(1-\gamma)i)} + \sqrt{\frac{8 \log\left(\frac{6TM|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta}\right)}{1+(1-\gamma)i}} \right)$$

$$+ b_\tau \left(\frac{4 \log \left(\frac{6TM |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}{3(1 + (1 - \gamma)t)} + \sqrt{\frac{8 \log \left(\frac{6TM |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}{1 + (1 - \gamma)t}} \right).$$

For analyzing the coefficients:

$$\begin{aligned} \gamma \eta_{t-1} \sum_{i=0}^{t-1} \frac{1}{1 + (1 - \gamma)i} + \frac{1}{1 + (1 - \gamma)t} &\leq \frac{\gamma}{1 + (1 - \gamma)t} \int_0^t \frac{1}{1 + (1 - \gamma)x} dx + \frac{1}{1 + (1 - \gamma)t} \\ &\leq \frac{\gamma}{1 + (1 - \gamma)t} \cdot \frac{\log(1 + (1 - \gamma)t)}{1 - \gamma} + \frac{1}{1 + (1 - \gamma)t} \\ &\leq \frac{2 \log(1 + (1 - \gamma)t)}{(1 - \gamma)^2 t}. \end{aligned}$$

$$\begin{aligned} \gamma \eta_{t-1} \sum_{i=0}^{t-1} \frac{1}{\sqrt{1 + (1 - \gamma)i}} + \frac{1}{\sqrt{1 + (1 - \gamma)t}} &\leq \frac{\gamma}{1 + (1 - \gamma)t} \int_0^t \frac{1}{\sqrt{1 + (1 - \gamma)x}} dx + \frac{1}{\sqrt{1 + (1 - \gamma)t}} \\ &\leq \frac{\gamma}{1 + (1 - \gamma)t} \frac{2(\sqrt{1 + (1 - \gamma)t} - 1)}{1 - \gamma} + \frac{1}{\sqrt{1 + (1 - \gamma)t}} \\ &\leq \frac{3}{(1 - \gamma)^{\frac{3}{2}} t^{\frac{1}{2}}}. \end{aligned}$$

Thus, the whole term satisfies:

$$\gamma \eta_{t-1} \sum_{k=0}^{t-1} \|P_k^a\|_\infty + \|P_t^a\|_\infty \leq \frac{8b_\tau \log(1 + (1 - \gamma)t) \log \left(\frac{6TM |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}{3(1 - \gamma)^2 t} + \frac{6b_\tau \sqrt{2 \log \left(\frac{6TM |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right)}}{(1 - \gamma)^{\frac{3}{2}} t^{\frac{1}{2}}}. \quad (74)$$

Absorbing the constants concludes our proof. \square

D.14 Proof of Lemma D.7

Proof. Recall the expression for the ratio $\frac{a_{t,k-1}}{a_{t,k}}$:

$$\frac{a_{t,k-1}}{a_{t,k}} = \frac{1 - \eta_k}{1 - (1 - \gamma)\eta_k} = \frac{(1 - \gamma)(k + 1)}{(1 - \gamma)(k + 1) + \gamma} = \frac{k + 1}{k + 1 + \frac{\gamma}{1 - \gamma}}.$$

Using this, we can express $a_{t,k-i}$ for any $1 \leq i < k$ as follows:

$$a_{t,k-i} = a_{t,k} \prod_{j=0}^{i-1} \frac{a_{t,k-j-1}}{a_{t,k-j}} = a_{t,k} \prod_{j=0}^{i-1} \frac{k + 1 - j}{k + 1 - j + \frac{\gamma}{1 - \gamma}}.$$

To simplify this product, note that:

$$\frac{k + 1 - j}{k + 1 - j + \frac{\gamma}{1 - \gamma}} \leq \frac{k + 1}{k + 1 + \frac{\gamma}{1 - \gamma}} \quad \forall j.$$

Thus, we can further bound $a_{t,k-i}$ by:

$$a_{t,k-i} \leq a_{t,k} \prod_{j=0}^{i-1} \frac{k+1}{k+1 + \frac{\gamma}{1-\gamma}} = a_{t,k} \left(\frac{k+1}{k+1 + \frac{\gamma}{1-\gamma}} \right)^i.$$

Letting $k = t - 1$, we can now bound $\sum_{k=0}^{t-1} a_{t,k}^2$ by considering the following square form:

$$\sum_{k=0}^{t-1} a_{t,k}^2 \leq a_{t,t-1}^2 \left(1 + \sum_{k=1}^{t-1} \left(\frac{t}{t + \frac{\gamma}{1-\gamma}} \right)^{2k} \right).$$

We analyze this term by considering different γ .

Case 1: $\gamma \geq \frac{1}{2}$

This series is geometric with a ratio $\left(\frac{t}{t + \frac{\gamma}{1-\gamma}} \right)^2 < 1$. The sum of a geometric series can be calculated as:

$$\sum_{k=0}^{\infty} r^k = \frac{1}{1-r} \quad \text{for } |r| < 1.$$

Applying this to our series:

$$\sum_{k=0}^{t-1} a_{t,k}^2 \leq a_{t,t-1}^2 \left(1 + \frac{\left(\frac{t}{t + \frac{\gamma}{1-\gamma}} \right)^2}{1 - \left(\frac{t}{t + \frac{\gamma}{1-\gamma}} \right)^2} \right) = a_{t,t-1}^2 \left(1 + \frac{(1-\gamma)^2 t^2}{\gamma(2t(1-\gamma) + \gamma)} \right).$$

Since $\gamma \geq \frac{1}{2}$, we have:

$$a_{t,t-1}^2 \left(1 + \frac{(1-\gamma)^2 t^2}{\gamma(2t(1-\gamma) + \gamma)} \right) \leq a_{t,t-1}^2 \left(1 + \frac{(1-\gamma)^2 t^2}{t(1-\gamma)} \right) = a_{t,t-1}^2 (1 + (1-\gamma)t)$$

Using $a_{t,t-1} = \frac{1}{1+(1-\gamma)t}$ yields that,

$$\sum_{k=0}^{t-1} a_{t,k}^2 \leq \frac{1}{1 + (1-\gamma)t}. \quad (75)$$

Case 2: $\gamma < \frac{1}{2}$

We use $\frac{t}{t + \frac{\gamma}{1-\gamma}} \leq 1$ and $1 - \gamma \geq \frac{1}{2}$ to obtain that:

$$a_{t,t-1}^2 \left(1 + \sum_{k=1}^{t-2} \left(\frac{t}{t + \frac{\gamma}{1-\gamma}} \right)^{2k} \right) \leq a_{t,t-1}^2 \left(1 + \sum_{k=1}^{t-2} 1 \right) \leq a_{t,t-1}^2 (1 + t).$$

Since $\gamma < \frac{1}{2}$, we have $1 - \gamma > \frac{1}{2}$ and $1 < 2(1 - \gamma)$. Thus

$$a_{t,t-1}^2 (1 + t) \leq a_{t,t-1}^2 (1 + 2(1 - \gamma)t) \leq 2a_{t,t-1}^2 (1 + (1 - \gamma)t).$$

Using $a_{t,t-1} = \frac{1}{1+(1-\gamma)(t-1)}$ yields that,

$$a_{t,t-1}^2 \left(1 + \frac{(1-\gamma)^2 t^2}{\gamma(2(1-\gamma)t + \gamma)} \right) \leq \frac{2}{1 + (1-\gamma)t}. \quad (76)$$

Combing Eq. (75) and (76) yields our result. \square

D.15 Proof to Lemma C.2

Proof. We proceed by induction on the epoch index τ . The base case when $\tau = 0$ is straightforward, as the initial approximation \bar{Q}_0 can satisfy the desired bound. Our focus is on the inductive step: assuming that the input \bar{Q}_τ in epoch τ satisfies

$$\|\bar{Q}_\tau - Q_\omega^*\|_\infty \leq \frac{1}{2^\tau \sqrt{1-\gamma}} := b'_\tau,$$

we aim to show that the output $\bar{Q}_{\tau+1}$ satisfies $\|\bar{Q}_{\tau+1} - Q_\omega^*\|_\infty \leq \frac{b'_\tau}{2}$ with probability at least $1 - \frac{\delta}{T}$.

To achieve this, we analyze the variance-reduced update rule, which can be expressed as

$$Q_{t+1} = (1 - \eta_t)Q_t + \eta_t \hat{\mathcal{F}}_t(Q_t),$$

where $\hat{\mathcal{F}}_t(Q_t)$ is the variance-reduced Bellman operator defined by

$$\hat{\mathcal{F}}_t(Q_t) = \hat{\mathcal{H}}_t(Q_t) - \hat{\mathcal{H}}_t(\bar{Q}_\tau) + \bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau).$$

Here, $\hat{\mathcal{H}}_t$ is the empirical Bellman operator based on samples at iteration t , and $\bar{\mathcal{H}}_{\tau+1}$ is the reference Bellman operator estimated using a larger reference dataset collected in epoch $\tau + 1$.

We additionally define the refined variance-reduced Bellman operator, eliminating the randomness due to sampling:

$$\mathcal{F}(Q_t) = \mathcal{H}(Q_t) - \mathcal{H}(\bar{Q}_\tau) + \bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau),$$

where \mathcal{H} is the true Bellman operator. Due to independent and identically distributed sampling, we have $\mathbb{E}[\hat{\mathcal{F}}_t(Q_t)] = \mathcal{F}(Q_t)$. Moreover, since \mathcal{H} is a γ -contraction, it follows that \mathcal{F} is also a γ -contraction mapping.

We introduce a reference Q -function \tilde{Q} , defined as the fixed point of \mathcal{F} :

$$\tilde{Q} = \mathcal{F}(\tilde{Q}) = \mathcal{H}(\tilde{Q}) - \mathcal{H}(\bar{Q}_\tau) + \bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau).$$

This function \tilde{Q} serves as an intermediary between Q_M and Q_ω^* , potentially closer to Q_M than Q_ω^* is.

Our strategy is to bound the error $\|Q_M - Q_\omega^*\|_\infty$ by decomposing it into two components:

$$\|Q_M - Q_\omega^*\|_\infty \leq \underbrace{\|Q_M - \tilde{Q}\|_\infty}_{=: B_1} + \underbrace{\|\tilde{Q} - Q_\omega^*\|_\infty}_{=: B_2}.$$

The term B_1 quantifies the distance between the iteratively computed Q_M and the reference \tilde{Q} , while B_2 measures the distance between \tilde{Q} and the optimal Q_ω^* .

We aim to bound B_1 and B_2 separately. Intuitively, B_1 will decrease with the number of iterations M within the epoch, and B_2 will be controlled by the size of the reference dataset N_τ . We show they can be well bounded when N_τ and M are properly selected:

Lemma D.8 (Proof in Appendix D.16). *In the τ -th epoch, with probability at least $1 - \frac{\delta}{2T}$, B_1 satisfies:*

$$B_1 \leq \frac{b'_\tau + B_2}{5},$$

if $N_\tau = c'_1 \frac{4^\tau \log(8T |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)}{(1-\gamma)^2}$ and $M = c'_2 \frac{\log(6T |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)(1-\gamma)^{-1} \epsilon^{-1}}{(1-\gamma)^3}$ with large enough c'_1 and c'_2 .

Lemma D.9 (Proof in Appendix D.17). *In the τ -th epoch, with probability at least $1 - \frac{\delta}{2T}$, B_2 satisfies:*

$$B_2 \leq \frac{b'_\tau}{4}.$$

if $N_\tau = c'_1 \frac{4^\tau \log(8T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|)}{(1-\gamma)^2}$ and $M = c'_2 \frac{\log(6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|(1-\gamma)^{-1}\epsilon^{-1})}{(1-\gamma)^3}$ with large enough c'_1 and c'_2 .

By combining Lemmas D.8 and D.9, we obtain:

$$\|Q_M - Q_\omega^*\|_\infty \leq B_1 + B_2 \leq \frac{b'_\tau}{5} + \frac{6B_2}{5} = \frac{b'_\tau}{5} + \frac{6}{5} \cdot \frac{b'_\tau}{4} \leq \frac{b'_\tau}{2}.$$

This completes the inductive step and hence the proof of Lemma C.2. \square

D.16 Proof to Lemma D.8

Proof. We aim to bound $B_1 = \|Q_M - \tilde{Q}\|_\infty$. Our approach mirrors the inductive case in the proof of Lemma C.1, with a key difference in the error decomposition.

Consider the difference between the iterates Q_{t+1} and the reference \tilde{Q} :

$$Q_{t+1} - \tilde{Q} = (1 - \eta_t)Q_t + \eta_t \hat{\mathcal{F}}_t(Q_t) - \tilde{Q}.$$

Subtracting and adding $\eta_t \mathcal{F}_t(\tilde{Q})$, and using $\mathcal{F}(\tilde{Q}) = \tilde{Q}$, we can rewrite this as

$$Q_{t+1} - \tilde{Q} = (1 - \eta_t)(Q_t - \tilde{Q}) + \eta_t \left(\hat{\mathcal{F}}_t(Q_t) - \hat{\mathcal{F}}_t(\tilde{Q}) \right) + \eta_t \left(\hat{\mathcal{F}}_t(\tilde{Q}) - \mathcal{F}(\tilde{Q}) \right).$$

The first term represents the contraction towards \tilde{Q} . The second is γ -contractive because

$$\left\| \hat{\mathcal{F}}_t(Q_t) - \hat{\mathcal{F}}_t(\tilde{Q}) \right\|_\infty \leq \gamma \left\| Q_t - \tilde{Q} \right\|_\infty,$$

which is easy to handle, and the third term is an *i.i.d.* zero-mean random variable across iterations t , with the magnitude can be bounded by:

$$\left\| \hat{\mathcal{F}}_t(\tilde{Q}) - \mathcal{F}(\tilde{Q}) \right\|_\infty = \left\| \hat{T}_t(\tilde{Q}) - \hat{T}_t(\bar{Q}_\tau) \right\|_\infty + \left\| T_t(\tilde{Q}) - T_t(\bar{Q}_\tau) \right\|_\infty \leq 2\gamma \left\| \tilde{Q} - \bar{Q}_\tau \right\|_\infty.$$

Hence, the iteration of $Q_{t+1} - \tilde{Q}$ mirrors the iteration in Eq. (50). Following the same routine as the proof of Lemma D.3, we establish that, for a step sizes $\eta_t = \frac{1}{1+(1-\gamma)(t+1)}$, with probability $1 - \frac{\delta}{2T}$,

$$\|Q_M - \tilde{Q}\|_\infty \leq c \left(\frac{1}{(1-\gamma)M} + \frac{1}{(1-\gamma)^{3/2}\sqrt{M}} \right) \log \left(\frac{6TM|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{\delta} \right) \|\tilde{Q} - \bar{Q}_\tau\|_\infty,$$

where c is a large enough constant.

By choosing $M = c_2 \frac{\log \left(\frac{6T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|}{(1-\gamma)\delta} \right)}{(1-\gamma)^3}$ with sufficiently large c_2 , we can ensure that the right-hand side is less than $\frac{1}{5} \|\tilde{Q} - \bar{Q}_\tau\|_\infty$.

Observing that

$$\|\tilde{Q} - \bar{Q}_\tau\|_\infty \leq \|\tilde{Q} - Q_\omega^*\|_\infty + \|\bar{Q}_\tau - Q_\omega^*\|_\infty = B_2 + b'_\tau,$$

we can conclude

$$B_1 = \|Q_M - \tilde{Q}\|_\infty \leq \frac{1}{5} \|\tilde{Q} - \bar{Q}_\tau\|_\infty \leq \frac{1}{5} (B_2 + b'_\tau).$$

This establishes the desired bound on B_1 , completing the proof of Lemma D.8. \square

D.17 Proof to Lemma D.9

Proof. Our goal is to bound the term $B_2 = \|\tilde{Q} - Q_\omega^*\|_\infty$, where \tilde{Q} is the fixed point of the operator \mathcal{F} defined as:

$$\tilde{Q} = \mathcal{F}(\tilde{Q}) = \mathcal{H}(\tilde{Q}) - \mathcal{H}(\bar{Q}_\tau) + \bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau).$$

Recall that $Q_\omega^* = \mathcal{H}(Q_\omega^*)$. Thus, \tilde{Q} can be viewed as the fixed point of a Bellman operator with a perturbed reward function, where the perturbation magnitude is $-\mathcal{H}(\bar{Q}_\tau) + \bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau)$. We seek to show that when the population size for estimating the reference Bellman operator is large enough, the term $-\mathcal{H}(\bar{Q}_\tau) + \bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau)$ converges to 0 and \mathcal{F} converges to the actual Bellman operator \mathcal{H} , allowing $\|\tilde{Q} - Q_\omega^*\|_\infty \leq \frac{b'_\tau}{4}$. Essentially, the error B_2 comes from the reward function perturbation $\bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau) - \mathcal{H}(\bar{Q}_\tau)$. In the first step, we bound the reward perturbation $\Delta r = -\mathcal{H}(\bar{Q}_\tau) + \bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau)$ as follows:

$$\begin{aligned} |\Delta r| &= |\bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau) - \mathcal{H}(\bar{Q}_\tau)| \leq |\bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau) - \bar{\mathcal{H}}_{\tau+1}(Q_\omega^*) + \mathcal{H}(Q_\omega^*) - \mathcal{H}(Q_\omega^*) + \bar{\mathcal{H}}_{\tau+1}(Q_\omega^*) - \mathcal{H}(\bar{Q}_\tau)| \\ &\leq |\bar{\mathcal{H}}_{\tau+1}(\bar{Q}_\tau) - \bar{\mathcal{H}}_{\tau+1}(Q_\omega^*)| + |\mathcal{H}(\bar{Q}_\tau) - \mathcal{H}(Q_\omega^*)| + |\bar{\mathcal{H}}_{\tau+1}(Q_\omega^*) - \mathcal{H}(Q_\omega^*)| \\ &\leq 2\gamma\|\bar{Q}_\tau - Q_\omega^*\|_\infty \cdot \mathbf{1} + |\bar{\mathcal{H}}_{\tau+1}(Q_\omega^*) - \mathcal{H}(Q_\omega^*)|. \end{aligned}$$

The first terms can be directly bounded through Hoeffding's inequality. Similar to the proof of Lemma D.4, with probability at least $1 - \frac{\delta}{4T}$,

$$2\gamma\|\bar{Q}_\tau - Q_\omega^*\|_\infty \cdot \mathbf{1} \leq 4b'_\tau \sqrt{\frac{\log(8T|\mathcal{X}[\cup_{k \in [K_\omega]} Z_k^P]|/\delta)}{N_\tau}}.$$

The second term can be bounded by applying the result of Eq. (73)³ in Lemma D.5, we have that with probability at least $1 - \frac{\delta}{4T}$,

$$|\bar{\mathcal{H}}_{\tau+1}(Q_\omega^*) - \mathcal{H}(Q_\omega^*)| \leq c \left\{ \frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta) \cdot \mathbf{1}}{N_\tau(1-\gamma)} + \sqrt{\frac{\sigma_\mathcal{E}^2 \log(8T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{N_\tau}} \right\},$$

where c is a large enough constant.

Combining these two conditions, we can derive that, with probability at least $1 - \frac{\delta}{2T}$,

$$|\Delta r| \leq c \left\{ \frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{N_\tau(1-\gamma)} \cdot \mathbf{1} + (b'_\tau \mathbf{1} + \sigma_\mathcal{E}) \sqrt{\frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{N_\tau}} \right\}, \quad (77)$$

where c is a large enough constant.

Now we have already obtained the element-wise error bound of the reward function, which is the only difference for inducing \tilde{Q} and Q_ω^* . Applying Lemma A.4 with Eq. (77), we have:

$$\|\tilde{Q} - Q_\omega^*\| \leq \max\{\Delta_1, \Delta_2\},$$

where

$$\Delta_1 = c(I - \gamma P_\omega^{\pi_\omega^*})^{-1} \left\{ \frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{N_\tau(1-\gamma)} \cdot \mathbf{1} + (b'_\tau \mathbf{1} + \sigma_\mathcal{E}) \sqrt{\frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{N_\tau}} \right\}, \quad (78)$$

³The only difference is to applying the bound on the vector $|\bar{\mathcal{H}}_{\tau+1}(Q_\omega^*) - \mathcal{H}(Q_\omega^*)|$, instead of the infinity norm.

$$\Delta_2 = c(I - \gamma P_{\omega}^{\tilde{\pi}^*})^{-1} \left\{ \frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|/\delta)}{N_{\tau}(1-\gamma)} \cdot \mathbf{1} + (b'_{\tau} \mathbf{1} + \sigma_{\mathcal{E}}) \sqrt{\frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|/\delta)}{N_{\tau}}} \right\}. \quad (79)$$

Bounding Δ_1 : For bounding Δ_1 , the key is to bound $(I - \gamma P_{\omega}^{\tilde{\pi}^*})^{-1} \sigma_{\mathcal{E}}$. We have:

$$\|(I - \gamma P_{\omega}^{\tilde{\pi}^*})^{-1} \sigma_{\mathcal{E}}\|_{\infty} = \gamma^2 \|(I - \gamma P_{\omega}^{\tilde{\pi}^*})^{-1} \text{Var}_{P_{\omega}^*}(V_{\omega}^*)\|_{\infty} \leq \frac{\sqrt{2}}{(1-\gamma)^{\frac{3}{2}}}, \quad (80)$$

where the second inequality comes from Eq. (48) in Appendix D.5. By using $\|(I - \gamma P_{\omega}^{\tilde{\pi}^*})^{-1}\|_{\infty} \leq \frac{1}{1-\gamma}$ and Eq. (80), we have:

$$\begin{aligned} \Delta_1 &\leq c \left\{ \frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|/\delta)}{N_{\tau}(1-\gamma)^2} \cdot \mathbf{1} + (b'_{\tau} + \frac{\sqrt{2}}{(1-\gamma)^{\frac{3}{2}}}) \sqrt{\frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|/\delta)}{N_{\tau}}} \cdot \mathbf{1} \right\} \\ &\leq cb'_{\tau} \left\{ \frac{2^{\tau} \log(8T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|/\delta)}{N_{\tau}(1-\gamma)^{\frac{3}{2}}} + \frac{2^{\tau+2}}{1-\gamma} \sqrt{\frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|/\delta)}{N_{\tau}}} \right\} \cdot \mathbf{1}, \end{aligned} \quad (81)$$

where the second inequality is obtained by using $b'_{\tau} = \frac{2^{-\tau}}{\sqrt{1-\gamma}}$.

Bounding Δ_2 : For bounding Δ_2 , the difference comes from bounding $(I - \gamma P_{\omega}^{\tilde{\pi}^*})^{-1} \sigma_{\mathcal{E}}$. We define $\sigma_{\tilde{\mathcal{E}}}^2$ as the variance of induced by Q function \tilde{Q} and policy $\tilde{\pi}^*$. Then, we have:

$$\begin{aligned} \|(I - \gamma P_{\omega}^{\tilde{\pi}^*})^{-1} \sigma_{\mathcal{E}}\|_{\infty} &= \|(I - \gamma P_{\omega}^{\tilde{\pi}^*})^{-1} (\sigma_{\tilde{\mathcal{E}}} - \sigma_{\tilde{\mathcal{E}}} + \sigma_{\mathcal{E}})\|_{\infty} \\ &= \|(I - \gamma P_{\omega}^{\tilde{\pi}^*})^{-1} \sigma_{\tilde{\mathcal{E}}}\|_{\infty} + \|(I - \gamma P_{\omega}^{\tilde{\pi}^*})^{-1} (\sigma_{\tilde{\mathcal{E}}} - \sigma_{\mathcal{E}})\|_{\infty} \\ &\leq \frac{3}{(1-\gamma)^{\frac{3}{2}}} + \frac{\|\tilde{Q} - Q_{\omega}^*\|_{\infty}}{1-\gamma}. \end{aligned} \quad (82)$$

Combining Eq. (82) with Eq. (79), we have:

$$\begin{aligned} \Delta_2 &\leq cb'_{\tau} \left\{ \frac{2^{\tau} \log(8T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|/\delta)}{N_{\tau}(1-\gamma)^{\frac{3}{2}}} + \frac{2^{\tau+2}}{1-\gamma} \sqrt{\frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|/\delta)}{N_{\tau}}} \right\} \cdot \mathbf{1} \\ &\quad + c \|\tilde{Q} - Q_{\omega}^*\|_{\infty} \sqrt{\frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|/\delta)}{(1-\gamma)^2 N_{\tau}}} \cdot \mathbf{1}. \end{aligned} \quad (83)$$

Combining Eq. (83) and (81), we can conclude that:

$$\begin{aligned} \|\tilde{Q} - Q_{\omega}^*\|_{\infty} &\leq cb'_{\tau} \left\{ \frac{2^{\tau} \log(8T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|/\delta)}{N_{\tau}(1-\gamma)^{\frac{3}{2}}} + \frac{2^{\tau+2}}{1-\gamma} \sqrt{\frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|/\delta)}{N_{\tau}}} \right\} \\ &\quad + c \|\tilde{Q} - Q_{\omega}^*\|_{\infty} \sqrt{\frac{\log(8T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|/\delta)}{(1-\gamma)^2 N_{\tau}}}. \end{aligned} \quad (84)$$

Using a sample size $N_{\tau} = c_3 4^{\tau} \log(6T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|) (1-\gamma)^{-2}$ with large enough c_3 , we can guarantee:

$$c \sqrt{\frac{4^{\tau} \log(8T|\mathcal{X}[\cup_{k=1}^{K_{\omega}} Z_k^P]|/\delta)}{N_{\tau}}} \leq \frac{1}{2}, \quad (85)$$

$$\frac{c \cdot 2^\tau \log(8T |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{(1-\gamma)^{\frac{3}{2}} N_\tau} \leq \frac{1}{8}, \quad (86)$$

$$c \sqrt{\frac{4^{\tau+1} \log(8T |\mathcal{X}[\cup_{k=1}^{K_\omega} Z_k^P]|/\delta)}{(1-\gamma)^2 N_\tau}} \leq \frac{1}{8}. \quad (87)$$

Substituting condition (85)-(87) into Eq. (84) leads to $\|\tilde{Q} - Q_\omega^*\|_\infty \leq \frac{b'_\tau}{4}$. This concludes our proof. \square

E Discussion on the Cost-Optimal Sampling Problem

E.1 Relating the Cost-Optimal Sampling Problem to Graph Coloring

The Graph Coloring Problem (GCP) is defined as follows: Given a graph $G = (V, E)$, where V is a set of vertices and E is a set of edges, the task is to assign a color to each vertex such that no two adjacent vertices (i.e., vertices connected by an edge) share the same color. The objective is to find the optimal coloring scheme that minimizes the number of colors used, known as the *chromatic number* of the graph (Erdős and Hajnal, 1966).

In the Cost-Optimal Sampling Problem (COSP), components with scopes $Z_1^P, Z_2^P, \dots, Z_{K_\omega}^P$ must be divided into different groups, where the scopes of components in the same subset do not overlap (i.e., $Z_i^P \cap Z_j^P = \emptyset$ for any i, j in the same group). This problem can be modeled as a variant of the GCP. Specifically, each component i represents a vertex in a graph, and two vertices i and j are connected by an edge if their scopes overlap, i.e., $Z_i^P \cap Z_j^P \neq \emptyset$. Assigning a color to a vertex corresponds to assigning the component to a group. Since components in the same subset must have disjoint scopes, no two connected vertices (representing components with overlapping scopes) can share the same color. The goal is to find a coloring scheme that minimizes the total costs across all groups, where the cost of each subset is determined by the scope with largest factor set space $|\mathcal{X}[Z_i^P]|$ in that group.

E.2 Proof for NP-Completeness of the COSP

NP-completeness is used to characterize a subset of problems that are computationally hard to solve. Famous NP-complete problems include the Traveling Salesman Problem (TSP) (Lin, 1965), the Knapsack Problem (Kellerer et al., 2004), the Hamiltonian Path Problem (Gary and Johnson, 1979), and the Satisfiability (SAT) Problem (Schaefer, 1978).

A problem is classified as *NP-complete* if it satisfies two conditions:

- *In NP*: The problem belongs to the class NP, meaning that given a proposed solution, we can verify its feasibility in polynomial time.
- *NP-hard*: The problem is at least as hard as any other problem in NP. This means that any problem in NP can be transformed or reduced to an NP-hard problem in polynomial time, then we could solve the NP-hard problem to get the solution of any NP problem.

Graph coloring has been shown to be NP-complete (Karp, 2010), as it is computationally difficult to find the minimum number of colors for an arbitrary graph. We now show that the COSP is NP-complete by proving: (1) it belongs to NP, and (2) it is at least as hard as the NP-complete GCP.

E.2.1 The COSP belongs to NP

In the COSP, given a solution (a partition of components into groups and their associated costs), we can verify its feasibility in polynomial time by:

- Verifying that the scopes within each subset are disjoint, which can be done by checking pairwise intersections within each group. This can be done in $\mathcal{O}(K_\omega^2(n+m))$ time.
- Calculating the total cost of a partition by identifying the largest scope size in each subset and summing these values, which can also be computed in polynomial time, specifically $\mathcal{O}(K_\omega)$.

E.2.2 The COSP is NP-hard

We prove NP-hardness by reducing the GCP to the COSP. We demonstrate that for any instance of the GCP, a corresponding instance of the COSP can be constructed, and based on the optimal solution of the COSP, we can derive the optimal solution to the GCP.

Given any instance of the GCP with a graph $G = (V, E)$, we construct a corresponding instance of the COSP as follows:

- *Step 1. Assign unique base dimensions for each vertex:* Let $|V|$ be the number of vertices, we initial $|V|$ components with the scope for each component $i \in [|V|]$ satisfies $Z_i^P = \{i\}$. This ensures that each component starts with a unique, non-overlapping dimension.
- *Step 2. Create shared dimensions for edges:* For each edge $(v_i, v_j) \in E$, we introduce a new and unique dimension $d_{i,j}$ that will belong to both Z_i^P and Z_j^P . Specifically, we updates the scopes of components i and j by

$$Z_i^P = Z_i^P \cup \{d_{i,j}\}, \quad Z_j^P = Z_j^P \cup \{d_{i,j}\}.$$

This ensures that for every edge in the graph, the corresponding components have overlapping scopes, mimicking the adjacency constraint in the GCP.

- *Step 3. Enforce disjointness for non-adjacent vertices:* If there is no edge between two vertices v_i and v_j , the corresponding scopes should remain disjoint, i.e., $Z_i^P \cap Z_j^P = \emptyset$. This property is automatically maintained because shared dimensions are only introduced for pairs of vertices that are connected by an edge.

We provide an example to illustrate the construction process. Consider a small example graph $G = (V, E)$ with three vertices $V = \{v_1, v_2, v_3\}$ and edges $E = \{(v_1, v_2), (v_2, v_3)\}$. The construction process is:

- *Step 1. Assign unique base dimensions for each vertex:* $Z_1^P = \{1\}$, $Z_2^P = \{2\}$ and $Z_3^P = \{3\}$.
- *Step 2. Create shared dimensions for edges:* For edge (v_1, v_2) , we introduce $d_{1,2} = 4$, so $Z_1^P = \{1, 4\}$, $Z_2^P = \{2, 4\}$. Also, for edge (v_2, v_3) , we introduce $d_{2,3} = 5$, so $Z_2^P = \{2, 4, 5\}$, $Z_3^P = \{3, 5\}$.
- *Step 3. Resulting scopes:* the resulting scopes are:

$$Z_1^P = \{1, 4\}, \quad Z_2^P = \{2, 4, 5\}, \quad Z_3^P = \{3, 5\}.$$

The objective of the COSP is to partition these components into disjoint groups (sets of components with no overlapping scopes) while minimizing the total sampling cost. Since the sampling cost for each component is set to 1, minimizing the total sampling cost is equivalent to minimizing the number of groups (colors) used, as each subset contributes 1 to the total cost. Therefore, solving the COSP on this constructed instance is equivalent to solving the GCP.

E.2.3 NP-completeness

Since the GCP can be reduced to the COSP in polynomial time, and solving the COSP provides a solution to the GCP, the COSP is NP-hard. Furthermore, since the COSP belongs to NP, it is NP-complete.

F Detailed Model of Wind Farm-equipped Storage Control

F.1 System Model

Consider a wind farm tasked with supplying power to the grid under an electricity supply contract. The predetermined supply quantity of wind power at any given time t is denoted by \hat{w}_t . This value is established based on wind power forecasts, which can be made from a minute to a day in advance, either by the wind farm itself or the Independent System Operator (ISO). As time progresses, the actual wind power generation w_t is sequentially disclosed in real-time. The wind farm is equipped with a storage system, allowing it to either store wind energy, denoted as u_t^+ , or release stored energy, denoted as u_t^- into the grid at any time t . As a result, the total power delivered to the grid, represented by g_t , is determined by

$$g_t = w_t + v_t^- - v_t^+.$$

When the mismatch exists between commitment \hat{w}_t and the delivered power g_t , the wind farm will be charged with a penalty cost $c_t(\hat{w}_t, g_t)$ as follows:

$$c_t(\hat{w}_t, g_t) = p_t^+ \max(g_t - \hat{w}_t, 0) + p_t^- \max(\hat{w}_t - g_t, 0),$$

where p_t^+ and p_t^- denote the unit penalty costs for wind generation surplus and shortage at time t , respectively.

The wind farm targets to minimize the accumulated mismatch penalty costs across all T time slots by reasonably controlling the storage system. Mathematically, the storage control problem can be formulated as follows:

$$(P1) \quad \min_{v_t^+, v_t^-, \forall t} \sum_{t=1}^T c_t(\hat{w}_t, g_t) \quad (88)$$

$$s.t. \quad g_t = w_t + v_t^- - v_t^+, \forall t, \quad (89)$$

$$SoC_1 = \frac{C}{2}, \quad (90)$$

$$SoC_{t+1} = SoC_t + \eta^+ v_t^+ - \eta^- v_t^-, \forall t, \quad (91)$$

$$\eta^+ v_t^+ \leq C - SoC_t, \forall t, \quad (92)$$

$$\eta^- v_t^- \leq SoC_t, \forall t, \quad (93)$$

$$v_t^+ \leq w_t, \forall t, \quad (94)$$

$$v_t^+, v_t^- \geq 0, \forall t, \quad (95)$$

$$v_t^+ v_t^- = 0, \forall t. \quad (96)$$

In the optimization, the decision variables at time t include:

- v_t^+ : generated wind power which is charged to the energy storage;
- v_t^- : discharged energy from the storage to the grid;

And the other functions, latent variables, and parameters include:

- $c_t(\cdot)$: total penalty cost at time t ;
- \widehat{w}_t : committed wind power supply at time t ;
- w_t : wind power generation at time t ;
- T : total duration of storage control decisions;
- p_t^+, p_t^- : unit grid penalty prices for power generation shortage and surplus at time t , respectively;
- g_t : actual supplied energy at time t ;
- SoC_t : state-of-charge (SOC) of storage at time t ;
- C : energy storage capacity;
- η^+, η^- : charging and discharging efficiencies of storage;

Constraint (89) represents the delivered power; constraints (90) and (91) characterize the dynamics of storage; and constraints (92) and (93) denote the storage capacity limits. Constraint (94) and (95) show the upper and lower limits of storage control actions, and constraint (96) indicates that the storage cannot be charged and discharged simultaneously.

Due to the inherent stochasticity of penalty prices and renewable energy production, it is impractical to obtain the optimal future storage control decisions. Therefore, in practice, we often consider sequential storage control. Specifically, at each time t , we determine the current storage control actions v_t^+, v_t^-, c_t based on the currently available information. Consequently, we establish the following online storage control problem at time t as follows:

$$\text{(P2)} \quad \min_{v_t^+, v_t^-} c_t(\widehat{w}_t, g_t) + \sum_{\tau=t+1}^{\infty} \gamma^{\tau-t} \mathbb{E}(c_{\tau}(\widehat{w}_{\tau}, g_{\tau})) \quad (97)$$

$$s.t. \quad g_t = w_t + v_t^- - v_t^+, \quad (98)$$

$$SoC_{t+1} = SoC_t + \eta^+ v_t^+ - \eta^- v_t^-, \quad (99)$$

$$\eta^+ v_t^+ \leq C - SoC_t, \quad (100)$$

$$\eta^- v_t^- \leq SoC_t, \quad (101)$$

$$v_t^+ \leq w_t, \quad (102)$$

$$v_t^+, v_t^- \geq 0, \quad (103)$$

$$v_t^+ v_t^- = 0. \quad (104)$$

F.2 Markov Decision Process Modeling

We highlight that problem (P2) can be transformed into MDP in the following manner:

Markov Decision Process $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$:

- States \mathcal{S} : Any state $s \in \mathcal{S}$ is composed of the penalty prices p^+, p^- , the committed and real wind power generation \widehat{w} and w , and state-of-charge SoC . Formally, $s = (p^+, p^-, \widehat{w}, w, SoC)$;
- Actions \mathcal{A} : Any action $a \in \mathcal{A}$ is composed of the charge amount v^+ and discharge amount v^- . Formally, $a = (v^+, v^-)$;
- Transition probability \mathcal{P} : \mathcal{P} is the transiting probability matrix $\mathcal{P}_a = \{\mathbf{Pr}(s_{t+1} = s' | s_t = s, a_t = a), \forall s, s' \in \mathcal{S}, a \in \mathcal{A}\}$, which includes the probability of transiting from state s to s' with action a for all s, s' and a ;

- Reward \mathcal{R} : \mathcal{R} is the immediate reward (penalty in our case) after transiting from state s to state s' due to action a , i.e., $\mathcal{R} = \{r(s, a), \forall s, a\}$. Specifically, in the one-shot decision problem, the penalty $r(s, a)$ equals the negative of the penalty, i.e.,

$$r(s, a) = p^+ \max(w - \hat{w}, 0) + p^- \max(\hat{w} - w, 0).$$

We can observe that, for the storage control problem, the state space \mathcal{S} and the action space \mathcal{A} are known. The reward \mathcal{R} is also known once the state s and action a are decided. The only unknown comes from the transition probability \mathcal{P} . However, some important observations for \mathcal{P} can simplify the problem. Specifically, we can divide the state variables $s = (p^+, p^-, \hat{w}, w, SoC)$ into one deterministic state and several random states. The deterministic state is SoC , which can be determined following Eq. (99) without any uncertainty. And the random states include p^+, p^-, \hat{w}, w , which are fully random⁴.

In our numerical study, we assume $p = p^+ = p^-$, and observe that only $\Delta = \hat{w} - w$ exists in the reward function. Therefore, the state can be rewritten as $\mathcal{S} = (p, \Delta w, SoC)$. For the factorization scheme ω , we use the following factorization to the transition kernel:

$$\hat{P}(p_{t+1}, \Delta_{t+1}, SoC_{t+1} | p_t, \Delta_t, SoC_t, a_t) = \hat{P}(p_{t+1} | p_t) \hat{P}(\Delta_{t+1} | \Delta_t) \hat{P}(SoC_{t+1} | SoC_t, a_t).$$

F.3 Parameter Settings

In the numerical study, we utilized the California aggregate wind power generation dataset from CAISO (California ISO, 2021) containing predicted and real wind power generation data with a 5-minute resolution spanning from January 2020 to December 2020. The penalty price equals the average electricity price of CASIO (California ISO, 2021) with the matching resolution and periods. We set $C = 500$ kWh, $\gamma = 0.9$. The discretization levels of p , Δw and SoC are set to be 8, 8, 50, respectively. The action set includes 3 discretized choices: charging (discharging) to satisfy 100%, 50%, and 0% of energy mismatches, respectively.

⁴Note that, the committed wind power \hat{w} is essentially not random. However, only the generation mismatch $w - \hat{w}$ exists in our problem, and such mismatch can be regarded as a random variable.