Branching Reinforcement Learning

Yihan Du [1] **Wei Chen** [2]

Abstract

In this paper, we propose a novel Branching Reinforcement Learning (Branching RL) model, and investigate both Regret Minimization (RM) and Reward-Free Exploration (RFE) metrics for this model. Unlike standard RL where the trajectory of each episode is a single H-step path, branching RL allows an agent to take multiple base actions in a state such that transitions branch out to multiple successor states correspondingly, and thus it generates a tree-structured trajectory. This model finds important applications in hierarchical recommendation systems and online advertising. For branching RL, we establish new Bellman equations and key lemmas, i.e., branching value difference lemma and branching law of total variance, and also bound the total variance by only $O(H^2)$ under an exponentially-large trajectory. For RM and RFE metrics, we propose computationally efficient algorithms `BranchVI` and `BranchRFE`, respectively, and derive nearly matching upper and lower bounds. Our regret and sample complexity results are polynomial in all problem parameters despite exponentially-large trajectories.

1. Introduction

Reinforcement Learning (RL) (Burnetas & Katehakis, 1997; Sutton & Barto, 2018) models a fundamental sequential decision making problem, where an agent interacts with the environment over time in order to maximize the obtained rewards. Standard RL (Jaksch et al., 2010; Agrawal & Jia, 2017; Azar et al., 2017; Jin et al., 2018; Zanette & Brunskill, 2019) considers taking only a single action in a state and formulates a single H-step path model. However, in many real-world applications such as recommendation systems (Fu et al., 2021) and online advertising (Kang et al., 2020), we often need to select multiple options at a time, and each option can trigger a corresponding successor state. For example, in category-based shopping recommendation (Fu et al., 2021), the recommendation system often displays a list of main categories at the first step, where each one has a probability to be clicked. If a main category is clicked, at the second step, the system further provides a list of sub-categories according to the clicked main category. By analogy, at the last step, the system provides a list of items according to the chosen category path. In this process, users can select (trigger) more than one category-item paths, e.g., one may buy IT accessories-printers-laser printers and IT accessories-scanners-document scanners at once.

To handle such scenarios involving multiple actions and successor states, we propose a novel Branching Reinforcement Learning (Branching RL) framework, which is an episodic tree-structured forward model. In each episode, an agent starts from an initial state and takes a *super action* that contains multiple *base actions*, where each base action in this state has a probability to be triggered. For each state-base action pair, if triggered successfully, the agent receives a reward and transitions to a next state; Otherwise, if it is not triggered, the agent receives zero reward and transitions to an absorbing state associated with zero reward. Thus, the transitions branch out to multiple successor states. At the second step, for each branched-out state, the agent also selects a super action that contains multiple base actions with trigger probabilities. She only obtains rewards from the triggered state-base action pairs, and each state-base action pair transitions to a corresponding next state. Then, the transitions at the second step branch out to more successor states. By analogy, till the last step, she traverses an H-layer tree-structured trajectory, and only collects rewards at the triggered state-base action pairs.

Different from standard episodic RL (Azar et al., 2017; Jin et al., 2018; Zanette & Brunskill, 2019) where each trajectory is a single H-step path, the trajectory of branching RL is an H-layer triggered tree with exponentially increasing states and actions in each layer. This model allows an agent to take multiple base actions at once and handle multiple successor states. It can be applied to many hierarchical decision making scenarios, such as category-based recommendation systems (Fu et al., 2021) and online advertising (Kang et al., 2020).

[1] IIIS, Tsinghua University, Beijing, China [2] Microsoft Research. Correspondence to: Yihan Du <duyh18@mails.tsinghua.edu.cn>, Wei Chen <weic@microsoft.com>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

Under the branching RL model, we investigate two popular metrics in the RL literature, i.e., Regret Minimization (RM) and Reward-Free Exploration (RFE). In regret minimization (Jaksch et al., 2010; Azar et al., 2017; Zanette & Brunskill, 2019), the agent aims to minimize the gap between the obtained reward and the reward that can obtained by always taking the optimal policy. In reward-free exploration (Jin et al., 2020a; Kaufmann et al., 2021; Ménard et al., 2021), the agent explores the unknown environment (model) without observation of rewards, in order to estimate the model accurately such that for any given reward function, she can plan a near-optimal policy using the estimated model. The performance in RFE is measured by the number of episodes used during exploration (i.e., sample complexity).

Our work faces several unique challenges: (i) Since branching RL is a tree-structured forward model which greatly differs from standard RL, existing analytical tools for standard RL, e.g., Bellman equations, value difference lemma and law of total variance, cannot be directly applied to our problem. (ii) With exponentially-large trajectories, it is challenging to analyze the total variance and derive tight (polynomial) regret and sample complexity guarantees. (iii) Since the number of possible super actions can be combinatorially large, how to design a computationally efficient algorithm that avoids naive enumeration over all super actions is another challenge.

To tackle the above challenges, we establish novel analytical tools, including branching Bellman equations, branching value difference lemma and branching law of total variance, and bound the total variance by only $O(H^2)$ under exponentially-large trajectories. We also propose computationally efficient algorithms for both RM and RFE metrics, and provide nearly matching upper and lower bounds, which are polynomial in all problem parameters despite exponentially-large trajectories.

To sum up, our contributions in this paper are as follows:

- We propose a novel Branching Reinforcement Learning (Branching RL) framework, which is an episodic H-layer tree-structured forward model and finds important applications in hierarchical recommendation systems and online advertising. Under branching RL, we investigate two popular metrics, i.e., Regret Minimization (RM) and Reward-Free Exploration (RFE).
- We establish new techniques for branching RL, including branching Bellman equations, branching value difference lemma and branching law of total variance, and bound the total variance by only $O(H^2)$ despite exponentially-large trajectories.
- For both RM and RFE metrics, we design computationally efficient algorithms `BranchVI` and `BranchRFE`, respectively, and build near-optimal upper and lower bounds, which are polynomial in all problem parameters even with exponentially-large trajectories. When our problem reduces to standard RL, our results match the state-of-the-arts.

Due to space limit, we defer all proofs to Appendix.

2. Related Work

Below we review the literature of standard (episodic and tabular) RL with regret minimization (RM) and reward-free exploration (RFE) metrics.

Standard RL-RM. For the regret minimization (RM) metric, Jaksch et al. (2010) propose an algorithm that adds optimistic bonuses on transition probabilities, and achieves a regret bound with a gap in factors H, S compared to the lower bound (Jaksch et al., 2010; Osband & Van Roy, 2016). Here H is the length of an episode, and S is the number of states. Agrawal & Jia (2017) use posterior sampling and obtain an improved regret bound. Azar et al. (2017) build confidence intervals directly for value functions rather than transition probabilities, and provide the first optimal regret. Zanette & Brunskill (2019) design an algorithm based on both optimistic and pessimistic value functions, and achieve a tighter problem-dependent regret bounds without requiring domain knowledge. The above works focus on model-based RL algorithms. There are also other works (Jin et al., 2018; Zhang et al., 2020) studying model-free algorithms based on Q-learning with exploration bonus or advantage functions.

Standard RL-RFE. Jin et al. (2020a) introduce the reward-free-exploration (RFE) metric and design an algorithm that runs multiple instances of existing RM algorithm (Zanette & Brunskill, 2019), and their sample complexity has a gap to the lower bound (Jin et al., 2020a; Domingues et al., 2021) in factors H, S. Kaufmann et al. (2021) propose an algorithm which builds upper confidence bounds for the estimation error of value functions, and improve the sample complexity of (Jin et al., 2020a). Ménard et al. (2021) achieve a near-optimal sample complexity by applying an empirical Bernstein inequality and upper bounding the overall estimation error.

There are huge differences between standard RL and our branching RL. The exponentially-large trajectory of branching RL brings unique challenges in developing Bellman equations and key lemmas, designing computationally efficient algorithms and deriving optimal (polynomial) bounds. Existing RL algorithms and analysis cannot be applied to solve our challenges.

3. Problem Formulation

In this section, we present the formal formulation of Branching Reinforcement Learning (Branching RL).

Branching Markov Decision Process (Branching MDP). We consider an episodic branching MDP defined by a tuple $\mathcal{M} = (\mathcal{S}, A^{\texttt{univ}}, \mathcal{A}, m, H, q, p, r)$. Here $\mathcal{S} = \mathcal{S}^{\texttt{reg}} \cup \{s_\perp\}$ is the state space with cardinality S. $\mathcal{S}^{\texttt{reg}}$ is the set of *regular states*, and $s_\perp$ is an *ending state*, which is an absorbing state with zero reward. $A^{\texttt{univ}}$ is the set of *base actions*, which represents the set of all feasible items in recommendation. Let $N := |A^{\texttt{univ}}|$ denote the number of base actions. A *super action* $A \subset A^{\texttt{univ}}$ consists of m ($m \leq N$) base actions, which stands for a recommended list. $\mathcal{A}$ is the collection of all feasible *super actions* and can be combinatorially large. H is the length of an episode. Throughout the paper, we call a super action an action for short, call $(s, a) \in \mathcal{S} \times A^{\texttt{univ}}$ a *state-base action pair*, and call $(s, a) \in \mathcal{S} \setminus \{s_\perp\} \times A^{\texttt{univ}}$ a *regular state-base action pair*.

$q(s, a)$ is the trigger probability of state-base action pair $(s, a) \in \mathcal{S} \times A^{\texttt{univ}}$. $p(s'|s, a)$ is the probability of transitioning to state s' on state-base action pair (s, a), for any $(s', s, a) \in \mathcal{S} \times \mathcal{S} \times A^{\texttt{univ}}$. $r(s, a) \in [0, 1]$ is the reward of pair $(s, a) \in \mathcal{S} \times A^{\texttt{univ}}$. We assume that reward function r is deterministic as many prior RL works (Azar et al., 2017; Jin et al., 2018; Zhang et al., 2020), and our work can generalize to stochastic rewards easily. Parameters q, p, r are time-homogeneous, i.e., have the same definitions for different step $h \in [H]$. The ending state $s_\perp$ has zero reward and always transitions back to itself, i.e., $q(s_\perp, a) = 0$, $p(s_\perp|s_\perp, a) = 1$ and $r(s_\perp, a) = 0$ for all $a \in A^{\texttt{univ}}$. We define a policy π as a collection of H functions $\{\pi_h : \mathcal{S} \mapsto \mathcal{A}\}_{h \in [H]}$, and K as the number of episodes.

String-based Notations. As shown in Figure 1, in branching RL, the trajectory of each episode is an m-ary tree, where there are H layers (steps), and each layer $h \in [H]$ has m^{h-1} states (nodes) and m^h state-base action pairs (edges). We use the following string-based notations to denote a trajectory: Each tree node in layer h has a string index $\langle i_1, \dots, i_{h-1} \rangle$, with the root node for layer 1 having the empty string $\emptyset$, and $i_1, \dots, i_{h-1} \in \{1, 2, \dots, m\}$. The m children of this node have indices that concatenate $i_h \in [m]$ to the string, making it $\langle i_1, \dots, i_{h-1}, i_h \rangle$, where i_h stands for that this node is the i_h-th child of the node $\langle i_1, \dots, i_{h-1} \rangle$. Operator $\oplus$ is the concatenation operation for strings, and $i^{\oplus h}$ denotes the concatenation of h strings $\langle i \rangle$ for any $i \in [m], h \in [H]$. For any string σ, $|\sigma|$ denotes its length, and thus state s_σ is at step $|\sigma| + 1$.

Online Game. In each episode $k \in [K]$, an agent selects a policy π^k at the beginning, and starts from an initial state $s_\emptyset$. At step 1, she chooses an action $A_\emptyset = \{a_{\langle 1 \rangle}, \dots, a_{\langle m \rangle}\}$ according to π^k_1. Each state-base action pair $(s_\emptyset, a_{\langle i \rangle})$ for $i \in [m]$ has probability $q(s_\emptyset, a_{\langle i \rangle})$ to be triggered. If triggered successfully, the agent obtains reward $r(s_\emptyset, a_{\langle i \rangle})$ and this state-base action pair transitions to a next state $s_{\langle i \rangle} \sim p(\cdot|s_\emptyset, a_{\langle i \rangle})$; Otherwise, if not triggered successfully, she obtains zero reward and this state-base action pair transitions to the ending state $s_\perp$. Hence, the transitions at step 1 branch out to m successor states $s_{\langle 1 \rangle}, \dots, s_{\langle m \rangle}$. At step 2, for each state $s_{\langle i \rangle}$ ($i \in [m]$), she chooses an action $A_{\langle i \rangle} = \{a_{\langle i,1 \rangle}, \dots, a_{\langle i,m \rangle}\}$ according to π^k_2. Then, there are m^2 state-base action pairs $\{(s_{\langle i \rangle}, a_{\langle i,j \rangle})\}_{i,j \in [m]}$ at step 2, and each of them is triggered with probability $q(s_{\langle i \rangle}, a_{\langle i,j \rangle})$. If triggered successfully, the agent receives reward $r(s_{\langle i \rangle}, a_{\langle i,j \rangle})$ and this pair transitions to a next state $s_{\langle i,j \rangle} \sim p(\cdot|s_{\langle i \rangle}, a_{\langle i,j \rangle})$; Otherwise, she receives zero reward and this pair transitions to $s_\perp$. Then, the transitions at step 2 branch out to m^2 successor states $\{s_{\langle i,j \rangle}\}_{i,j \in [m]}$. The episode proceeds by analogy at the following steps $3, \dots, H$. In the trajectory tree, once the agent reaches $s_\perp$ at some node, she obtains no reward throughout this branch (This is so-called "ending state").

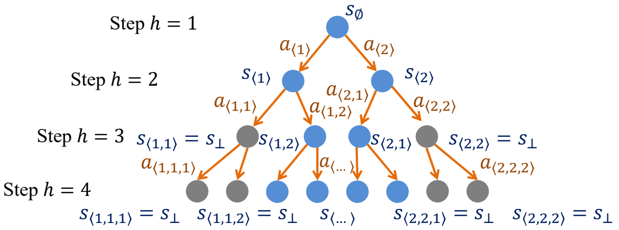
Step $h = 1$
Step $h = 2$
Step $h = 3$
Step $h = 4$
$s_\emptyset$
$a_{\langle 1 \rangle}$
$a_{\langle 2 \rangle}$
$s_{\langle 1 \rangle}$
$s_{\langle 2 \rangle}$
$a_{\langle 2,1 \rangle}$
$a_{\langle 1,1 \rangle}$
$a_{\langle 1,2 \rangle}$
$a_{\langle 2,2 \rangle}$
$s_{\langle 1,1 \rangle} = s_\perp$
$s_{\langle 1,2 \rangle}$
$s_{\langle 2,1 \rangle}$
$s_{\langle 2,2 \rangle} = s_\perp$
$a_{\langle 1,1,1 \rangle}$
$a_{\langle \dots \rangle}$
$a_{\langle 2,2,2 \rangle}$
$s_{\langle 1,1,1 \rangle} = s_\perp$
$s_{\langle 1,1,2 \rangle} = s_\perp$
$s_{\langle \dots \rangle}$
$s_{\langle 2,2,1 \rangle} = s_\perp$
$s_{\langle 2,2,2 \rangle} = s_\perp$

Figure 1. Illustrating example with $m = 2$ for branching RL.

Branching Value functions and Bellman Equations. For any policy π, we define value function $V^\pi_h : \mathcal{S} \mapsto \mathbb{R}$, so that

$$V^\pi_h(s) = \mathbb{E}_{q,p,\pi} \Bigg[\sum_{\sigma'=\emptyset}^{m^{\oplus(H-h)}} \sum_{\ell=1}^{m} q(s_{\sigma \oplus \sigma'}, a_{\sigma \oplus \sigma' \oplus \ell}) \cdot r(s_{\sigma \oplus \sigma'}, a_{\sigma \oplus \sigma' \oplus \ell}) | s_\sigma = s \Bigg] \quad (1)$$

gives the expected cumulative reward starting from some state s at step h till the end of this branch, under policy π. Here σ is the index string for an arbitrary state at step h, and thus $\sigma \in \{1^{\oplus(h-1)}, \dots, m^{\oplus(h-1)}\}$. $\sum_{\sigma'=\emptyset}^{m^{\oplus(H-h)}}$ denotes the summation over strings $\sigma' = \emptyset, \langle 1 \rangle, \dots, \langle m \rangle, \langle 1, 1 \rangle, \dots, \langle m, m \rangle, \dots, m^{\oplus(H-h)}$, which effectively enumerates all tree nodes of $H - h + 1$ layers. The expectation is taken with respect to the trajectory, which is dependent on trigger distribution q, transition distribution p and policy π.

Accordingly, we also define Q-value function $Q^\pi_h : \mathcal{S} \times \mathcal{A} \mapsto$

$\mathbb{R}$, so that

$$Q^\pi_h(s,A) = \mathbb{E}_{q,p,\pi}\Big[\sum_{\sigma'=\emptyset}^{m^{\oplus(H-h)}} \sum_{\ell=1}^{m} q(s_{\sigma\oplus\sigma'}, a_{\sigma\oplus\sigma'\oplus\ell}) \cdot r(s_{\sigma\oplus\sigma'}, a_{\sigma\oplus\sigma'\oplus\ell}) | s_\sigma = s, A_\sigma = A\Big]$$

denotes the expected cumulative reward starting from some state-action pair (s,A) at step h till the end of this branch, under policy π. From the definitions of r,p,q for ending state $s_\perp$, we have $V^\pi_h(s_\perp) = Q^\pi_h(s_\perp, A) = 0$ for any $A \in \mathcal{A}, h \in [H], \pi$.

Since $\mathcal{S}$, $\mathcal{A}$ and H are all finite, there exists a deterministic optimal policy π^* which has the optimal value $V^*_h(s) = \sup_\pi V^\pi_h(s)$ for any $s \in \mathcal{S}$ and $h \in [H]$. Then, we can establish the Bellman (optimality) equations as follows:

$$\begin{cases} Q^\pi_h(s,A) = \sum_{a\in A} q(s,a)\left(r(s,a) + p(\cdot|s,a)^\top V^\pi_{h+1}\right) \\ V^\pi_h(s) = Q^\pi_h(s,\pi_h(s)) \\ V^\pi_{H+1}(s) = 0,\ \forall s \in \mathcal{S}, \end{cases}$$

$$\begin{cases} Q^*_h(s,A) = \sum_{a\in A} q(s,a)\left(r(s,a) + p(\cdot|s,a)^\top V^*_{h+1}\right) \\ V^*_h(s) = \max_{A\in\mathcal{A}} Q^*_h(s,A) \\ V^*_{H+1}(s) = 0,\ \forall s \in \mathcal{S}. \end{cases}$$

Under the framework of branching RL, we consider two important RL settings, i.e., regret minimization (branching RL-RM) and reward-free exploration (branching RL-RFE).

Regret Minimization (RM). In branching RL-RM, the agent plays the branching RL game for K episodes, and the goal is to minimize the following regret

$$\texttt{Regret}(K) = \sum_{k=1}^{K} \left(V^*_1(s^k_\emptyset) - V^{\pi_k}_1(s^k_\emptyset)\right).$$

Reward-Free Exploration (RFE). Branching RL-RFE consists of two phases, i.e., exploration and planning. (i) In the exploration phase, given a fixed initial state $s_\emptyset$, the agent plays the branching RL game *without* the observation of reward function r, and estimates a trigger and transition model $(\hat{q},\hat{p})$. (ii) In the planning phase, the agent is given reward function r, and computes the optimal policy $\hat{\pi}^*$ under her estimated model $(\hat{q},\hat{p})$ with respect to r. Given an accuracy parameter ε and a confidence parameter δ, she needs to guarantee that for *any* given reward function r, the policy $\hat{\pi}^*$ with respect to r is *ε-optimal*, i.e.,

$$V^{\hat{\pi}^*}_1(s_\emptyset; r) \geq V^{\pi^*}_1(s_\emptyset; r) - \varepsilon,$$

with probability at least $1-\delta$. We measure the performance by *sample complexity*, i.e., the number of episodes used in the exploration phase to guarantee an ε-optimal policy for any given r.

In order to ensure that the number of triggered state-base action pairs will not increase exponentially and designing sample efficient algorithms is possible in branching RL, we introduce the following assumption.

Assumption 1 (Bounded Trigger Probability)**.** For any $(s,a) \in \mathcal{S} \times A^{\texttt{univ}}$, we have $q(s,a) \leq \frac{1}{m}$.

To justify the necessity of Assumption 1, we provide a rigorous lower bound to show that once relaxing the threshold of trigger probability, any branching RL algorithm must suffer an exponential regret.

Theorem 2. *Suppose that for any $(s,a) \in \mathcal{S} \times A^{\texttt{univ}}$, $q(s,a) \leq \bar{q}$ for some threshold parameter $\bar{q} > \frac{1}{m}$. Then, there exists an instance of branching RL with $H > 1$, where the regret of any algorithm is bounded by $\Omega(\frac{m\bar{q}((m\bar{q})^{H-1}-1)}{m\bar{q}-1}\sqrt{SNK})$.*

We describe the intuition behind this lower bound, and defer the full proof to Appendix C.2.2. Consider a branching MDP, where at an early step the agent has to distinguish the optimal action that has trigger probability $\bar{q}$, from the sub-optimal actions that have trigger probabilities only $\bar{q}-\eta$. Once the agent takes a sub-optimal action at the early step, such trigger sub-optimality will impact exponentially many states in the following steps, and she will suffer a regret of $m\eta \cdot \left(m\bar{q} + m^2\bar{q}^2 + \cdots + m^{H-1}\bar{q}^{H-1}\right)$ in this episode, which is the sum of a geometric progression with common ratio $m\bar{q}$. If $\bar{q} > \frac{1}{m}$, summing over all episodes, the total regret is exponentially large with respect to H.

Besides this lower bound, Assumption 1 is also mild in practice, since in real-world applications such as recommendation systems (Fu et al., 2021) and online advertising (Kang et al., 2020), it is often the case that users are only attracted to and click on a few items in a recommended list. In addition, in multi-step (e.g., category-based) recommendation, the interests of users usually converge to a single branch in the end (in expectation).

When $m = 1$, our branching RL reduces to standard episodic RL (Azar et al., 2017; Jin et al., 2018; Zanette & Brunskill, 2019) with transition probability $p^{\texttt{aug}}$, such that $p^{\texttt{aug}}(s_\perp|s,a) = 1 - q(s,a)$ and $p^{\texttt{aug}}(s'|s,a) = q(s,a)p(s'|s,a)$ for any $s' \neq s_\perp$. In this case, our results match the state-of-the-art results for standard RL in both RM (Azar et al., 2017; Zanette & Brunskill, 2019) and RFE (Ménard et al., 2021; Zhang et al., 2021) settings.

4. Properties of the Branching Markov Decision Process

Before introducing our algorithms for branching RL, in this section, we first investigate special structural properties of branching MDP, which are critical to deriving tight (polynomial) regret and sample complexity guarantees.

4.1. Branching Value Difference Lemma and Law of Total Variance

Different from standard episodic MDP (Azar et al., 2017; Zanette & Brunskill, 2019) where a trajectory is an H-step path, the trajectory of branching MDP is an m-ary tree with each node a state and each edge a state-base action pair. Thus, many analytical tools in standard MDP, e.g., value difference lemma (Dann et al., 2017) and law of total variance (Jin et al., 2018; Zanette & Brunskill, 2019), cannot be directly applied to branching MDP. To handle this problem, we establish new fundamental techniques for branching MDP, including branching value difference lemma and branching law of total variance.

First, we present a branching value difference lemma.

Lemma 3 (Branching Value Difference Lemma)**.** *For any two branching MDP $\mathcal{M}'(\mathcal{S}, A^{\texttt{univ}}, \mathcal{A}, m, H, q', p', r)$ and $\mathcal{M}''(\mathcal{S}, A^{\texttt{univ}}, \mathcal{A}, m, H, q'', p'', r)$, the difference in values under the same policy π satisfies that*

$$V_h'^{\pi}(s) - V_h''^{\pi}(s) = \sum_{\sigma'=\emptyset}^{m^{\oplus(H-h)}} \sum_{\ell=1}^{m} \mathbb{E}_{q'',p'',\pi}\Big[\big(q'(s_\tau, a_{\tau\oplus\ell}) \\ - q''(s_\tau, a_{\tau\oplus\ell})\big) \cdot r(s_\tau, a_{\tau\oplus\ell}) + \big(q'(s_\tau, a_{\tau\oplus\ell}) p'(s_\tau, a_{\tau\oplus\ell}) \\ - q''(s_\tau, a_{\tau\oplus\ell}) p''(s_\tau, a_{\tau\oplus\ell})\big)^\top V_{|\tau\oplus\ell|+1}'^{\pi} | s_\sigma = s\Big],$$

where $\tau := \sigma \oplus \sigma'$.

Using Lemma 3 with $\mathcal{M}'$ and $\mathcal{M}''$ being the optimistic and true models, respectively, we can bound the difference between optimistic and true values by the deviations between optimistic and true trigger and transition probabilities, in expectation with respect to the true model.

Next, we provide a branching law of total variance, which is critical to analyzing the estimation error of transition.

Lemma 4 (Branching Law of Total Variance)**.** *For any policy π,*

$$\mathbb{E}_{q,p,\pi}\Big[\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \mathrm{Var}_{q,p}\left(V_{|\sigma\oplus\ell|+1}^{\pi}(s_{\sigma\oplus\ell}) | s_\sigma, a_{\sigma\oplus\ell}\right)\Big]$$

$$= \mathbb{E}_{q,p,\pi}\Big[\Big(\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} q(s_\sigma, a_{\sigma\oplus\ell}) r(s_\sigma, a_{\sigma\oplus\ell}) - V_1^{\pi}(s_\emptyset)\Big)^2\Big] \quad (2)$$

$$\leq \mathbb{E}_{q,p,\pi}\Big[\Big(\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \mathbb{1}\left\{s_\sigma \neq s_\perp\right\}\Big)^2\Big]. \quad (3)$$

Here $\mathrm{Var}_{q,p}\left(V_{|\sigma\oplus\ell|+1}^{\pi}(s_{\sigma\oplus\ell}) | s_\sigma, a_{\sigma\oplus\ell}\right)$ denotes the variance of value $V_{|\sigma\oplus\ell|+1}^{\pi}(s_{\sigma\oplus\ell})$ with respect to $s_{\sigma\oplus\ell}$, which depends on trigger probability $q(s_\sigma, a_{\sigma\oplus\ell})$ and transition probability $p(\cdot|s_\sigma, a_{\sigma\oplus\ell})$, conditioning on $(s_\sigma, a_{\sigma\oplus\ell})$.

Remark 1. Lemma 4 exhibits that under branching MDP, the sum of conditional variances over all state-base action pairs is equal to the overall variance considering the whole trajectory, shown by Eq. (2). Furthermore, the overall variance can be bounded by the total number of regular (triggered) states, revealed by Eq. (3).

From Lemma 4, we have that to bound the estimation error of transition, which is related to the sum of conditional variances, it suffices to bound the total number of triggered states in a trajectory tree (discussed in the following).

4.2. The Number of Triggered States

In this subsection, we show that with Assumption 1 that only constrains the first moment of trigger distribution, we can bound both the first and second moments of the number of triggered states in a trajectory tree.

Lemma 5 (The Number of Triggered States)**.** *For any policy π,*

$$\mathbb{E}_{q,p,\pi}\Big[\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \mathbb{1}\left\{s_\sigma \neq s_\perp\right\}\Big] \leq H, \quad (4)$$

$$\mathbb{E}_{q,p,\pi}\Big[\Big(\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \mathbb{1}\left\{s_\sigma \neq s_\perp\right\}\Big)^2\Big] \leq 3H^2. \quad (5)$$

Remark 2. Eq. (4) gives a universal upper bound of value function as $V_h^{\pi}(s) \leq \Big[\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \mathbb{1}\left\{s_\sigma \neq s_\perp\right\}\Big] \leq H$ for any $s \in \mathcal{S}, h \in [H], \pi$. Moreover, Eq. (5) provides a sharp upper bound for overall variance, as well as the sum of conditional variances of transition (by plugging Eq. (5) into Lemma 4). To our best knowledge, this second moment result is novel.

Lemma 5 shows that despite the exponentially increasing nodes in branching MDP, its value and overall variance (estimation error) will not explode. This critical property enables us to avoid an exponentially-large regret or sample complexity.

Novel Analysis for Triggered States. The analysis of Lemma 5 is highly non-trivial. We first relax all regular trigger probabilities to $\frac{1}{m}$, and then investigate the number of triggered states for each step individually. While we can show that the number of triggered states at each step is a conditional Binomial random variable, the distribution of their sum is too complex to express. This incurs a nontrivial challenge on analyzing the second moment of the total number of triggered states. To tackle this challenge, we investigate the correlation of triggered states between any two steps, by exploiting the structure of branching MDP.

Proof sketch. Under Assumption 1, to bound the total number of triggered (regular) states for any branching MDP and policy π, it suffices to bound it under a relaxed model $\mathcal{M}^*$ with $q(s,a)=q^*:=\frac{1}{m}$ for all $(s,a)\in\mathcal{S}\setminus\{s_\perp\}\times A^{\texttt{univ}}$. Let ω_h denote the number of triggered states at each step h under $\mathcal{M}^*$, and $\omega:=\sum_{h=1}^{H}\omega_h$. Below we prove that $\mathbb{E}[\omega]\leq H$ and $\mathbb{E}[\omega^2]\leq 3H^2$.

For $h=1$, $\omega_h=1$ deterministically. For $h\geq 2$, $\omega_h|\omega_{h-1}\sim\texttt{Binomial}(m\omega_{h-1},q^*)$. According to the properties of Binomial distribution and $q^*:=\frac{1}{m}$, for $h\geq 2$,

$$\begin{aligned}\mathbb{E}\left[\omega_h\right] &= mq^*\mathbb{E}\left[\omega_{h-1}\right]=1,\\ \mathbb{E}\left[(\omega_h)^2\right] &= mq^*(1-q^*)\mathbb{E}\left[\omega_{h-1}\right]+m^2(q^*)^2\mathbb{E}\left[(\omega_{h-1})^2\right]\\ &= (1-q^*)+\mathbb{E}\left[(\omega_{h-1})^2\right]\leq h.\end{aligned}$$

Hence, we have that $\mathbb{E}[\omega]=\sum_{h=1}^{H}\mathbb{E}[\omega_h]=H$, and

$$\begin{aligned}\mathbb{E}[\omega^2] &= \sum_{h=1}^{H}\mathbb{E}[(\omega_h)^2]+2\sum_{1<i,j<H}\mathbb{E}[\omega_i\omega_j]\\ &\leq \frac{H(H+1)}{2}+2\sum_{1<i,j<H}\mathbb{E}[\omega_i\omega_j]. \qquad (6)\end{aligned}$$

Now, the challenge falls on how to bound $\mathbb{E}[\omega_i\omega_j]$ for any $1<i,j<H$. Let W_σ be a Bernoulli random variable denoting whether state s_σ is triggered for any index string σ. Then, we can write $\mathbb{E}[\omega_i\omega_j]$ as

$$\begin{aligned}&\mathbb{E}\Big[\Big(\sum_{\sigma=1^{\oplus(i-1)}}^{m^{\oplus(i-1)}}W_\sigma\Big)\cdot\Big(\sum_{\sigma'=1^{\oplus(j-1)}}^{m^{\oplus(j-1)}}W_{\sigma'}\Big)\Big]\\ \overset{\text{(a)}}{=}&m^{i-1}\mathbb{E}\Big[W_{1^{\oplus(i-1)}}\Big(\sum_{\sigma'=1^{\oplus(j-1)}}^{m^{\oplus(j-1)}}W_{\sigma'}\Big)\Big]\\ =&m^{i-1}\mathbb{E}\Big[W_{1^{\oplus(i-1)}}\Big(\sum_{\substack{\sigma'=1^{\oplus(j-1)}\\ \sigma'\text{starts with }1^{\oplus(i-1)}}}^{m^{\oplus(j-1)}}W_{\sigma'}+\sum_{\substack{\sigma'=1^{\oplus(j-1)}\\ \sigma'\text{does not start with }1^{\oplus(i-1)}}}^{m^{\oplus(j-1)}}W_{\sigma'}\Big)\Big]\\ \overset{\text{(b)}}{\leq}&m^{i-1}\left(m^{j-i}(q^*)^{i-1+j-i}+m^{j-1}(q^*)^{i-1+j-1}\right)\\ =&2. \qquad (7)\end{aligned}$$

Here (a) comes from the symmetry of trajectory tree. (b) is due to that at step j, the children states of $s_{1^{\oplus(i-1)}}$ have dependency on it and the other states are independent of it, and $\mathbb{E}[W_\sigma W_{\sigma'}]=\Pr[W_\sigma=1,W_{\sigma'}=1]$ for any σ,σ'. By plugging Eq. (7) into Eq. (6), we have $\mathbb{E}[\omega^2]\leq\frac{H(H+1)}{2}+4\frac{H(H-1)}{2}\leq 3H^2$. Thus, we obtain Lemma 5. □

Algorithm 1 `BranchVI`

```
Input: confidence parameter $\delta$, $\delta' := \frac{1}{6}\delta$, $L := \log(\frac{SNH(m^H \vee K)}{\delta'})$. Initialize $\bar{V}^k_h(s_\perp) = \underline{V}^k_h(s_\perp) = 0,\ \forall h \in [H], k$.
for $k = 1, 2, \dots$ do
    for $h = H, H-1, \dots, 1$ do
        for $s \in \mathcal{S} \setminus \{s_\perp\}$ do
            for $a \in A^{\texttt{univ}}$ do
                $\hat{q}^k(s,a) \leftarrow \frac{J^k_{\texttt{sum}}(s,a)}{n^k(s,a)}$. $b^{k,q}(s,a) \leftarrow 4\sqrt{\frac{L}{n^k(s,a)}}$;
                $\hat{p}^k(s'|s,a) \leftarrow \frac{P^k_{\texttt{sum}}(s'|s,a)}{J^k_{\texttt{sum}}(s,a)},\ \forall s' \in \mathcal{S}$;
                $b^{k,qpV}(s,a) \leftarrow 4\sqrt{\frac{\mathrm{Var}_{s'}\left(\bar{V}^k_{h+1}(s')\right)L}{n^k(s,a)}} + 4\sqrt{\frac{\mathbb{E}_{s'}\left[\left(\bar{V}^k_{h+1}(s') - \underline{V}^k_{h+1}(s')\right)^2\right]L}{n^k(s,a)}} + \frac{36HL}{n^k(s,a)}$;
                $f^k_h(s,a) \leftarrow (\hat{q}^k(s,a) + b^{k,q}(s,a))r(s,a) + \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \bar{V}^k_{h+1} + b^{k,qpV}(s,a)$;
            end for
            $\bar{V}^k_h(s) \leftarrow \min\{\max_{A \in \mathcal{A}} \sum_{a \in A} f^k_h(s,a),\ H\}$;
            $\pi^k_h(s) \leftarrow \mathrm{argmax}_{A \in \mathcal{A}} \sum_{a \in A} f^k_h(s,a)$;
            $\underline{V}^k_h(s) \leftarrow \max\{\sum_{a \in \pi^k_h(s)}((\hat{q}^k(s,a) - b^{k,q}(s,a))r(s,a) + \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \underline{V}^k_{h+1} - b^{k,qpV}(s,a)),\ 0\}$;
        end for
    end for
    Take policy $\pi^k$ and observe the trajectory;
end for
```

5. Branching Reinforcement Learning with Regret Minimization

In this section, we study branching RL-RM, and propose an efficient algorithm `BranchVI` with a near-optimal regret guarantee for large enough K. A lower bound is also established to validate the optimality of `BranchVI`.

5.1. Algorithm `BranchVI`

The algorithm design for branching RL faces two unique *challenges*: (i) Computation efficiency. Since the action space $\mathcal{A}$ can be combinatorially large, it is inefficient to explicitly maintain Q function as in standard RL (Azar et al., 2017; Zanette & Brunskill, 2019); (ii) Tight optimistic estimator. Naively adapting standard RL algorithms (Azar et al., 2017; Zanette & Brunskill, 2019) and adding optimistic bonuses on trigger and transition probabilities, respectively, will lead to a loose regret bound (see Appendix A for details). To handle these challenges, `BranchVI` only maintains the components of Q function, and uses a maximization oracle to directly calculate V function. In addition, `BranchVI` considers trigger and transition distributions as a whole and adds a composite bonus.

The procedure of `BranchVI` (Algorithm 1) is as follows. In each episode k, we first calculate the empirical trigger and transition probabilities $\hat{q}^k, \hat{p}^k$, a bonus for trigger probability $b^{k,q}$, and a composite bonus for trigger and transition probabilities $b^{k,qpV}$ (Lines 1-1). Here $n^k(s,a)$, $J^k_{\texttt{sum}}(s,a)$ and $P^k_{\texttt{sum}}(s'|s,a)$ denote the number of times (s,a) was visited, the number of times (s,a) was successfully triggered, and the number of times the agent transitioned to s' from (s,a) up to episode k, respectively. Then, we calculate a component function $f^k_h(s,a)$, which represents the contribution to value function from each (s,a) (Line 1).

We allow `BranchVI` to access a maximization oracle which can efficiently calculate $\max_{A \in \mathcal{A}} \sum_{a \in A} w(a)$ and $\operatorname{argmax}_{A \in \mathcal{A}} \sum_{a \in A} w(a)$ for any vector $\boldsymbol{w} \in \mathbb{R}^N$ ($N := |A^{\texttt{univ}}|$). Since the objective function $\sum_{a \in A} w(a)$ is linear, such oracle exists for many combinatorial decision classes, e.g., all m-cardinality subsets and m-cardinality matchings. By utilizing this oracle with $f^k_h(s,a)$, we can efficiently calculate the optimistic value function $\bar{V}^k_h(s)$ and policy π^k_h, and further compute the pessimistic value function $\underline{V}^k_h$ (Lines 1-1). After figuring out $\bar{V}^k_h(s), \underline{V}^k_h, \pi^k_h(s)$ for all s, h, we execute policy π^k in episode k (Line 1).

Computation Efficiency. We remark that the computation complexity of `BranchVI` is $O(S^2 N)$, instead of expensive $O(S^2 |\mathcal{A}|)$ as one may suffer by naively adapting standard RL algorithms (Azar et al., 2017; Zanette & Brunskill, 2019). This advantage is due to that `BranchVI` only maintains the component function instead of the Q function, and utilizes a maximization oracle to directly compute V function.

5.2. Regret Upper Bound for `BranchVI`

Now we provide the regret guarantee for `BranchVI`.

Theorem 6 (Regret Upper Bound)**.** *With probability at least $1-\delta$, for any episode $K > 0$, the regret of algorithm* `BranchVI` *is bounded by $O(H\sqrt{SNK}\log(\frac{SNH(m^H \vee K)}{\delta}))$. In particular, when $K \geq m^H$, the regret is bounded by*

$$O\left(H\sqrt{SNK}\log\left(\frac{SNHK}{\delta}\right)\right)$$

Remark 3. Theorem 6 shows that, despite the exponentially-large trajectory of branching RL, `BranchVI` enjoys a regret only polynomial in problem parameters. For large enough K such that $K \geq m^H$, Theorem 6 matches the lower bound (presented in Section 5.3) up to logarithmic factors. In addition, when branching RL reduces to standard RL (i.e., $m=1$), our result also matches the state-of-the-arts (Azar et al., 2017; Zanette & Brunskill, 2019).

Branching Regret Analysis. In contrast to standard RL (Azar et al., 2017; Zanette & Brunskill, 2019), we derive a tree-structured regret analysis based on special structural properties of branching RL in Section 4.

Using the branching value difference lemma (Lemma 3), we can decompose the regret into the estimation error from each regular state-base action pair as follows:

$$\begin{aligned}
\texttt{Regret}(K) \leq & \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a), s \neq s_\perp} w_{k,\sigma,\ell}(s,a) \cdot \\
& \Big[\underbrace{\big(\tilde{q}^k(s,a) - q(s,a)\big)\, r(s,a)}_{\texttt{Term 1: Triggered reward}} \\
& + \underbrace{\big(\tilde{q}^k(s,a)\tilde{p}^k(\cdot|s,a) - \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)\big)^\top \bar{V}^k_{|\sigma\oplus\ell|+1}}_{\texttt{Term 2: Triggered transition optimism}} \\
& + \underbrace{\big(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a) - q(s,a)p(\cdot|s,a)\big)^\top V^*_{|\sigma\oplus\ell|+1}}_{\texttt{Term 3: Triggered transition estimation}} \\
& + \underbrace{\big(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a) - q(s,a)p(\cdot|s,a)\big)^\top \Big(\bar{V}^k_{|\sigma\oplus\ell|+1} - V^*_{|\sigma\oplus\ell|+1}\Big)}_{\texttt{Term 4: Lower order term}} \Big].
\end{aligned}$$

Here $w_{k,\sigma,\ell}(s,a)$ denotes the probability that $(s_\sigma, a_{\sigma\oplus\ell}) = (s,a)$ in episode k, and $\tilde{q}^k$ and $\tilde{p}^k$ represent optimistic trigger and transition probabilities, respectively. In this decomposition, we address the estimation error for triggered reward (`Term 1`) and triggered transition (`Terms 2, 3`) separately, and consider trigger and transition probabilities as a whole distribution (in `Terms 2, 3`). The dominant terms are `Terms 2, 3`, which stand for the estimation error for triggered transition and depend on the sum of conditional variances. Using branching law of total variance (Lemma 4) and the second moment bound of triggered states (Lemma 5), we can bound `Terms 2, 3` by only $O(H^2)$ despite exponential state-base action pairs.

We note that it is necessary to separately address triggered reward and triggered transition, and consider trigger and transition as a whole distribution in bonus design (Line 1) and regret decomposition. A naive adaption of standard RL algorithm (Zanette & Brunskill, 2019), which adds bonuses on trigger and transition probabilities respectively, i.e., replacing Line 1 with $f^k_h(s,a) \leftarrow \big(\hat{q}^k(s,a) + b^{k,q}(s,a)\big) \cdot \big(r(s,a) + \hat{p}^k(\cdot|s,a)^\top \bar{V}^k_{h+1} + b^{k,pV}(s,a)\big)$, will suffer an extra factor $\sqrt{H}$ in the regret bound. Please see Appendix A for more discussion.

5.3. Regret Lower Bound

In this subsection, we provide a lower bound for branching RL-RM which is polynomial in problem parameters and

Algorithm 2 `BranchRFE`

```
1: Input: $s_\emptyset$, $\varepsilon$, $\delta$, $\beta(t,\kappa) := \log(SN/\kappa) + S\log(8e(t+1))$ for any $t \in \mathbb{N}$ and $\kappa \in (0,1)$. Initialize $B_h^k(s_\perp) = 0, \forall h \in [H], k$ and $B_{H+1}^k(s) = 0, \forall s \in \mathcal{S}, k$.
2: for $k = 1, 2, \dots$ do
3:   for $h = H, H-1, \dots, 1$ do
4:     for $s \in \mathcal{S} \setminus \{s_\perp\}$ do
5:       for $a \in A^{\texttt{univ}}$ do
6:         $\hat{q}^k(s,a) \leftarrow \frac{J^k_{\texttt{sum}}(s,a)}{n^k(s,a)}$;
7:         $\hat{p}^k(s'|s,a) \leftarrow \frac{P^k_{\texttt{sum}}(s'|s,a)}{J^k_{\texttt{sum}}(s,a)}$, $\forall s' \in \mathcal{S}$;
8:         $g_h^k(s,a) \leftarrow 12H^2 \frac{\beta(n^k(s,a),\delta)}{n^k(s,a)} + (1+\frac{1}{H})\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top B_{h+1}^k(s)$;
9:       end for
10:      $\pi_h^k(s) \leftarrow \mathrm{argmax}_{A \in \mathcal{A}} \sum_{a \in A} g_h^k(s,a)$;
11:      $B_h^k(s) \leftarrow \min\{\max_{A \in \mathcal{A}} \sum_{a \in A} g_h^k(s,a),\ H\}$;
12:    end for
13:  end for
14:  if $4e\sqrt{B_1^k(s_\emptyset)} + B_1^k(s_\emptyset) \le \frac{\varepsilon}{2}$, return $(\hat{q}^k, \hat{p}^k)$;
15:  else Take policy $\pi^k$ and observe the trajectory;
16: end for
```

demonstrates the optimality of `BranchVI`.

Theorem 7 (Regret Lower Bound)**.** *There exists an instance of branching RL-RM, where any algorithm must have* $\Omega(H\sqrt{SNK})$ *regret.*

Remark 4. Theorem 7 validates that the regret of `BranchVI` (Theorem 6) is near-optimal for large enough K, and reveals that, a polynomial regret is achievable and tight even with exponentially-large trajectories in branching RL.

Branching Regret Lower Bound Analysis. Unlike prior standard episodic RL works (Azar et al., 2017; Jin et al., 2018) which directly adapt the diameter-based lower bound (Jaksch et al., 2010) to the episodic setting, we derive a new lower bound analysis for branching RL-RM. We construct a hard instance, where an agent uniformly enters one of $\Theta(S)$ "bandit states", i.e., states with an optimal action and sub-optimal actions, and hereafter always transitions to a "homogeneous state", i.e., a state with homogeneous actions. Then, if the agent makes a mistake in a bandit state, she will suffer $\Omega(H)$ regret in this episode. By bounding the KL-divergence between this hard instance and uniform instance, we can derive a desired lower bound.

6. Branching Reinforcement Learning with Reward-free Exploration

In this section, we investigate branching RL-RFE, and develop an efficient algorithm `BranchRFE` and nearly matching upper and lower bounds of sample complexity.

6.1. Algorithm `BranchRFE`

In each episode, `BranchRFE` estimates the trigger and transition probabilities, and computes the estimation error $B_h^k(s)$, which stands for the cumulative variance of trigger and transition from step h to H. Once the total estimation error $B_1^k(s_\emptyset)$ is shrunk below the required accuracy, `BranchRFE` returns the estimated model $(\hat{q}^k, \hat{p}^k)$. Given any reward function r, the optimal policy $\hat{\pi}^*$ under $(\hat{q}^k, \hat{p}^k)$ with respect to r is ε-optimal, i.e., $V_1^{\hat{\pi}^*}(s_\emptyset; r) \ge V_1^*(s_\emptyset; r) - \varepsilon$, with probability at least $1-\delta$.

We describe `BranchRFE` (Algorithm 2) as follows: In each episode k, for each step h, `BranchRFE` first estimates the trigger and transition probabilities $\hat{q}^k, \hat{p}^k$ (Lines 2,2), and calculates the component estimation error $g_h^k(s,a)$ for each state-base action pair (Line 2). Then, similar to `BranchVI`, we utilize a maximization oracle to efficiently find the action with the maximum estimation error (the most necessary action for exploration) to be $\pi_h^k(s)$, and assign such maximum error to $B_h^k(s)$ (Lines 2,2). If the square root of total estimation error $\sqrt{B_1^k(s_\emptyset)}$, which represents the standard deviation of trigger and transition for whole trajectory, is smaller than $\varepsilon/2$, `BranchRFE` stops and outputs the estimated model $(\hat{q}^k, \hat{p}^k)$ (Line 2); Otherwise, it continues to explore according to the computed policy π^k (Line 2).

The computation complexity of `BranchRFE` is also $O(S^2N)$ instead of $O(S^2|\mathcal{A}|)$, since it only computes the component estimation error for state-base action pairs rather than enumerating super actions, and utilizes a maximization oracle to calculate $\pi_h^k(s)$ and $B_h^k(s)$.

6.2. Sample Complexity Upper Bound for `BranchRFE`

Now we present the sample complexity for `BranchRFE`. We say an algorithm for branching RL-RFE is (δ, ε)-correct, if it returns an estimated model $(\hat{q}, \hat{p})$ such that given any reward function r, the optimal policy under $(\hat{q}, \hat{p})$ with respect to r is ε-optimal with probability at least $1-\delta$.

Theorem 8 (Sample Complexity Upper Bound)**.** *For any* $\varepsilon > 0$ *and* $\delta \in (0,1)$*, algorithm* `BranchRFE` *is* (δ, ε)*-correct. Moreover, with probability* $1-\delta$*, the number of episodes used in* `BranchRFE` *is bounded by*

$$O\left(\frac{H^2SN}{\varepsilon^2}\left(\log\left(\frac{SN}{\delta}\right) + S\log\left(e \cdot m^H\right)\right) \cdot C^2\right),$$

where $C = \log\left(\left(\log\left(\frac{SN}{\delta}\right) + S\log\left(e \cdot m^H\right)\right) \cdot \frac{HSN}{\varepsilon}\right)$.

Remark 5. Theorem 6 exhibits that even with exponentially-large trajectories in branching RL, `BranchRFE` only needs polynomial episodes to ensure an ε-optimal policy for any reward function. This sample complexity is optimal for small enough δ within logarithmic factors (compared to the lower bound presented in Section 6.3). In addition, when degenerating to standard RL (i.e., $m = 1$), our result also matches the state-of-the-arts (Ménard et al., 2021; Zhang et al., 2021).

Branching Sample Complexity Analysis. Unlike standard RL (Ménard et al., 2021; Zhang et al., 2021) which only bounds the estimation error in a single H-step path, in branching RL we need to unfold and analyze the estimation error for all state-base action pairs in a trajectory tree. In our analysis, we utilize special structural properties of branching MDP, e.g., branching law of total variance (Lemmas 4) and the second moment bound of triggered states (Lemmas 5), to skillfully bound the total estimation error throughout the trajectory tree. Despite exponentially-large trajectories, we obtain sample complexity only polynomial in problem parameters.

6.3. Sample Complexity Lower Bound

Theorem 9 (Sample Complexity Lower Bound)**.** *There exists an instance of branching RL-RFE, where any* (δ, ε)*-correct algorithm requires* $\Omega(\frac{H^2 SN}{\varepsilon^2} \log \delta^{-1})$ *trajectories.*

Remark 6. Theorem 9 demonstrates that `BranchRFE` (Theorem 8) achieves a near-optimal sample complexity for small enough δ, and also, a polynomial sample complexity is achievable and sharp for branching RL-RFE.

7. Experiments

In this section, we conduct experiments for branching RL. We set $K = 5000$, $\delta = 0.005$, $H = 6$, $m = 2$, $N \in \{10, 15\}$, $\mathcal{S} = \{s_{\perp}, s_1, \dots, s_5\}$. $\mathcal{A}$ is the collection of all m-cardinality subsets of $A^{\texttt{univ}} = \{a_1, \dots, a_N\}$, and thus $|\mathcal{A}| = \binom{N}{m} \in \{45, 105\}$. The reward function $r(s, a) = 1$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. The trigger probability $q(s, a) = \frac{1}{m}$ for any $(s, a) \in \mathcal{S} \times \{a_{N-1}, a_N\}$, and $q(s, a) = \frac{1}{2m}$ for any $(s, a) \in \mathcal{S} \times A^{\texttt{univ}} \setminus \{a_{N-1}, a_N\}$. We set s_1 as the initial state for each episode. Under all actions $a \in A^{\texttt{univ}}$, the transition probability $q(s'|s_1, a) = 0.5$ for any $s' \in \{s_2, s_3\}$, and $q(s'|s, a) = 0.5$ for any $(s, s') \in \{s_2, s_3\} \times \{s_4, s_5\}$ or $(s, s') \in \{s_4, s_5\} \times \{s_2, s_3\}$. We perform 50 independent runs, and report the average regrets and running times (in legends) across runs.

Since we study a new problem and there is no existing algorithm for branching RL, we compare our algorithm `BranchVI` with two adaptations from standard RL, i.e., `Euler-Adaptation` (Zanette & Brunskill, 2019) and ε-`Greedy` ($\varepsilon = 0.01$). The former uses individual exploration bonuses for trigger and transition, the latter uses ε-greedy in exploration, and both of them explicitly maintain Q-functions. As shown in Figure 2, `BranchVI` enjoys a lower regret and a faster running time than the baselines, which demonstrates the effectiveness of our exploration strategy and the computation efficiency of our algorithm.

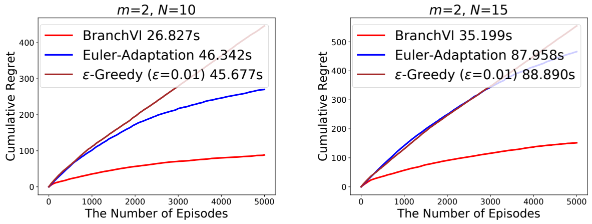

Figure 2. Experiments for branching RL.

8. Conclusion and Future Work

In this paper, we formulate a novel branching RL model, and consider both regret minimization and reward-free exploration metrics. Different from standard RL where each episode is a single H-step path, branching RL is a tree-structured forward model which allows multiple base actions in a state and multiple successor states. For branching RL, we build novel fundamental analytical tools and carefully bound the overall variance. We design efficient algorithms and provide near-optimal upper and lower bounds.

There are many interesting directions to explore. One direction is to improve the dependency on H in our regret upper bound (Theorem 6) for small K and close the gap on factors H, S between sample complexity upper and lower bounds (Theorems 8,9). Another direction is to extend branching RL from the tabular setting to function approximation, e.g., representing the value function in a linear form with respect to the feature vectors of state-action pairs (Jin et al., 2020b; Zhou et al., 2021).

References

Agrawal, S. and Jia, R. Posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pp. 1184–1194, 2017.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

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

Burnetas, A. N. and Katehakis, M. N. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in*

Neural Information Processing Systems, pp. 2818–2826, 2015.

Dann, C., Lattimore, T., and Brunskill, E. Unifying PAC and regret: uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 5717–5727, 2017.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

Fu, M., Agrawal, A., Irissappane, A. A., Zhang, J., Huang, L., and Qu, H. Deep reinforcement learning framework for category-based item recommendation. *IEEE Transactions on Cybernetics*, 2021.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4868–4878, 2018.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020a.

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

Kang, S., Jeong, C., and Chung, K. Tree-based real-time advertisement recommendation system in online broadcasting. *IEEE Access*, 8:192693–192702, 2020.

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. Adaptive reward-free exploration. In *International Conference on Algorithmic Learning Theory*, pp. 865–891. PMLR, 2021.

Mannor, S. and Tsitsiklis, J. N. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

Ménard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., and Valko, M. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pp. 7599–7608. PMLR, 2021.

Osband, I. and Van Roy, B. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

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

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 2020.

Zhang, Z., Du, S. S., and Ji, X. Nearly minimax optimal reward-free reinforcement learning. *International Conference on Machine Learning*, 2021.

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

Appendix

A. Discussion for Algorithm `BranchVI`

We remark that it is necessary to separately address triggered reward and triggered transition, and consider trigger and transition as a whole distribution in bonus design (Line 1 in Algorithm 1) and regret analysis (Eq. (13)). A counter example is discussed below to support this statement.

If one naively adapts standard RL algorithms (Azar et al., 2017; Zanette & Brunskill, 2019), she/he may directly add bonuses on trigger and transition probabilities, respectively, without separating triggered reward and triggered transition. In this case, $f_h^k(s,a)$ (Line 1) becomes:

$$f_h^k(s,a) \leftarrow \left(\hat{q}^k(s,a) + b^{k,q}(s,a)\right) \left(r(s,a) + \hat{p}^k(\cdot|s,a)^\top \bar{V}_{h+1}^k + b^{k,pV}(s,a)\right)$$

Then, in regret decomposition, we will obtain

$$\texttt{Regret}(K) \overset{\text{(a)}}{\approx} O(1) \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a), s \neq s_\perp} w_{k\sigma\ell}(s,a) \Big[b^{k,q}(s,a) \Big(r(s,a) + \underbrace{p(\cdot|s,a)^\top \bar{V}_{|\sigma\oplus\ell|+1}^k}_{\texttt{Term } \Lambda} \Big) + b^{k,p\bar{V}}(s,a) q(s,a) \Big],$$

where (a) omits second order terms. Since `Term Λ` already reaches $\Theta(h)$ order, further summing over all episodes k and steps h, we will suffer an extra $\sqrt{H}$ factor in the final result.

One can see from this counter example that, our bonus design and analysis, which separately handle triggered reward and triggered transition and consider trigger and transition as a whole distribution, are sharp and enable a near-optimal regret.

B. Proofs for Properties of Branching MDP

In this section, we present the proofs for structural properties of branching MDP, including branching value difference lemma (Lemma 3), branching law of total variance (Lemma 4) and the upper bounds of the number of triggered states (Lemma 5).

B.1. Proof of Lemma 3

Proof of Lemma 3. This proof adapts the analysis of Lemma E.15 in (Dann et al., 2017) to branching RL. According to branching Bellman equations,

$$\begin{aligned}
& V_h'^{\pi}(s_\sigma) - V_h''^{\pi}(s_\sigma) \\
\overset{\text{(a)}}{=} & \sum_{\ell=1}^{m} \Big(q'(s_\sigma, a_{\sigma\oplus\ell}) r(s_\sigma, a_{\sigma\oplus\ell}) - q''(s_\sigma, a_{\sigma\oplus\ell}) r(s_\sigma, a_{\sigma\oplus\ell}) \\
& + q'(s_\sigma, a_{\sigma\oplus\ell}) p'(s_\sigma, a_{\sigma\oplus\ell})^\top V_{h+1}'^{\pi} - q''(s_\sigma, a_{\sigma\oplus\ell}) p''(s_\sigma, a_{\sigma\oplus\ell})^\top V_{h+1}''^{\pi} \Big) \\
= & \sum_{\ell=1}^{m} \Big(q'(s_\sigma, a_{\sigma\oplus\ell}) r(s_\sigma, a_{\sigma\oplus\ell}) - q''(s_\sigma, a_{\sigma\oplus\ell}) r(s_\sigma, a_{\sigma\oplus\ell}) \\
& + \left(q'(s_\sigma, a_{\sigma\oplus\ell}) p'(s_\sigma, a_{\sigma\oplus\ell}) - q''(s_\sigma, a_{\sigma\oplus\ell}) p''(s_\sigma, a_{\sigma\oplus\ell}) \right)^\top V_{h+1}'^{\pi} \\
& + q''(s_\sigma, a_{\sigma\oplus\ell}) p''(s_\sigma, a_{\sigma\oplus\ell})^\top \left(V_{h+1}'^{\pi} - V_{h+1}''^{\pi} \right) \Big) \\
= & \sum_{\ell=1}^{m} \Big(q'(s_\sigma, a_{\sigma\oplus\ell}) r(s_\sigma, a_{\sigma\oplus\ell}) - q''(s_\sigma, a_{\sigma\oplus\ell}) r(s_\sigma, a_{\sigma\oplus\ell}) \\
& + \left(q'(s_\sigma, a_{\sigma\oplus\ell}) p'(s_\sigma, a_{\sigma\oplus\ell}) - q''(s_\sigma, a_{\sigma\oplus\ell}) p''(s_\sigma, a_{\sigma\oplus\ell}) \right)^\top V_{h+1}'^{\pi} \Big) \\
& + \sum_{\ell=1}^{m} \mathbb{E}_{q'',p'',\pi} \left[V_{h+1}'^{\pi}(s_{\sigma\oplus\ell}) - V_{h+1}''^{\pi}(s_{\sigma\oplus\ell}) | s_h = s \right] \\
= & \sum_{\sigma'=\emptyset}^{m} \sum_{\ell=1}^{m} \mathbb{E}_{q'',p'',\pi} \Big[q'(s_{\sigma\oplus\sigma'}, a_{\sigma\oplus\sigma'\oplus\ell}) r(s_{t\ell}, a_{\sigma\oplus\sigma'\oplus\ell}) - q''(s_{t\ell}, a_{\sigma\oplus\sigma'\oplus\ell}) r(s_{t\ell}, a_{\sigma\oplus\sigma'\oplus\ell})
\end{aligned}$$

$$+ \left(q'(s_{t\ell}, a_{\sigma\oplus\sigma'\oplus\ell})p'(s_{t\ell}, a_{\sigma\oplus\sigma'\oplus\ell}) - q''(s_{t\ell}, a_{\sigma\oplus\sigma'\oplus\ell})p''(s_{t\ell}, a_{\sigma\oplus\sigma'\oplus\ell})\right)^\top V'^{\pi}_{h+1}\Big]$$

$$+ \sum_{\sigma'=1^{\oplus 2}}^{m^{\oplus 2}} \mathbb{E}_{q'',p'',\pi}\left[V'^{\pi}_{h+2,\ell}(s_{\sigma'}) - V''^{\pi}_{h+2,\ell}(s_{\sigma'})|s_h\right]$$

$$= \sum_{\sigma'=\emptyset}^{m^{\oplus(H-h)}} \sum_{\ell=1}^{m} \mathbb{E}_{q'',p'',\pi}\Big[q'(s_{\sigma\oplus\sigma'}, a_{\sigma\oplus\sigma'\oplus\ell})r(s_{\sigma\oplus\sigma'}, a_{\sigma\oplus\sigma'\oplus\ell}) - q''(s_{\sigma\oplus\sigma'}, a_{\sigma\oplus\sigma'\oplus\ell})r(s_{\sigma\oplus\sigma'}, a_{\sigma\oplus\sigma'\oplus\ell})$$

$$+ \left(q'(s_{\sigma\oplus\sigma'}, a_{\sigma\oplus\sigma'\oplus\ell})p'(s_{\sigma\oplus\sigma'}, a_{\sigma\oplus\sigma'\oplus\ell}) - q''(s_{\sigma\oplus\sigma'}, a_{\sigma\oplus\sigma'\oplus\ell})p''(s_{\sigma\oplus\sigma'}, a_{\sigma\oplus\sigma'\oplus\ell})\right)^\top V'^{\pi}_{|\sigma\oplus\sigma'\oplus\ell|+1}\Big]$$

□

B.2. Proof of Lemma 4

Proof of Lemma 4. First, we prove the equality. This proof adapts the analysis of standard law of total variance in (Jin et al., 2018; Zanette & Brunskill, 2019) to branching RL.

$$\mathbb{E}_{q,p,\pi}\left[\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \mathrm{Var}_{s_{\sigma\oplus\ell}\sim q,p}\left(V^{\pi}_{|\sigma\oplus\ell|+1}(s_{\sigma\oplus\ell})|s_\sigma, a_{\sigma\oplus\ell}\right)\right]$$

$$=\mathbb{E}_{q,p,\pi}\left[\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \left(V^{\pi}_{|\sigma\oplus\ell|+1}(s_{\sigma\oplus\ell}) - q(s_\sigma, a_{\sigma\oplus\ell})p(s_\sigma, a_{\sigma\oplus\ell})^\top V^{\pi}_{|\sigma\oplus\ell|+1}\right)^2\right]$$

$$=\mathbb{E}_{q,p,\pi}\Bigg[\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \Bigg(q(s_\sigma, a_{\sigma\oplus\ell})r(s_\sigma, a_{\sigma\oplus\ell}) + V^{\pi}_{|\sigma\oplus\ell|+1}(s_{\sigma\oplus\ell})$$

$$- \left(q(s_\sigma, a_{\sigma\oplus\ell})r(s_\sigma, a_{\sigma\oplus\ell}) + q(s_\sigma, a_{\sigma\oplus\ell})p(s_\sigma, a_{\sigma\oplus\ell})^\top V^{\pi}_{|\sigma\oplus\ell|+1}\right)\Bigg)^2\Bigg]$$

$$\overset{(a)}{=}\mathbb{E}_{q,p,\pi}\Bigg[\Bigg(\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \Bigg(q(s_\sigma, a_{\sigma\oplus\ell})r(s_\sigma, a_{\sigma\oplus\ell}) + V^{\pi}_{|\sigma\oplus\ell|+1}(s_{\sigma\oplus\ell})$$

$$- \left(q(s_\sigma, a_{\sigma\oplus\ell})r(s_\sigma, a_{\sigma\oplus\ell}) + q(s_\sigma, a_{\sigma\oplus\ell})p(s_\sigma, a_{\sigma\oplus\ell})^\top V^{\pi}_{|\sigma\oplus\ell|+1}\right)\Bigg)\Bigg)^2\Bigg]$$

$$\overset{(b)}{=}\mathbb{E}_{q,p,\pi}\left[\left(\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} q(s_\sigma, a_{\sigma\oplus\ell})r(s_\sigma, a_{\sigma\oplus\ell}) + \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} V^{\pi}_{|\sigma\oplus\ell|+1}(s_{\sigma\oplus\ell}) - \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} V^{\pi}_{|\sigma|+1}(s_\sigma)\right)^2\right]$$

$$=\mathbb{E}_{q,p,\pi}\left[\left(\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} q(s_\sigma, a_{\sigma\oplus\ell})r(s_\sigma, a_{\sigma\oplus\ell}) - V^{\pi}_{1}(s_\emptyset)\right)^2\right]$$

Here (a) comes from the Markov property and that conditioning on the filtration of step h, the triggers and transitions of state-base action pairs at step $h+1$ are independent. (b) is due to that $V^{\pi}_{|\sigma|+1}(s_\sigma) = \sum_{\ell=1}^{m}\left(q(s_\sigma, a_{\sigma\oplus\ell})r(s_\sigma, a_{\sigma\oplus\ell}) + q(s_\sigma, a_{\sigma\oplus\ell})p(s_\sigma, a_{\sigma\oplus\ell})^\top V^{\pi}_{|\sigma\oplus\ell|+1}\right)$ and we can merge the m terms of state-base action value into $V^{\pi}_{|\sigma|+1}(s_\sigma)$.

Now, we prove the inequality.

$$\mathbb{E}_{q,p,\pi}\left[\left(\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} q(s_\sigma, a_{\sigma\oplus\ell})r(s_\sigma, a_{\sigma\oplus\ell}) - V^{\pi}_{1}(s_\emptyset)\right)^2\right]$$

$$
\begin{aligned}
&\leq \mathbb{E}_{q,p,\pi}\left[\left(\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m} q(s_\sigma, a_{\sigma\oplus\ell})\right)^2\right]\\
&= \mathbb{E}\left[\left(\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\Big(q(s_\sigma, a_{\sigma\oplus\ell})\mathbb{1}\left\{s_\sigma \neq s_\perp\right\} + q(s_\sigma, a_{\sigma\oplus\ell})\mathbb{1}\left\{s_\sigma = s_\perp\right\}\Big)\right)^2\right]\\
&\overset{(a)}{\leq} \mathbb{E}\left[\left(\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\frac{1}{m}\mathbb{1}\left\{s_\sigma \neq s_\perp\right\}\right)^2\right]\\
&= \mathbb{E}\left[\left(\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\mathbb{1}\left\{s_\sigma \neq s_\perp\right\}\right)^2\right]
\end{aligned}
$$

where (a) is due to Assumption 1 and $q(s_\perp, a) = 0$ for any $a \in A^{\texttt{univ}}$.

□

B.3. Proof of Lemma 5

Proof of Lemma 5. Under Assumption 1, to bound the total number of triggered (regular) states for any branching MDP and policy π, it suffices to bound it under a relaxed model $\mathcal{M}^*$ with $q(s,a) = q^* := \frac{1}{m}$ for all $(s,a) \in \mathcal{S} \setminus \{s_\perp\} \times A^{\texttt{univ}}$. Let ω_h denote the number of triggered states at each step h under $\mathcal{M}^*$, and $\omega := \sum_{h=1}^{H}\omega_h$. Below we prove that $\mathbb{E}[\omega] \leq H$ and $\mathbb{E}[\omega^2] \leq 3H^2$.

For $h = 1$, $\omega_h = 1$ deterministically. For $h \geq 2$, $\omega_h | \omega_{h-1} \sim \texttt{Binomial}(m\omega_{h-1}, q^*)$. According to the properties of Binomial distribution and $q^* := \frac{1}{m}$, for $h \geq 2$,

$$
\begin{aligned}
\mathbb{E}\left[\omega_h\right] &= mq^*\mathbb{E}\left[\omega_{h-1}\right] = 1,\\
\mathbb{E}\left[(\omega_h)^2\right] &= mq^*(1-q^*)\mathbb{E}\left[\omega_{h-1}\right] + m^2(q^*)^2\mathbb{E}\left[(\omega_{h-1})^2\right]\\
&= (1-q^*) + \mathbb{E}\left[(\omega_{h-1})^2\right] \leq h.
\end{aligned}
$$

Hence, we have that $\mathbb{E}[\omega] = \sum_{h=1}^{H}\mathbb{E}[\omega_h] = H$, and

$$
\begin{aligned}
\mathbb{E}[\omega^2] &= \sum_{h=1}^{H}\mathbb{E}[(\omega_h)^2] + 2\sum_{1<i,j<H}\mathbb{E}[\omega_i\omega_j]\\
&\leq \frac{H(H+1)}{2} + 2\sum_{1<i,j<H}\mathbb{E}[\omega_i\omega_j].
\end{aligned} \tag{8}
$$

Now, the challenge falls on how to bound $\mathbb{E}[\omega_i\omega_j]$ for any $1 < i, j < H$. Let W_σ be a Bernoulli random variable denoting whether state s_σ is triggered for any index string σ. Then, we can write $\mathbb{E}[\omega_i\omega_j]$ as

$$
\begin{aligned}
&\mathbb{E}\Big[\Big(\sum_{\sigma=1^{\oplus(i-1)}}^{m^{\oplus(i-1)}} W_\sigma\Big)\cdot\Big(\sum_{\sigma'=1^{\oplus(j-1)}}^{m^{\oplus(j-1)}} W_{\sigma'}\Big)\Big]\\
\overset{(a)}{=}& m^{i-1}\mathbb{E}\Big[W_{1^{\oplus(i-1)}}\Big(\sum_{\sigma'=1^{\oplus(j-1)}}^{m^{\oplus(j-1)}} W_{\sigma'}\Big)\Big]\\
=& m^{i-1}\mathbb{E}\Big[W_{1^{\oplus(i-1)}}\Big(\sum_{\substack{\sigma'=1^{\oplus(j-1)}\\ \sigma'\text{starts with } 1^{\oplus(i-1)}}}^{m^{\oplus(j-1)}} W_{\sigma'} + \sum_{\substack{\sigma'=1^{\oplus(j-1)}\\ \sigma'\text{does not start with } 1^{\oplus(i-1)}}}^{m^{\oplus(j-1)}} W_{\sigma'}\Big)\Big]\\
\overset{(b)}{\leq}& m^{i-1}\left(m^{j-i}(q^*)^{i-1+j-i} + m^{j-1}(q^*)^{i-1+j-1}\right)
\end{aligned}
$$

$$=2. \tag{9}$$

Here (a) comes from the symmetry of trajectory tree. (b) is due to that at step j, the children states of $s_{1\oplus(i-1)}$ have dependency on it and the other states are independent of it, and $\mathbb{E}[W_\sigma W_{\sigma'}] = \Pr[W_\sigma = 1, W_{\sigma'} = 1]$ for any σ, σ'. By plugging Eq. (9) into Eq. (8), we have $\mathbb{E}[\omega^2] \le \frac{H(H+1)}{2} + 4\frac{H(H-1)}{2} \le 3H^2$. Thus, we obtain Lemma 5. □

C. Proofs for Branching RL with Regret Minimization

In this section, we prove the regret upper bound for algorithm `BranchVI` (Theorem 6) and the regret lower bounds for Branching RL-RM in cases with Assumption 1 (Theorem 7) and without Assumption 1 (Theorem 2).

We first introduce some notations. Let Bernoulli random variable $X_{k\sigma\ell}(s,a)$ denote whether (s,a) was visited at indices σ and $\sigma \oplus \ell$, respectively, in episode k, and $w_{k\sigma\ell}(s,a) := \Pr[X_{k\sigma\ell}(s,a)]$. Let $X_k(s,a) := \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} X_{k\sigma\ell}(s,a)$ denote the number of times that (s,a) was visited in episode k, and $w_k(s,a) := \mathbb{E}[X_k(s,a)] = \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} w_{k\sigma\ell}(s,a)$. Let $n^k_{\sigma\ell}(s,a) := \sum_{k'<k} X_{k'\sigma\ell}(s,a)$ denote the cumulative number of times that (s,a) was visited at indices σ and $\sigma \oplus \ell$, respectively, up to episode k. Let $n^k(s,a) := \sum_{k'<k} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} X_{k\sigma\ell}(s,a)$ denote the *cumulative* number of times that (s,a) was visited up to episode k.

In the following proofs, we make the convention that when $m = 1$, $\frac{m^{H+1}-m}{m-1} := H$. Then, we have $X_k(s,a) \le m + m^2 + \cdots + m^H = \frac{m^{H+1}-m}{m-1}$ for any $m \ge 1$.

C.1. Proof of Regret Upper Bound

C.1.1. Concentration

In the following, we present several important concentration lemmas and define concentration events.

Lemma 10 (Concentration of Trigger)**.**

$$\Pr\left[\left|\hat{q}^k(s,a) - q(s,a)\right| \le 4\sqrt{\frac{\log\left(\frac{SNH(m^H \vee K)}{\delta'}\right)}{n^k(s,a)}},\ \forall (s,a) \in \mathcal{S} \setminus \{s_\perp\} \times A^{\texttt{univ}}, \forall k \in [K]\right] \ge 1 - 2\delta'$$

Proof of Lemma 10. Since $n^k(s,a) \le \frac{m^{H+1}-m}{m-1}K$, using the Hoeffding inequality with a union bound over (s,a) and $n^k(s,a)$, we have

$$\Pr\left[\left|\hat{q}^k(s,a) - q(s,a)\right| \le 2\sqrt{\frac{\log\left(\frac{SN}{\delta'} \cdot \frac{m^{H+1}-m}{m-1}K\right)}{n^k(s,a)}},\ \forall (s,a) \in \mathcal{S} \setminus \{s_\perp\} \times A^{\texttt{univ}}, \forall k \in [K]\right] \ge 1 - 2\delta'$$

If $\frac{m^{H+1}-m}{m-1} \le K$, then we have

$$\log\left(\frac{SN}{\delta'} \cdot \frac{m^{H+1}-m}{m-1}K\right) \le 2\log\left(\frac{SNHK}{\delta'}\right)$$

If $\frac{m^{H+1}-m}{m-1} \ge K$, then using $\frac{m^{H+1}-m}{m-1} \le Hm^{2H}$, we have

$$\begin{aligned}\log\left(\frac{SN}{\delta'} \cdot \frac{m^{H+1}-m}{m-1}K\right) \le& 2\log\left(\frac{SNHm^{2H}}{\delta'}\right)\\ \le& 4\log\left(\frac{SNHm^H}{\delta'}\right)\end{aligned}$$

Combining the above two cases, we have

$$\log\left(\frac{SN}{\delta'} \cdot \frac{m^{H+1}-m}{m-1}K\right) \le 4\log\left(\frac{SNH(m^H \vee K)}{\delta'}\right)$$

Therefore, we have

$$\Pr\left[\left|\hat{q}^k(s,a)-q(s,a)\right| \leq 4\sqrt{\frac{\log\left(\frac{SNH(m^H \vee K)}{\delta'}\right)}{n^k(s,a)}},\ \forall (s,a)\in\mathcal{S}\setminus\{s_\perp\}\times A^{\texttt{univ}}, \forall k\in[K]\right] \geq 1-2\delta'$$

□

Lemma 11 (Concentration of Triggered Transition)**.**

$$\Pr\Bigg[\left|\left(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)-q(s,a)p(\cdot|s,a)\right)^\top V^*_{h+1}\right| \leq 4\sqrt{\frac{\mathrm{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)\log\left(\frac{SNH(m^H\vee K)}{\delta'}\right)}{n^k(s,a)}}$$
$$+\frac{4H\log\left(\frac{SNH(m^H\vee K)}{\delta'}\right)}{n^k(s,a)},\ \forall (s,a)\in\mathcal{S}\setminus\{s_\perp\}\times A^{\texttt{univ}}, \forall k\in[K]\Bigg] \geq 1-2\delta'$$

Proof of Lemma 11. Using the similar analytical procedure as Lemma 10 and the Bernstein's inequality, we can obtain this lemma. □

Lemma 12 (Concentration of Variance)**.**

$$\Pr\Bigg[\left|\sqrt{\mathrm{Var}_{s'\sim\hat{q},\hat{p}}\left(\bar{V}^k_{h+1}(s')\right)}-\sqrt{\mathrm{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)}\right| \leq \sqrt{\mathbb{E}_{s'\sim\hat{q},\hat{p}}\left[\left(\bar{V}^k_{h+1}(s')-V^*_{h+1}(s')\right)^2\right]}$$
$$+8H\sqrt{\frac{\log\left(\frac{SNH(m^H\vee K)}{\delta'}\right)}{n^k(s,a)}},\ \forall (s,a)\in\mathcal{S}\setminus\{s_\perp\}\times A^{\texttt{univ}}, \forall k\in[K]\Bigg] \geq 1-2\delta' \tag{10}$$

Proof of Lemma 12. Using the similar analytical procedure as Lemma 10 and Proposition 2 (in particular, Eq. (53)) in (Zanette & Brunskill, 2019), we can obtain

$$\Pr\Bigg[\left|\sqrt{\mathrm{Var}_{s'\sim\hat{q},\hat{p}}\left(V^*_{h+1}(s')\right)}-\sqrt{\mathrm{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)}\right| \leq 8H\sqrt{\frac{\log\left(\frac{SNH(m^H\vee K)}{\delta'}\right)}{n^k(s,a)}},$$
$$\forall (s,a)\in\mathcal{S}\setminus\{s_\perp\}\times A^{\texttt{univ}}, \forall k\in[K]\Bigg] \geq 1-2\delta' \tag{11}$$

With probability $1-2\delta'$, for any $(s,a)\in\mathcal{S}\setminus\{s_\perp\}\times A^{\texttt{univ}}$ and $k\in[K]$, we have

$$\begin{aligned}
&\left|\sqrt{\mathrm{Var}_{s'\sim\hat{q},\hat{p}}\left(\bar{V}^k_{h+1}(s')\right)}-\sqrt{\mathrm{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)}\right| \\
\leq&\left|\sqrt{\mathrm{Var}_{s'\sim\hat{q},\hat{p}}\left(\bar{V}^k_{h+1}(s')\right)}-\sqrt{\mathrm{Var}_{s'\sim\hat{q},\hat{p}}\left(V^*_{h+1}(s')\right)}\right| \\
&+\left|\sqrt{\mathrm{Var}_{s'\sim\hat{q},\hat{p}}\left(V^*_{h+1}(s')\right)}-\sqrt{\mathrm{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)}\right| \\
\overset{(a)}{\leq}&\sqrt{\mathrm{Var}_{s'\sim\hat{q},\hat{p}}\left(\bar{V}^k_{h+1}(s')-V^*_{h+1}(s')\right)}+8H\sqrt{\frac{\log\left(\frac{SNH(m^H\vee K)}{\delta'}\right)}{n^k(s,a)}} \\
\leq&\sqrt{\mathbb{E}_{s'\sim\hat{q},\hat{p}}\left[\left(\bar{V}^k_{h+1}(s')-V^*_{h+1}(s')\right)^2\right]}+8H\sqrt{\frac{\log\left(\frac{SNH(m^H\vee K)}{\delta'}\right)}{n^k(s,a)}}
\end{aligned}$$

where (a) uses Proposition 2 (in particular, Eqs. (52)) in (Zanette & Brunskill, 2019) and Eq. (11).

□

To summarize the concentration lemmas used, we define the following concentration events:

$$
\begin{aligned}
\mathcal{E}_{\text{tri}} :=& \left[\left|\hat{q}^k(s,a) - q(s,a)\right| \le 4\sqrt{\frac{\log\left(\frac{SNH(m^H \vee K)}{\delta'}\right)}{n^k(s,a)}},\ \forall (s,a) \in \mathcal{S}\setminus\{s_\perp\} \times A^{\mathtt{univ}}, \forall k \in [K] \right] \\
\mathcal{E}_{\text{trans}} :=& \Bigg[\left| \left(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a) - q(s,a)p(\cdot|s,a)\right)^\top V^*_{h+1} \right| \le 4\sqrt{\frac{\mathrm{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)\log\left(\frac{SNH(m^H \vee K)}{\delta'}\right)}{n^k(s,a)}} \\
&+ 4\frac{H\log\left(\frac{SNH(m^H \vee K)}{\delta'}\right)}{n^k(s,a)},\ \forall (s,a) \in \mathcal{S}\setminus\{s_\perp\} \times A^{\mathtt{univ}}, \forall k \in [K] \Bigg] \\
\mathcal{E}_{\text{var}} :=& \Bigg[\left| \sqrt{\mathrm{Var}_{s'\sim \hat{q},\hat{p}}\left(\bar{V}^k_{h+1}(s')\right)} - \sqrt{\mathrm{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)} \right| \le \sqrt{\mathbb{E}_{s'\sim\hat{q},\hat{p}}\left[\bar{V}^k_{h+1}(s') - V^*_{h+1}(s')\right]^2} \\
&+ 8H\sqrt{\frac{\log\left(\frac{SNH(m^H \vee K)}{\delta'}\right)}{n^k(s,a)}},\ \forall (s,a) \in \mathcal{S}\setminus\{s_\perp\} \times A^{\mathtt{univ}}, \forall k \in [K] \Bigg] \\
\mathcal{E} :=& \mathcal{E}_{\text{tri}} \cap \mathcal{E}_{\text{trans}} \cap \mathcal{E}_{\text{var}}
\end{aligned}
$$

Lemma 13. *Letting $\delta' := \frac{\delta}{6}$, the concentration event $\mathcal{E}$ satisfies that*

$$
\Pr\left[\mathcal{E}\right] \ge 1 - 6\delta' = 1 - \delta
$$

Proof of Lemma 13. We can obtain this lemma by combining Lemmas 10-12. □

C.1.2. Visitation

Below we present an important bound on visitation, which will be used in the proof of Theorem 6.

Lemma 14 (Regret Bound of Visitation)**.** *Suppose that the concentration event $\mathcal{E}$ holds. Then, it holds that*

$$
\sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a), s\neq s_\perp} \frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)} \le SN\log\left(KH\right)
$$

Proof of Lemma 14.

$$
\begin{aligned}
\sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a), s\neq s_\perp} \frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)} =& \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \mathbb{E}_{(s_\sigma, a_{\sigma\oplus\ell})\sim\pi^k}\left[\frac{1}{n^k(s_\sigma, a_{\sigma\oplus\ell})} \cdot \mathbb{1}\left\{s_\sigma \neq s_\perp\right\} \right] \\
=& \sum_{k=1}^{K} \mathbb{E}_{X_k\sim\pi^k}\left[\sum_{(s,a), s\neq s_\perp} X_k(s,a) \frac{1}{n^k(s,a)} \right] \\
=& \mathbb{E}_{X_k\sim\pi^k}\left[\sum_{(s,a), s\neq s_\perp} \sum_{k=1}^{K} X_k(s,a) \frac{1}{n^k(s,a)} \right] \\
=& \mathbb{E}_{X_k\sim\pi^k}\left[\sum_{(s,a), s\neq s_\perp} \sum_{k=1}^{K} X_k(s,a) \frac{1}{\sum_{k'<k} X_{k'}(s,a)} \right] \\
\le& \mathbb{E}_{X_k\sim\pi^k}\left[\sum_{(s,a), s\neq s_\perp} \log\left(\sum_{k=1}^{K} X_k(s,a) \right) \right]
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(a)}{\le} \sum_{(s,a),s\neq s_\perp} \log\left(\mathbb{E}_{X_k\sim\pi^k}\left[\sum_{k=1}^{K} X_k(s,a)\right]\right)\\
&\overset{(b)}{\le} SN\log\left(KH\right),
\end{aligned}
$$

where (a) uses Jensen's inequality, and (b) is due to that for a fixed (s,a) such that $s \neq s_\perp$, $\mathbb{E}[X_k(s,a)] \le \mathbb{E}[\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \mathbb{1}\left\{s_\sigma \neq s_\perp\right\}] \le H$, since a single base action cannot be chosen twice in a state.

□

C.1.3. Optimism and Pessimism

Next, we prove the optimism and pessimism of the constructed value functions in algorithm `BranchVI`, and bound the gap between optimistic and pessimistic value functions. Recall that $L := \log\left(\frac{SNH(m^H \vee K)}{\delta'}\right)$.

Lemma 15 (Optimism)**.** *Suppose that the concentration event $\mathcal{E}$ holds. Then,*

$$
\underline{V}_h^k(s) \le V_h^*(s) \le \bar{V}_h^k(s),\ \forall s\in\mathcal{S}, h\in[H], k\in[K]
$$

Proof of Lemma 15. We prove this lemma by induction. Since $\underline{V}_{H+1}^k(s) = V_{H+1}^*(s) = \bar{V}_{H+1}^k(s) = 0, \forall s \in \mathcal{S}$, it suffices to prove that if $\underline{V}_{h+1}^k(s) \le V_h^*(s) \le \bar{V}_{h+1}^k(s), \forall s\in\mathcal{S}$, then $\underline{V}_h^k(s) \le V_h^*(s) \le \bar{V}_h^k(s), \forall s\in\mathcal{S}$.

First, we prove the optimistic direction, i.e., $\bar{V}_h^k(s) \ge V_h^*(s),\ \forall s\in\mathcal{S}$. In the following, we prove $\bar{Q}_h^{\pi^k}(s,A) \ge Q_h^*(s,A)$ for any $s\in\mathcal{S}$ and $A\in\mathcal{A}$. If $\bar{Q}_h^{\pi^k}(s,A) = H$, then $\bar{Q}_h^{\pi^k}(s,A) = H \ge Q_h^*(s,A)$ trivially holds. Otherwise,

$$
\begin{aligned}
\bar{Q}_h^{\pi^k}(s,A) \ge& \sum_{a\in A}\left(\left(\hat{q}^k(s,a)+b_k^q(s,a)\right)r(s,a)+\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \bar{V}_{h+1}^k + b_k^{qpV}(s,a)\right)\\
=& \sum_{a\in A}\Bigg(\left(\hat{q}^k(s,a)+4\sqrt{\frac{L}{n^k(s,a)}}\right)r(s,a)+\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \bar{V}_{h+1}^k + 4\sqrt{\frac{\mathrm{Var}_{s'\sim\hat{q},\hat{p}}\left(\bar{V}_{h+1}^k(s')\right)L}{n^k(s,a)}}\\
&+4\sqrt{\frac{\mathbb{E}_{s'\sim\hat{q},\hat{p}}\left[\left(\bar{V}_{h+1}^k(s')-\underline{V}_{h+1}^k(s')\right)^2\right]L}{n^k(s,a)}}+\frac{36HL}{n^k(s,a)}\Bigg)\\
\ge& \sum_{a\in A}\Bigg(q(s,a)r(s,a)+\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \bar{V}_{h+1}^k + 4\sqrt{\frac{L}{n^k(s,a)}}\Bigg(\sqrt{\mathrm{Var}_{s'\sim\hat{q},\hat{p}}\left(\bar{V}_{h+1}^k(s')\right)}\\
&+\sqrt{\mathbb{E}_{s'\sim\hat{q},\hat{p}}\left[\left(\bar{V}_{h+1}^k(s')-\underline{V}_{h+1}^k(s')\right)^2\right]}+8H\sqrt{\frac{L}{n^k(s,a)}}\Bigg)+\frac{4HL}{n^k(s,a)}\Bigg)\\
\overset{(a)}{\ge}& \sum_{a\in A}\left(q(s,a)r(s,a)+\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top V_{h+1}^* + 4\sqrt{\frac{\mathrm{Var}_{s'\sim q,p}\left(V_{h+1}^*(s')\right)L}{n^k(s,a)}}+\frac{4HL}{n^k(s,a)}\right)\\
\overset{(b)}{\ge}& \sum_{a\in A}\left(q(s,a)r(s,a)+q(s,a)p(\cdot|s,a)^\top V_{h+1}^*\right)\\
=&Q_h^*(s,A),
\end{aligned}
$$

where (a) is due to the induction hypothesis and Lemma 12, and (b) comes from Lemma 11.

Then, we have

$$
\bar{V}_h^k(s) = \bar{Q}_h^k(s,\pi^k(s)) \ge \bar{Q}_h^k(s,\pi^*(s)) \ge Q_h^*(s,\pi^*(s)) = V_h^*(s)
$$

Now, we prove the pessimistic direction, i.e., $\underline{V}_h^k(s) \leq V_h^*(s),\ \forall s \in \mathcal{S}$. If $\underline{V}_h^k(s) = 0$, then $\underline{V}_h^k(s) = 0 \leq V_h^*(s)$ trivially holds. Otherwise,

$$\begin{aligned}
\underline{V}_h^k(s) =& \sum_{a \in A} \left(\left(\hat{q}^k(s,a) - b_k^q(s,a) \right) r(s,a) + \hat{q}^k(s,a) \hat{p}^k(\cdot|s,a)^\top \underline{V}_{h+1}^k - b_k^{qpV}(s,a) \right) \\
=& \sum_{a \in \pi^k(s)} \Bigg(\left(\hat{q}^k(s,a) - 4\sqrt{\frac{L}{n^k(s,a)}} \right) r(s,a) + \hat{q}^k(s,a) \hat{p}^k(\cdot|s,a)^\top \underline{V}_{h+1}^k - 4\sqrt{\frac{\mathrm{Var}_{s' \sim \hat{q},\hat{p}} \left(\bar{V}_{h+1}^k(s') \right) L}{n^k(s,a)}} \\
& - 4\sqrt{\frac{\mathbb{E}_{s' \sim \hat{q},\hat{p}} \left[\left(\bar{V}_{h+1}^k(s') - \underline{V}_{h+1}^k(s') \right)^2 \right] L}{n^k(s,a)}} - \frac{36HL}{n^k(s,a)} \Bigg) \\
\leq& \sum_{a \in \pi^k(s)} \Bigg(q(s,a) r(s,a) + \hat{q}^k(s,a) \hat{p}^k(\cdot|s,a)^\top \underline{V}_{h+1}^k - 4\sqrt{\frac{L}{n^k(s,a)}} \Bigg(\sqrt{\mathrm{Var}_{s' \sim \hat{q},\hat{p}} \left(\bar{V}_{h+1}^k(s') \right)} \\
& + \sqrt{\mathbb{E}_{s' \sim \hat{q},\hat{p}} \left[\left(\bar{V}_{h+1}^k(s') - \underline{V}_{h+1}^k(s') \right)^2 \right]} + 8H\sqrt{\frac{L}{n^k(s,a)}} \Bigg) - \frac{4HL}{n^k(s,a)} \Bigg) \\
\leq& \sum_{a \in \pi^k(s)} \left(q(s,a) r(s,a) + \hat{q}^k(s,a) \hat{p}^k(\cdot|s,a)^\top V_{h+1}^* - 4\sqrt{\frac{\mathrm{Var}_{s' \sim q,p} \left(V_{h+1}^*(s') \right) L}{n^k(s,a)}} - \frac{4HL}{n^k(s,a)} \right) \\
\leq& \sum_{a \in \pi^k(s)} \left(q(s,a) r(s,a) + q(s,a) p(\cdot|s,a)^\top V_{h+1}^* \right) \\
=& Q_h^*(s, \pi^k(s)) \\
\leq& Q_h^*(s, \pi^*(s)) \\
=& V_h^*(s)
\end{aligned}$$

□

Lemma 16 (Gap between Optimism and Pessimism)**.** *Suppose that the concentration event $\mathcal{E}$ holds. Then, it holds that*

$$\bar{V}_h^{\pi^k}(s) \leq \sum_{\sigma'=\emptyset}^{m^{\oplus(H-h)}} \sum_{\ell=1}^{m} \mathbb{E} \left[170HL\sqrt{\frac{S}{n^k(s_{\sigma \oplus \sigma'}, a_{\sigma \oplus \sigma' \oplus \ell})}} \cdot \mathbb{1}\left\{ s_{\sigma \oplus \sigma'} \neq s_\perp \right\} \middle| s_\sigma = s, \pi^k \right].$$

In particular,

$$\bar{V}_1^{\pi^k}(s) - \underline{V}_1^{\pi^k}(s) \leq \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \mathbb{E} \left[170HL\sqrt{\frac{S}{n^k(s_\sigma, a_{\sigma \oplus \ell})}} \cdot \mathbb{1}\left\{ s_\sigma \neq s_\perp \right\} \middle| s_\emptyset = s, \pi^k \right]$$

Proof of Lemma 16. According to the construction of optimistic and pessimistic value functions, we have

$$\begin{cases} \bar{V}_h^{\pi^k}(s) \leq \sum_{a \in A} \left(\left(\hat{q}^k(s,a) + b_k^q(s,a) \right) r(s,a) + \hat{q}^k(s,a) \hat{p}^k(\cdot|s,a)^\top \bar{V}_{h+1}^k + b_k^{qpV}(s,a) \right) \\ \underline{V}_h^{\pi^k}(s) \geq \sum_{a \in A} \left(\left(\hat{q}^k(s,a) - b_k^q(s,a) \right) r(s,a) + \hat{q}^k(s,a) \hat{p}^k(\cdot|s,a)^\top \underline{V}_{h+1}^k - b_k^{qpV}(s,a) \right) \end{cases} \tag{12}$$

Then,

$$\begin{aligned}
\bar{V}_h^{\pi^k}(s) - \underline{V}_h^{\pi^k}(s) \leq& \sum_{a \in A} \left(2b_k^q(s,a) r(s,a) + \hat{q}^k(s,a) \hat{p}^k(\cdot|s,a)^\top \left(\bar{V}_{h+1}^k - \underline{V}_{h+1}^k \right) + 2b_k^{qpV}(s,a) \right) \\
\leq& \sum_{a \in A} \left(2b_k^q(s,a) r(s,a) + 2b_k^{qpV}(s,a) + \hat{q}^k(s,a) \hat{p}^k(\cdot|s,a)^\top \left(\bar{V}_{h+1}^k - \underline{V}_{h+1}^k \right) \right)
\end{aligned}$$

$$
\begin{aligned}
&\le \sum_{a \in A} \Bigg(8\sqrt{\frac{L}{n^k(s,a)}} r(s,a) + 8\sqrt{\frac{\mathrm{Var}_{s' \sim \hat{q},\hat{p}} \left(\bar{V}^k_{h+1}(s')\right) L}{n^k(s,a)}} \\
&\quad + 8\sqrt{\frac{\mathbb{E}_{s' \sim \hat{q},\hat{p}} \left[\left(\bar{V}^k_{h+1}(s') - \underline{V}^k_{h+1}(s') \right)^2 \right] L}{n^k(s,a)}} + \frac{72HL}{n^k(s,a)} + \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \left(\bar{V}^k_{h+1} - \underline{V}^k_{h+1} \right) \Bigg) \\
&\le \sum_{a \in A} \Bigg(8\sqrt{\frac{L}{n^k(s,a)}} r(s,a) + 8\sqrt{\frac{\mathrm{Var}_{s' \sim q,p} \left(V^*_{h+1}(s')\right) L}{n^k(s,a)}} \\
&\quad + 16\sqrt{\frac{\mathbb{E}_{s' \sim \hat{q},\hat{p}} \left[\left(\bar{V}^k_{h+1}(s') - \underline{V}^k_{h+1}(s') \right)^2 \right] L}{n^k(s,a)}} + \frac{136HL}{n^k(s,a)} + q(s,a)p(\cdot|s,a)^\top \left(\bar{V}^k_{h+1} - \underline{V}^k_{h+1} \right) \\
&\quad + \left(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a) - q(s,a)p(\cdot|s,a) \right)^\top \left(\bar{V}^k_{h+1} - \underline{V}^k_{h+1} \right) \Bigg) \\
&\le \sum_{a \in A} \Bigg(8\sqrt{\frac{L}{n^k(s,a)}} r(s,a) + 8\sqrt{\frac{\mathrm{Var}_{s' \sim q,p} \left(V^*_{h+1}(s')\right) L}{n^k(s,a)}} \\
&\quad + 16\sqrt{\frac{\mathbb{E}_{s' \sim \hat{q},\hat{p}} \left[\left(\bar{V}^k_{h+1}(s') - \underline{V}^k_{h+1}(s') \right)^2 \right] L}{n^k(s,a)}} + \frac{136HL}{n^k(s,a)} + q(s,a)p(\cdot|s,a)^\top \left(\bar{V}^k_{h+1} - \underline{V}^k_{h+1} \right) \\
&\quad + \| \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a) - q(s,a)p(\cdot|s,a) \|_1 \| \bar{V}^k_{h+1} - \underline{V}^k_{h+1} \|_\infty \Bigg) \\
&\le \sum_{a \in A} \Bigg(8\sqrt{\frac{L}{n^k(s,a)}} r(s,a) + 8\sqrt{\frac{\mathrm{Var}_{s' \sim q,p} \left(V^*_{h+1}(s')\right) L}{n^k(s,a)}} \\
&\quad + 16\sqrt{\frac{\mathbb{E}_{s' \sim \hat{q},\hat{p}} \left[\left(\bar{V}^k_{h+1}(s') - \underline{V}^k_{h+1}(s') \right)^2 \right] L}{n^k(s,a)}} + \frac{136HL}{n^k(s,a)} + q(s,a)p(\cdot|s,a)^\top \left(\bar{V}^k_{h+1} - \underline{V}^k_{h+1} \right) \\
&\quad + H\sqrt{\frac{2SL}{n^k(s,a)}} \Bigg) \\
&\le \sum_{a \in A} \left(170HL\sqrt{\frac{S}{n^k(s,a)}} + q(s,a)p(\cdot|s,a)^\top \left(\bar{V}^k_{h+1} - \underline{V}^k_{h+1} \right) \right) \\
&\le \sum_{\sigma'=\emptyset}^{m^{\oplus(H-h)}} \sum_{\ell=1}^{m} \mathbb{E} \left[170HL\sqrt{\frac{S}{n^k(s_{\sigma\oplus\sigma'}, a_{\sigma\oplus\sigma'\oplus\ell})}} \cdot \mathbb{1}\left\{ s_{\sigma\oplus\sigma'} \neq s_\perp \right\} \Big| s_\sigma = s, \pi^k \right]
\end{aligned}
$$

Thus,

$$
\bar{V}^{\pi^k}_1(s) - \underline{V}^{\pi^k}_1(s) \le \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \mathbb{E} \left[170HL\sqrt{\frac{S}{n^k(s_\sigma, a_{\sigma\oplus\ell})}} \cdot \mathbb{1}\left\{ s_\sigma \neq s_\perp \right\} \Big| s_\emptyset = s, \pi^k \right]
$$

□

Lemma 17 (Cumulative Gap between Optimism and Pessimism)**.** *Suppose that the concentration event $\mathcal{E}$ holds. Then, it holds that*

$$
\sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a), s \neq s_\perp} w_{k\sigma\ell}(s,a) \sum_{s'} q(s,a)p(s'|s,a) \left(\bar{V}^k_{h+1}(s') - \underline{V}^k_{h+1}(s') \right)^2 \le 28900 m H^4 L^3 S^3 N^2
$$

Proof of Lemma 17.

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\sum_{s'\neq s_\perp} q(s,a)p(s'|s,a)\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2\\
=&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\sum_{s'\neq s_\perp} w_{k\sigma\ell}(s,a)q(s,a)p(s'|s,a)\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2\\
=&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\sum_{s'\neq s_\perp} w_{k\sigma\ell}(s',s,a)\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2\\
=&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{s'\neq s_\perp} \tilde{w}_{k\sigma\ell}(s')\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2\\
=&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\mathbb{E}_{s_{\sigma\oplus\ell}\sim\pi^k}\left[\left(\bar{V}^k_{h+1}(s_{\sigma\oplus\ell})-\underline{V}^k_{h+1}(s_{\sigma\oplus\ell})\right)^2\right]\\
\leq&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\mathbb{E}_{s_\sigma\sim\pi^k}\left[\left(\bar{V}^k_{h}(s_\sigma)-\underline{V}^k_{h}(s_\sigma)\right)^2\right]\\
=&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\mathbb{E}_{s_\sigma\sim\pi^k}\left[\left(\left(\bar{V}^k_{h}(s_\sigma)-\underline{V}^k_{h}(s_\sigma)\right)\cdot\mathbb{1}\left\{s_\sigma\neq s_\perp\right\}\right)^2\right]\\
\overset{(a)}{\leq}&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\mathbb{E}_{s_\sigma\sim\pi^k}\left[\left(\mathbb{1}\left\{s_\sigma\neq s_\perp\right\}\cdot\sum_{\sigma'=\emptyset}^{m^{\oplus(H-h)}}\sum_{\ell=1}^{m}\mathbb{E}\left[170HL\sqrt{\frac{S}{n^k(s_{\sigma\oplus\sigma'},a_{\sigma\oplus\sigma'\oplus\ell})}}\cdot\mathbb{1}\left\{s_{\sigma\oplus\sigma'}\neq s_\perp\right\}\Big|s_\sigma,\pi^k\right]\right)^2\right]\\
\leq&28900H^2L^2\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\mathbb{E}_{s_\sigma\sim\pi^k}\left[\left(\mathbb{1}\left\{s_\sigma\neq s_\perp\right\}\cdot\mathbb{E}\left[\sum_{(s,a),s\neq s_\perp}X_k(s,a)\sqrt{\frac{S}{n^k(s,a)}}\right]\right)^2\right]\\
=&28900H^2L^2\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\mathbb{E}_{s_\sigma\sim\pi^k}\left[\left(\mathbb{1}\left\{s_\sigma\neq s_\perp\right\}\right)^2\cdot\left(\mathbb{E}\left[\sum_{(s,a),s\neq s_\perp}X_k(s,a)\sqrt{\frac{S}{n^k(s,a)}}\right]\right)^2\right]\\
=&28900H^2L^2\sum_{k=1}^{K}\left(\mathbb{E}\left[\sum_{(s,a),s\neq s_\perp}X_k(s,a)\sqrt{\frac{S}{n^k(s,a)}}\right]\right)^2\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\mathbb{E}_{s_\sigma\sim\pi^k}\left[\left(\mathbb{1}\left\{s_\sigma\neq s_\perp\right\}\right)^2\right]\\
\leq&28900H^3L^2\sum_{k=1}^{K}\left(\mathbb{E}\left[\sum_{(s,a),s\neq s_\perp}X_k(s,a)\right]\right)^2\frac{S}{n^k(s,a)}\\
\overset{(b)}{\leq}&28900mH^4L^2\sum_{k=1}^{K}\mathbb{E}\left[\sum_{(s,a),s\neq s_\perp}X_k(s,a)\right]\frac{S}{n^k(s,a)}\\
=&28900mH^4L^2S\,\mathbb{E}\left[\sum_{(s,a),s\neq s_\perp}\sum_{k=1}^{K}X_k(s,a)\frac{1}{n^k(s,a)}\right]\\
\overset{(c)}{\leq}&28900mH^4L^3S^2N
\end{aligned}
$$

where (a) uses Lemma 16, (b) is due to $\mathbb{E}\left[\sum_{(s,a),s\neq s_\perp}X_k(s,a)\right]\leq m\mathbb{E}[\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\mathbb{1}\left\{s_\sigma\neq s_\perp\right\}]\leq mH$ and (c) comes from Lemma 14.

□

C.1.4. Proof of Theorem 6

Now we prove the regret upper bound (Theorem 6) for algorithm `BranchVI`.

Proof of Theorem 6. Suppose that the concentration event $\mathcal{E}$ holds.

For any $k \in [K]$ and $(s,a) \in \mathcal{S} \setminus \{s_\perp\} \times A^{\texttt{univ}}$, let $\tilde{q}^k(s,a) := \hat{q}^k(s,a) + b^q_k(s,a)$, $\tilde{q}^k(s,a)\tilde{p}^k(\cdot|s,a)^\top \bar{V}^k_{h+1} := \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \bar{V}^k_{h+1} + b^{qpV}_k(s,a)$.

Step 1: Regret decomposition. Using Lemma 3, we can decompose $\texttt{Regret}(K)$ as follows:

$$
\begin{aligned}
\texttt{Regret}(K) =& \sum_{k=1}^{K} \left(V_1^*(s) - V_1^{\pi^k}(s) \right) \\
=& \sum_{k=1}^{K} \left(\bar{V}_1^k(s) - V_1^{\pi^k}(s) \right) \\
=& \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \mathbb{E}\Big[\left(\tilde{q}^k(s_\sigma, a_{\sigma\oplus\ell}) - q(s_\sigma, a_{\sigma\oplus\ell}) \right) r(s_\sigma, a_{\sigma\oplus\ell}) \\
& + \left(\tilde{q}^k(s_\sigma, a_{\sigma\oplus\ell}) \tilde{p}^k(\cdot|s_\sigma, a_{\sigma\oplus\ell}) - q(s_\sigma, a_{\sigma\oplus\ell}) p(\cdot|s_\sigma, a_{\sigma\oplus\ell}) \right)^\top \bar{V}^k_{h+1} \Big] \\
=& \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a), s\neq s_\perp} w_{k\sigma\ell}(s,a) \cdot \\
& \Big[\left(\tilde{q}^k(s,a) - q(s,a) \right) r(s,a) + \left(\tilde{q}^k(s,a)\tilde{p}^k(\cdot|s,a) - q(s,a)p(\cdot|s,a) \right)^\top \bar{V}^k_{h+1} \Big] \\
=& \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a), s\neq s_\perp} w_{k\sigma\ell}(s,a) \Big[\underbrace{\left(\tilde{q}^k(s,a) - q(s,a) \right) r(s,a)}_{\texttt{Term 1}} \\
& + \underbrace{\left(\tilde{q}^k(s,a)\tilde{p}^k(\cdot|s,a) - \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a) \right)^\top \bar{V}^k_{h+1}}_{\texttt{Term 2}} + \underbrace{\left(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a) - q(s,a)p(\cdot|s,a) \right)^\top V^*_{h+1}}_{\texttt{Term 3}} \\
& + \underbrace{\left(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a) - q(s,a)p(\cdot|s,a) \right)^\top \left(\bar{V}^k_{h+1} - V^*_{h+1} \right)}_{\texttt{Term 4}} \Big]
\end{aligned}
\tag{13}
$$

Step 2: Bound the bonus term for triggered rewards – `Term 1`.

$$
\begin{aligned}
\texttt{Term 1} =& \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a), s\neq s_\perp} w_{k\sigma\ell}(s,a) \left(\left(\tilde{q}^k(s,a) - q(s,a) \right) r(s,a) \right) \\
=& \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a), s\neq s_\perp} w_{k\sigma\ell}(s,a) \left(\left(\hat{q}^k(s,a) + 4\sqrt{\frac{L}{n^k(s,a)}} - q(s,a) \right) r(s,a) \right) \\
=& 8 \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a), s\neq s_\perp} w_{k\sigma\ell}(s,a) \sqrt{\frac{L}{n^k(s,a)}} r(s,a) \\
\leq& 8\sqrt{L} \sqrt{ \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a), s\neq s_\perp} \frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)} } \sqrt{ \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a), s\neq s_\perp} w_{k\sigma\ell}(s,a) r^2(s,a) }
\end{aligned}
$$

$$\begin{aligned}
&\le 8L\sqrt{SN\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\mathbb{E}\left[\mathbb{1}\left\{s^k_{\sigma\ell}\neq s_\perp\right\}\right]}\\
&\overset{(a)}{\le} 8L\sqrt{mSNHK},
\end{aligned} \tag{14}$$

where (a) uses Lemma 5.

Step 3: Bound the bonus term for triggered future values – `Term 2`.

$$\begin{aligned}
\texttt{Term 2} =& \sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\left(\tilde{q}^k(s,a)\tilde{p}^k(\cdot|s,a)-\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)\right)^\top \bar{V}^k_{h+1}\\
=&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\Bigg(4\sqrt{\frac{\text{Var}_{s'\sim\hat{q},\hat{p}}\left(\bar{V}^k_{h+1}(s')\right)L}{n^k(s,a)}}+4\sqrt{\frac{\mathbb{E}_{s'\sim\hat{q},\hat{p}}\left[\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2\right]L}{n^k(s,a)}}\\
&+\frac{36HL}{n^k(s,a)}\Bigg)\\
\overset{\text{(a)}}{\le}&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\Bigg(4\sqrt{\frac{\text{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)L}{n^k(s,a)}}+8\sqrt{\frac{\mathbb{E}_{s'\sim\hat{q},\hat{p}}\left[\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2\right]L}{n^k(s,a)}}\\
&+\frac{68HL}{n^k(s,a)}\Bigg)\\
=&4\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\sqrt{\frac{\text{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)L}{n^k(s,a)}}+68HL\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)}\\
&+8\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\sqrt{\frac{\mathbb{E}_{s'\sim\hat{q},\hat{p}}\left[\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2\right]L}{n^k(s,a)}}\\
\le&4\sqrt{L}\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)}}\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\text{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)}\\
&+68HL\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)}\\
&+8\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\sqrt{\frac{\mathbb{E}_{s'\sim\hat{q},\hat{p}}\left[\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2\right]L}{n^k(s,a)}}\\
\overset{(b)}{\le}&4L\sqrt{SN\sum_{k=1}^{K}\mathbb{E}_{\pi^k}\left[\sum_{h=1}^{H}\sum_{\ell=1}^{m^h}\text{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)\right]}+68SNHL^2\\
&+8\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\sqrt{\frac{\mathbb{E}_{s'\sim\hat{q},\hat{p}}\left[\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2\right]L}{n^k(s,a)}}
\end{aligned}$$

$$\overset{(c)}{\leq} 4L\sqrt{3SNKH^2} + 68SNHL^2 + 8\underbrace{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\sqrt{\frac{\mathbb{E}_{s'\sim\hat{q},\hat{p}}\left[\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2\right]L}{n^k(s,a)}}}_{\texttt{Term 2.1}}, \tag{15}$$

where (a) uses Lemma 12, (b) comes from Lemma 14 and (c) is due to Lemma 4.

Then, we bound `Term 2.1` as follows:

$$\begin{aligned}
&\texttt{Term 2.1} \qquad (16)\\
=&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\sqrt{\frac{\mathbb{E}_{s'\sim\hat{q},\hat{p}}\left[\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2\right]L}{n^k(s,a)}}\\
\leq&\sqrt{L}\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)}}\cdot\\
&\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2}\\
\leq&\sqrt{L}\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)}}\cdot\\
&\Bigg(\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)q(s,a)p(\cdot|s,a)^\top\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2}+\\
&\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\left(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)-q(s,a)p(\cdot|s,a)\right)^\top\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2}\Bigg)\\
\leq&\sqrt{L}\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)}}\cdot\\
&\Bigg(\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)q(s,a)p(\cdot|s,a)^\top\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2}+\\
&\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\left|\left(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)-q(s,a)p(\cdot|s,a)\right)^\top\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2\right|}\Bigg)\\
\leq&\sqrt{L}\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)}}\cdot\\
&\Bigg(\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)q(s,a)p(\cdot|s,a)^\top\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2}+
\end{aligned}$$

$$
\begin{aligned}
&\sqrt{H}\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\left|\left(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)-q(s,a)p(\cdot|s,a)\right)^\top\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)\right|}\Bigg)\\
&\overset{(a)}{\leq}\sqrt{L}\sqrt{SNL}\left(\sqrt{28900mH^4L^3S^2N}+\sqrt{H}\sqrt{170S^2NH^2L^2\sqrt{mL}}\right)\\
&\leq 184SNH^2L^2\sqrt{mSL} \qquad (17)
\end{aligned}
$$

where (a) comes from Lemmas 14,17 and Eq. (20) (the upper bound of `Term` 4).

Plugging Eq. (17) into Eq. (15), we obtain

$$
\begin{aligned}
\texttt{Term 2} =&4L\sqrt{3SNKH^2}+68SNHL^2+8\underbrace{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\sqrt{\frac{\mathbb{E}_{s'\sim\hat{q},\hat{p}}\left[\left(\bar{V}^k_{h+1}(s')-\underline{V}^k_{h+1}(s')\right)^2\right]L}{n^k(s,a)}}}_{\texttt{Term 2.1}}\\
\leq&8HL\sqrt{SNK}+68SNHL^2+1472SNH^2L^2\sqrt{mSL}\\
\leq&8HL\sqrt{SNK}+1540SNH^2L^2\sqrt{mSL} \qquad (18)
\end{aligned}
$$

Step 4: Bound the estimate deviation term for triggered future values – `Term` **3.**

$$
\begin{aligned}
\texttt{Term 3} =&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\left(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)-q(s,a)p(\cdot|s,a)\right)^\top V^*_{h+1}\\
\leq&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\left(4\sqrt{\frac{\text{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)L}{n^k(s,a)}}+\frac{4HL}{n^k(s,a)}\right)\\
\leq&4\sqrt{L}\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)}}\sqrt{\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\text{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)}\\
&+4HL\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)}\\
\leq&4L\sqrt{SN\sum_{k=1}^{K}\mathbb{E}_{\pi^k}\left[\sum_{h=1}^{H}\sum_{\ell=1}^{m^h}\text{Var}_{s'\sim q,p}\left(V^*_{h+1}(s')\right)\right]}+4SNHL^2\\
\overset{(a)}{\leq}&4L\sqrt{3SNKH^2}+4SNHL^2\\
\leq&8HL\sqrt{SNK}+4SNHL^2, \qquad (19)
\end{aligned}
$$

where (a) comes from Lemma 4.

Step 5: Bound the second order term. – `Term` **4.**

$$
\begin{aligned}
\texttt{Term 4} =&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\left(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)-q(s,a)p(\cdot|s,a)\right)^\top\left(\bar{V}^k_{h+1}-V^*_{h+1}\right)\\
=&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\sum_{s'}\left(\hat{q}^k(s,a)\hat{p}^k(s'|s,a)-q(s,a)p(s'|s,a)\right)\left(\bar{V}^k_{h+1}(s')-V^*_{h+1}(s')\right)\\
\leq&\sum_{k=1}^{K}\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\sum_{s'}\left|\hat{q}^k(s,a)\hat{p}^k(s'|s,a)-q(s,a)p(s'|s,a)\right|\cdot\left(\bar{V}^k_{h+1}(s')-V^*_{h+1}(s')\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a)\cdot \\
&\quad \sum_{s'} \left(\sqrt{\frac{q(s,a)p(s'|s,a)\left(1-q(s,a)p(s'|s,a)\right)L}{n^k(s,a)}} + \frac{L}{n^k(s,a)} \right) \left(\bar{V}^k_{h+1}(s') - \underline{V}^k_{h+1}(s') \right) \\
&\leq \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a) \sum_{s'} \sqrt{\frac{q(s,a)p(s'|s,a)L}{n^k(s,a)}} \left(\bar{V}^k_{h+1}(s') - \underline{V}^k_{h+1}(s') \right) \\
&\quad + \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a) \sum_{s'} \frac{HL}{n^k(s,a)} \\
&\leq \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a) \sqrt{SL \cdot \sum_{s'} \frac{q(s,a)p(s'|s,a)}{n^k(s,a)} \left(\bar{V}^k_{h+1}(s') - \underline{V}^k_{h+1}(s') \right)^2} \\
&\quad + \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a) \frac{SHL}{n^k(s,a)} \\
&\leq \sqrt{SL} \sqrt{\sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a),s\neq s_\perp} \frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)}} \cdot \\
&\quad \sqrt{\underbrace{\sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a) \sum_{s'} q(s,a)p(s'|s,a) \left(\bar{V}^k_{h+1}(s') - \underline{V}^k_{h+1}(s') \right)^2}_{\texttt{Term 4.1}}} \\
&\quad + SHL \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a),s\neq s_\perp} \frac{w_{k\sigma\ell}(s,a)}{n^k(s,a)} \\
&\overset{(a)}{\leq} \sqrt{SL}\sqrt{SNL}\sqrt{28900 m H^4 L^3 S^2 N} + S^2 NHL^2 \\
&\leq 170 S^2 N H^2 L^2 \sqrt{mL}
\end{aligned} \tag{20}
$$

where (a) is due to Lemmas 14,17.

Finally, we combine the upper bounds of Term 1, Term 2, Term 3, Term 4 and the minimal regret contribution to bound the total regret.

Plugging Eqs. (14),(18),(19) and (20) into Eq. (13), we have

$$
\begin{aligned}
&\texttt{Regret}(K) \\
\leq& \sum_{k=1}^{K} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \sum_{(s,a),s\neq s_\perp} w_{k\sigma\ell}(s,a) \Big[\underbrace{\left(\tilde{q}^k(s,a) - q(s,a) \right) r(s,a)}_{\texttt{Term 1}} + \underbrace{\left(\tilde{q}^k(s,a)\tilde{p}^k(\cdot|s,a) - \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a) \right)^\top \bar{V}^k_{h+1}}_{\texttt{Term 2}} \\
&+ \underbrace{\left(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a) - q(s,a)p(\cdot|s,a) \right)^\top V^*_{h+1}}_{\texttt{Term 3}} + \underbrace{\left(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a) - q(s,a)p(\cdot|s,a) \right)^\top \left(\bar{V}^k_{h+1} - V^*_{h+1} \right)}_{\texttt{Term 4}} \Big] \\
\leq& 8L\sqrt{mSNHK} + 8HL\sqrt{SNK} + 1540 SNH^2L^2\sqrt{mSL} + 8HL\sqrt{SNK} + 4SNHL^2 \\
&+ 170 S^2NH^2L^2\sqrt{mL} \\
=& O\left(HL\sqrt{SNK} \right)
\end{aligned}
$$

□

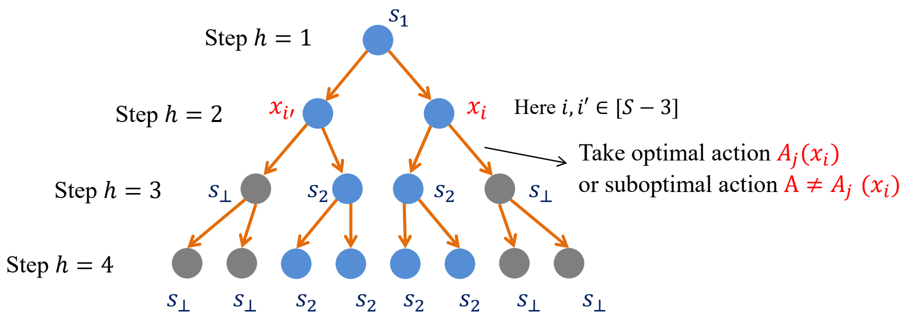

Figure 3. The constructed instance with $m = 2$ in regret lower bound analysis.

C.2. Regret Lower Bound

In this subsection, we prove the regret lower bounds for branching RL-RM in cases with Assumption 1 (Theorem 7) and without Assumption 1 (Theorem 2).

C.2.1. Proof of Theorem 7

Proof of Theorem 7. As shown in Figure 3, consider a random instance as follows: There are N base actions, i.e., $A^{\texttt{univ}} = \{a_1, \dots, a_N\}$, and $d := \frac{N}{m}$ super actions, i.e., $A_1 = \{a_1, \dots, a_m\}, A_2 = \{a_{m+1}, \dots, a_{2m}\}, \dots, A_d = \{a_{m(d-1)+1}, \dots, a_{md}\}$. The action set is $\mathcal{A} = \{A_1, \dots, A_d\}$. The state set is $\mathcal{S} = \{s_\perp, s_1, s_2, x_1, \dots, x_{S-3}\}$.

The trigger probabilities are as follows: $q(s_1, a) = q(s_2, a) = \alpha + \eta$ and $q(s_\perp, a) = 0$ for any $a \in A^{\texttt{univ}}$.

The transition distributions are as follows: s_1 is the initial state. $p(x_i|s_1, a) = \frac{1}{S-3}$ for any $a \in A^{\texttt{univ}}$. $p(s_2|x_i, a) = 1$ for any $a \in A^{\texttt{univ}}, i \in [S-3]$. $p(s_2|s_2, a) = 1$ and $p(s_\perp|s_\perp, a) = 1$ for any $a \in A^{\texttt{univ}}$.

The reward function is only dependent on the current state. $r(s, a) = 1$ for any $s \in \mathcal{S} \setminus \{s_\perp\}, a \in A^{\texttt{univ}}$, and $r(s_\perp, a) = 0$ for any $a \in A^{\texttt{univ}}$.

The randomness of this instance is as follows: for each $x_i \in \{x_1, \dots, x_{S-3}\}$, we uniformly choose an action $A_j(x_i)$ from $A_1, \dots, A_d$ as the optimal action. Let $q(x_i, a) = \alpha + \eta$ for all $a \in A_j(x_i)$, and $q(x_i, a) = \alpha$ for all $a \notin A_j(x_i)$. Let $\alpha + \eta = \frac{1}{m}$.

In words, in each episode, at step 1, an agent starts from state s_1 and takes an action that contains m base actions, where each base action has trigger probability $\alpha + \eta$. For each state-base action pair at step 1, if triggered successfully, it transitions to $x_i \in \{x_1, \dots, x_{S-3}\}$ with probability $\frac{1}{S-3}$; Otherwise, if triggered successfully, it transitions to ending state $s_\perp$. At step 1, in a bandit state x_i, the agent takes an action $A \in \mathcal{A}$ that contains m base actions, where each base action has trigger probability $\alpha + \eta$ if $A = A_j(x_i)$, and has trigger probability only α if $A \neq A_j(x_i)$. At step 2, for each state-base action pair, if triggered successfully, it transitions to s_2; Otherwise, it transitions to $s_\perp$. At step 3, starting from state s_2, the agent takes an action where each contained base action has trigger probability $\alpha + \eta$. For each state-base action pair at step 3, if triggered successfully, it still transitions back to s_2; Otherwise, it transitions to $s_\perp$. The following steps $4, \dots, H$ are similar to step 3, where the agent starts from s_2 and transitions back to s_2 or transitions to $s_\perp$.

The optimal policy π_* is to take action $A_j(x_i)$ at state x_i, and we have

$$\mathbb{E}[\texttt{Reward}^{\pi_*}] = \mathbb{E}\left[\sum_{k=1}^{K} V_1^*(s_1)\right] = K\left(m(\alpha+\eta) + m^2(\alpha+\eta)^2 + \cdots + m^H(\alpha+\eta)^H\right) = HK \tag{21}$$

Fix an algorithm $\mathbb{A}$. Let π^k denote the policy taken by $\mathbb{A}$ in episode k. For each $x_i \in \{x_1, \dots, x_{S-3}\}$, let $T_{x_i, A_j(x_i)} := \sum_{k=1}^{K} \mathbb{1}\left\{\pi^k(x_i) = A_j(x_i)\right\}$ denote the number of episodes where $\mathbb{A}$ chooses $A_j(x_i)$ in state x_i. Then, the number of episodes where $\mathbb{A}$ chooses suboptimal actions in state x_i is $K - T_{x_i, A_j(x_i)}$.

$$\mathbb{E}[\texttt{Reward}^{\mathbb{A}}] = \mathbb{E}\left[\sum_{k=1}^{K} V_1^{\pi^k}(s_1)\right]$$

$$
\begin{aligned}
=&Km(\alpha+\eta)+\frac{1}{(S-3)d}\sum_{i=1}^{S-3}\sum_{j=1}^{d}\mathbb{E}\Big[T_{x_i,A_j(x_i)}\left(m^2(\alpha+\eta)^2+\cdots+m^H(\alpha+\eta)^H\right)\\
&+\left(K-T_{x_i,A_j(x_i)}\right)\left(m^2(\alpha+\eta)\alpha+m^3(\alpha+\eta)\alpha(\alpha+\eta)+\cdots+m^H(\alpha+\eta)\alpha(\alpha+\eta)^{H-2}\right)\Big]\\
=&Km(\alpha+\eta)+\frac{1}{(S-3)d}\sum_{i=1}^{S-3}\sum_{j=1}^{d}\mathbb{E}\Big[T_{x_i,A_j(x_i)}\left(m^2(\alpha+\eta)^2+\cdots+m^H(\alpha+\eta)^H\right)\\
&+\left(K-T_{x_i,A_j(x_i)}\right)\left(m^2(\alpha+\eta)\alpha+m^3(\alpha+\eta)^2\alpha+\cdots+m^H(\alpha+\eta)^{H-1}\alpha\right)\Big]
\end{aligned} \tag{22}
$$

Subtracting Eq. (22) by Eq. (21), we have

$$
\begin{aligned}
\mathbb{E}[\texttt{Regret}^{\mathbb{A}}]=&\mathbb{E}[\texttt{Reward}^{\pi_*}]-\mathbb{E}[\texttt{Reward}^{\mathbb{A}}]\\
=&\mathbb{E}\left[\sum_{k=1}^{K}\left(V_1^*(s_1)-V_1^{\pi^k}(s_1)\right)\right]\\
=&\frac{1}{(S-3)d}\sum_{i=1}^{S-3}\sum_{j=1}^{d}\mathbb{E}\Big[\left(K-T_{x_i,A_j(x_i)}\right)\left(m(\alpha+\eta)m\eta+m^2(\alpha+\eta)^2m\eta+\cdots+m^{H-1}(\alpha+\eta)^{H-1}m\eta\right)\Big]\\
=&\frac{1}{(S-3)d}\sum_{i=1}^{S-3}\sum_{j=1}^{d}\mathbb{E}\Big[\left(K-T_{x_i,A_j(x_i)}\right)(H-1)m\eta\Big]\\
=&m\eta(H-1)\left(K-\frac{1}{(S-3)d}\sum_{i=1}^{S-3}\sum_{j=1}^{d}\mathbb{E}\left[T_{x_i,A_j(x_i)}\right]\right)
\end{aligned} \tag{23}
$$

Let $\mathbb{E}_{A_j(x_i)}[\cdot]$ denote the expectation operator under the instance $\mathcal{I}_{A_j(x_i)}$ where the optimal action of state x_i is $A_j(x_i)$.

Let $\mathbb{E}_{x_i,\texttt{unif}}[\cdot]$ denote the expectation operator under the instance where all actions $A\in\mathcal{A}$ at state x_i have the same trigger probability, i.e., $q(x_i,a)=\alpha$ for any $a\in A^{\mathrm{univ}}$, and other distribution settings are the same as $\mathcal{I}_{A_j(x_i)}$.

Note that the KL-divergence between the above two instances is $m\cdot\mathrm{KL}\left(\texttt{Bernoulli}(\alpha)||\texttt{Bernoulli}(\alpha+\eta)\right)\le m\cdot\frac{\eta^2}{(\alpha+\eta)(1-(\alpha+\eta))}\le m\cdot\frac{\eta^2}{c_1}$ with $(\alpha+\eta)(1-(\alpha+\eta))\ge c_1$ for some absolute positive constant c_1. After a pull of $(x_i,A_j(x_i))$, we receive an observation of such difference between the two instances.

Using Lemma A.1 in (Auer et al., 2002), we have

$$
\begin{aligned}
\mathbb{E}_{A_j(x_i)}\left[T_{x_i,A_j(x_i)}\right]\le&\mathbb{E}_{x_i,\texttt{unif}}\left[T_{x_i,A_j(x_i)}\right]\\
&+\frac{K}{2}\sqrt{\frac{1}{2}\cdot\frac{1}{S-3}\mathbb{E}_{x_i,\texttt{unif}}\left[T_{x_i,A_j(x_i)}\right]\cdot m\mathrm{KL}\left(\texttt{Bernoulli}(\alpha)||\texttt{Bernoulli}(\alpha+\eta)\right)}\\
\overset{\text{(a)}}{\le}&\mathbb{E}_{x_i,\texttt{unif}}\left[T_{x_i,A_j(x_i)}\right]+\frac{K}{2}\sqrt{\frac{m}{2c_1(S-3)}\mathbb{E}_{x_i,\texttt{unif}}\left[T_{x_i,A_j(x_i)}\right]\eta^2}\\
=&\mathbb{E}_{x_i,\texttt{unif}}\left[T_{x_i,A_j(x_i)}\right]+\frac{K\eta}{2}\sqrt{\frac{m}{2c_1(S-3)}\mathbb{E}_{x_i,\texttt{unif}}\left[T_{x_i,A_j(x_i)}\right]}
\end{aligned}
$$

Since $\sum_{j=1}^{d}\mathbb{E}_{x_i,\texttt{unif}}\left[T_{x_i,A_j(x_i)}\right]=\sum_{j=1}^{d}\sum_{k=1}^{K}\mathbb{E}_{x_i,\texttt{unif}}\left[\pi^k(A_j(x_i)|x_i)\right]=K$, we have

$$
\begin{aligned}
\sum_{j=1}^{d}\mathbb{E}_{A_j(x_i)}\left[T_{x_i,A_j(x_i)}\right]\le&\sum_{j=1}^{d}\mathbb{E}_{x_i,\texttt{unif}}\left[T_{x_i,A_j(x_i)}\right]+\frac{K\eta}{2}\sum_{j=1}^{d}\sqrt{\frac{m}{2c_1(S-3)}\mathbb{E}_{x_i,\texttt{unif}}\left[T_{x_i,A_j(x_i)}\right]}\\
\le&\sum_{j=1}^{d}\mathbb{E}_{x_i,\texttt{unif}}\left[T_{x_i,A_j(x_i)}\right]+\frac{K\eta}{2}\sqrt{\frac{dm}{2c_1(S-3)}\sum_{j=1}^{d}\mathbb{E}_{x_i,\texttt{unif}}\left[T_{x_i,A_j(x_i)}\right]}
\end{aligned}
$$

$$=K + \frac{K\eta}{2}\sqrt{\frac{dmK}{2c_1(S-3)}}$$

and thus

$$\begin{aligned}\frac{1}{(S-3)d}\sum_{i=1}^{S-1}\sum_{j=1}^{d}\mathbb{E}_{A_j(x_i)}\left[T_{x_i,A_j(x_i)}\right] \leq& \frac{1}{d}\left(K + \frac{K\eta}{2}\sqrt{\frac{dmK}{2c_1(S-3)}}\right)\\ =& K\left(\frac{1}{d} + \frac{\eta}{2}\sqrt{\frac{mK}{2c_1 d(S-3)}}\right) \qquad (24)\end{aligned}$$

Plugging Eq. (24) into Eq. (23), we obtain

$$\begin{aligned}\mathbb{E}[\texttt{Regret}^{\mathbb{A}}] \geq& m\eta(H-1)\left(K - \frac{1}{(S-3)d}\sum_{i=1}^{S-3}\sum_{j=1}^{d}\mathbb{E}\left[T_{x_i,A_j(x_i)}\right]\right)\\ \geq& m\eta(H-1)K\left(1 - \frac{1}{d} - \frac{\eta}{2}\sqrt{\frac{mK}{2c_1 d(S-3)}}\right) \qquad (25)\end{aligned}$$

Let $\eta = c_2\sqrt{\frac{d(S-3)}{mK}}$ for some small enough constant c_2, we have

$$\begin{aligned}\mathbb{E}[\texttt{Regret}^{\mathbb{A}}] =& \Omega\left((H-1)\sqrt{(S-3)dmK}\right)\\ =& \Omega\left(H\sqrt{SNK}\right)\end{aligned}$$

□

C.2.2. Proof of Theorem 2

Proof of Theorem 2. This proof uses the same instance and analytical procedure as the proof of Theorem 7, except that we set $\alpha + \eta = \bar{q}$ for some trigger probability threshold $\bar{q} > \frac{1}{m}$.

Then, Eq. (23) becomes

$$\begin{aligned}\mathbb{E}[\texttt{Regret}^{\mathbb{A}}] =& \mathbb{E}[\texttt{Reward}^{\pi_*}] - \mathbb{E}[\texttt{Reward}^{\mathbb{A}}]\\ =& \mathbb{E}\left[\sum_{k=1}^{K}\left(V_1^*(s_1) - V_1^{\pi^k}(s_1)\right)\right]\\ =& \frac{1}{(S-3)d}\sum_{i=1}^{S-3}\sum_{j=1}^{d}\mathbb{E}\Big[\left(K - T_{x_i,A_j(x_i)}\right)\left(m\bar{q}\cdot m\eta + m^2\bar{q}^2\cdot m\eta + \cdots + m^{H-1}\bar{q}^{H-1}\cdot m\eta\right)\Big]\\ =& \frac{1}{(S-3)d}\sum_{i=1}^{S-3}\sum_{j=1}^{d}\mathbb{E}\Big[\left(K - T_{x_i,A_j(x_i)}\right)\frac{m\bar{q}\left((m\bar{q})^{H-1}-1\right)}{m\bar{q}-1}\cdot m\eta\Big]\\ =& \frac{m\bar{q}\left((m\bar{q})^{H-1}-1\right)}{m\bar{q}-1}\cdot m\eta\left(K - \frac{1}{(S-3)d}\sum_{i=1}^{S-3}\sum_{j=1}^{d}\mathbb{E}\left[T_{x_i,A_j(x_i)}\right]\right) \qquad (26)\end{aligned}$$

and Eq (25) becomes

$$\mathbb{E}[\texttt{Regret}^{\mathbb{A}}] \geq \frac{m\bar{q}\left((m\bar{q})^{H-1}-1\right)}{m\bar{q}-1}\cdot m\eta\left(K - \frac{1}{(S-3)d}\sum_{i=1}^{S-3}\sum_{j=1}^{d}\mathbb{E}\left[T_{x_i,A_j(x_i)}\right]\right) \qquad (27)$$

$$\geq \frac{m\bar{q}\left((m\bar{q})^{H-1}-1\right)}{m\bar{q}-1} \cdot m\eta K \left(1-\frac{1}{d}-\frac{\eta}{2}\sqrt{\frac{mK}{2c_1 d(S-3)}}\right) \tag{28}$$

Let $\eta = c_2\sqrt{\frac{d(S-3)}{mK}}$ for some small enough constant c_2, we have

$$\begin{aligned}
\mathbb{E}[\texttt{Regret}^{\mathbb{A}}] =& \Omega\left(\frac{m\bar{q}\left((m\bar{q})^{H-1}-1\right)}{m\bar{q}-1}\sqrt{(S-3)dmK}\right) \\
=& \Omega\left(\frac{m\bar{q}\left((m\bar{q})^{H-1}-1\right)}{m\bar{q}-1}\sqrt{SNK}\right)
\end{aligned}$$

Therefore, when relaxing the trigger probability threshold in Assumption 1 to some $\bar{q} > \frac{1}{m}$, any algorithm for branching RL-RM must suffer an exponential regret.

□

D. Proofs for Branching RL with Reward-Free Exploration

In this section, we prove sample complexity upper and lower bounds (Theorems 8,9) for branching RL-RFE.

D.1. Proof for Sample Complexity Upper Bound

D.1.1. Augmented Transition Distribution

First, we introduce an augmented transition distribution $p^{\texttt{aug}}(\cdot|s,a)$ and connect it with trigger distribution $q(s,a)$ and transition distribution $p(\cdot|s,a)$.

For any $(s,a) \in \mathcal{S} \times A^{\texttt{univ}}$, let $p^{\texttt{aug}}(\cdot|s,a)$ denote the augmented transition distribution on $\mathcal{S}$, which satisfies that

$$\begin{aligned}
p^{\texttt{aug}}(s_\perp|s,a) &= 1-q(s,a), \\
p^{\texttt{aug}}(s'|s,a) &= q(s,a)p(s'|s,a), \ \forall s' \in \mathcal{S}\setminus\{s_\perp\}.
\end{aligned}$$

For any episode k, we can also define the empirical augmented transition distribution as

$$\begin{aligned}
\hat{p}^{\texttt{aug},k}(s_\perp|s,a) &= 1-\hat{q}^k(s,a), \\
\hat{p}^{\texttt{aug},k}(s'|s,a) &= \hat{q}^k(s,a)\hat{p}^k(s'|s,a), \ \forall s' \in \mathcal{S}\setminus\{s_\perp\}.
\end{aligned}$$

Lemma 18. *For any function $f(\cdot)$ defined on $\mathcal{S}$ such that $f(s_\perp)=0$ (e.g., f can be the value function $V_h^\pi(\cdot)$, $\forall h \in [H], \pi$), it holds that*

$$p^{\texttt{aug}}(\cdot|s,a)^\top f = q(s,a)p(\cdot|s,a)^\top f, \tag{29}$$

$$\hat{p}^{\texttt{aug},k}(\cdot|s,a)^\top V_{h+1} = \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top f. \tag{30}$$

Proof of Lemma 18. We prove Eq. (29) as follows:

$$\begin{aligned}
p^{\texttt{aug}}(\cdot|s,a)^\top f &= \sum_{s'\in\mathcal{S}} p^{\texttt{aug}}(s'|s,a)f(s') \\
&= \sum_{s'\in\mathcal{S}\setminus\{s_\perp\}} q(s,a)p(s'|s,a)f(s') \\
&= q(s,a)p(\cdot|s,a)^\top f
\end{aligned}$$

Eq. (30) can be proved in a similar manner. □

D.1.2. CONCENTRATION

In the following, we introduce a concentration lemma.

Lemma 19 (KL-divergence Based Concentration of Triggered Transition)**.** *Defining event*

$$\mathcal{G} := \left\{ \mathrm{KL}\left(\hat{p}^{\texttt{aug},k}(\cdot|s,a) \| p^{\texttt{aug}}(\cdot|s,a)\right) \leq \frac{\log\left(\frac{SN}{\delta'}\right) + S\log\left(8e(n^k(s,a)+1)\right)}{n^k(s,a)},\ \forall (s,a) \in \mathcal{S} \times A^{\texttt{univ}}, \forall k \right\},$$

it holds that

$$\Pr\left[\mathcal{G}\right] \geq 1 - \delta.$$

Proof of Lemma 19. Using Theorem 3 and Lemma 3 in (Ménard et al., 2021), we can obtain this lemma. □

D.1.3. KL DIVERGENCE-BASED TECHNICAL TOOLS

Below, we present several useful KL divergence-based technical tools.

Lemma 20 (Lemma 10 in (Ménard et al., 2021))**.** *Let p_1 and p_2 be two distributions on $\mathcal{S}$ such that* $\mathrm{KL}(p_1, p_2) \leq \alpha$*. Let f be a function defined on $\mathcal{S}$ such that for any $s \in \mathcal{S}$, $0 \leq f(s) \leq b$. Then,*

$$|p_1 f - p_2 f| \leq \sqrt{2\mathrm{Var}_{p_2}(f)\alpha} + \frac{2}{3}b\alpha$$

Lemma 21 (Lemma 11 in (Ménard et al., 2021))**.** *Let p_1 and p_2 be two distributions on $\mathcal{S}$ such that* $\mathrm{KL}(p_1, p_2) \leq \alpha$*. Let f be a function defined on $\mathcal{S}$ such that for any $s \in \mathcal{S}$, $0 \leq f(s) \leq b$. Then,*

$$\mathrm{Var}_{p_2}(f) \leq 2\mathrm{Var}_{p_1}(f) + 4b^2\alpha$$
$$\mathrm{Var}_{p_1}(f) \leq 2\mathrm{Var}_{p_2}(f) + 4b^2\alpha$$

Lemma 22 (Lemma 12 in (Ménard et al., 2021))**.** *Let p_1 and p_2 be two distributions on $\mathcal{S}$ such that* $\mathrm{KL}(p_1, p_2) \leq \alpha$*. Let f, g be two functions defined on $\mathcal{S}$ such that for any $s \in \mathcal{S}$, $0 \leq f(s), g(s) \leq b$. Then,*

$$\begin{aligned}\mathrm{Var}_{p_1}(f) \leq& 2\mathrm{Var}_{p_1}(g) + 2bp_1|f - g| \\ \mathrm{Var}_{p_2}(f) \leq& \mathrm{Var}_{p_1}(f) + 3b^2\alpha\|p_1 - p_2\|_1\end{aligned}$$

D.1.4. ESTIMATION ERROR

Next, we state an important lemma on estimation error.

Lemma 23 (Estimation Error)**.** *Suppose that the concentration event $\mathcal{G}$ holds. Then, for any episode k, policy π and reward function r,*

$$\left|\hat{V}_1^{k,\pi}(s;r) - V_1^{\pi}(s;r)\right| \leq 4e\sqrt{B_1^k(s)} + B_1^k(s).$$

Proof of Lemma 23. For any $t \in \mathbb{N}, \kappa \in (0,1)$, let $\beta(t,\kappa) := \log(SN/\kappa) + S\log(8e(t+1))$. Then, for any π, r and $(s, A) \in \mathcal{S} \setminus \{s_\perp\} \times \mathcal{A}$,

$$\begin{aligned}&\left|\hat{Q}_h^{k,\pi}(s,A;r) - Q_h^{\pi}(s,A;r)\right| \\ =&\sum_{a \in A}\left|\left(\hat{q}^k(s,a) - q(s,a)\right)r(s,a) + \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \hat{V}_{h+1}^{k,\pi} - q(s,a)p(\cdot|s,a)^\top V_{h+1}^{\pi}\right| \\ \leq&\sum_{a \in A}\Bigg(\left|\hat{q}^k(s,a) - q(s,a)\right|r(s,a) + \left|\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a) - q(s,a)p(\cdot|s,a)\right|^\top V_{h+1}^{\pi} \\ &+ \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top\left|\hat{V}_{h+1}^{k,\pi} - V_{h+1}^{\pi}\right|\Bigg)\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(a)}}{\le} \sum_{a \in A} \Bigg(2\sqrt{\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}} r(s,a) + \sqrt{\frac{2\text{Var}_{s' \sim q,p}\left(V^*_{h+1}(s')\right)\beta(n^k(s,a),\delta')}{n^k(s,a)}} + \frac{2}{3}H\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} \\
&\quad + \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \left| \hat{V}^{k,\pi}_{h+1} - V^{\pi}_{h+1} \right| \Bigg) \\
&\overset{\text{(b)}}{\le} \sum_{a \in A} \Bigg(2\sqrt{\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}} r(s,a) + \sqrt{\frac{4\text{Var}_{s' \sim \hat{q},\hat{p}}\left(V^*_{h+1}(s')\right)\beta(n^k(s,a),\delta')}{n^k(s,a)} + 8H^2\left(\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}\right)^2} \\
&\quad + \frac{2}{3}H\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} + \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \left| \hat{V}^{k,\pi}_{h+1} - V^{\pi}_{h+1} \right| \Bigg) \\
&\overset{\text{(c)}}{\le} \sum_{a \in A} \Bigg(2\sqrt{\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}} r(s,a) \\
&\quad + \sqrt{\frac{8\text{Var}_{s'}(\hat{V}^{k,\pi}_{h+1}(s'))\beta(n^k(s,a),\delta')}{n^k(s,a)} + 8H\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \left| V^{\pi}_{h+1} - \hat{V}^{k,\pi}_{h+1} \right| \frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} + 8H^2\left(\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}\right)^2} \\
&\quad + \frac{2}{3}H\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} + \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \left| \hat{V}^{k,\pi}_{h+1} - V^{\pi}_{h+1} \right| \Bigg) \\
&\overset{\text{(d)}}{\le} \sum_{a \in A} \Bigg(2\sqrt{\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}} r(s,a) + \sqrt{\frac{8\text{Var}_{s' \sim \hat{q},\hat{p}}\left(\hat{V}^{k,\pi}_{h+1}(s')\right)\beta(n^k(s,a),\delta')}{n^k(s,a)}} + \sqrt{8H^2\left(\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}\right)^2} \\
&\quad + \sqrt{\frac{1}{H}\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \left| V^{\pi}_{h+1} - \hat{V}^{k,\pi}_{h+1} \right| 8H^2\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}} + \frac{2}{3}H\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} \\
&\quad + \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \left| \hat{V}^{k,\pi}_{h+1} - V^{\pi}_{h+1} \right| \Bigg) \\
&\overset{\text{(e)}}{\le} \sum_{a \in A} \Bigg(2\sqrt{\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}} r(s,a) + \sqrt{\frac{8\text{Var}_{s' \sim \hat{q},\hat{p}}\left(\hat{V}^{k,\pi}_{h+1}(s')\right)\beta(n^k(s,a),\delta')}{n^k(s,a)}} \\
&\quad + \frac{1}{H}\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \left| V^{\pi}_{h+1} - \hat{V}^{k,\pi}_{h+1} \right| + 12H^2\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} + \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \left| \hat{V}^{k,\pi}_{h+1} - V^{\pi}_{h+1} \right| \Bigg) \\
&= \sum_{a \in A} \Bigg(2\sqrt{\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}} r(s,a) + \sqrt{\frac{8\text{Var}_{s' \sim \hat{q},\hat{p}}\left(\hat{V}^{k,\pi}_{h+1}(s')\right)\beta(n^k(s,a),\delta')}{n^k(s,a)}} + 12H^2\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} \\
&\quad + \left(1 + \frac{1}{H}\right)\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top \left| \hat{V}^{k,\pi}_{h+1} - V^{\pi}_{h+1} \right| \Bigg),
\end{aligned}
$$

Here (a)(b)(c) use Lemmas 20,21,22, respectively. (d) is due to that $\sqrt{x+y+z} \le \sqrt{x} + \sqrt{y} + \sqrt{z}$ for $x,y,z \ge 0$. (e) comes from that $\sqrt{xy} \le x + y$ for $x,y \ge 0$.

Then, unfolding $\left| \hat{Q}^{k,\pi}_1(s,a;r) - Q^{\pi}_1(s,a;r) \right|$, we have

$$
\begin{aligned}
&\left| \hat{Q}^{k,\pi}_1(s,A;r) - Q^{\pi}_1(s,A;r) \right| \\
\le & \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \left(1 + \frac{1}{H}\right)^{h-1} \mathbb{E}_{\hat{q},\hat{p},\pi} \Bigg[\Bigg(2\sqrt{\frac{\beta(n^k(s_\sigma, a_{\sigma\oplus\ell}),\delta')}{n^k(s_\sigma, a_{\sigma\oplus\ell})}} r(s_\sigma, a_{\sigma\oplus\ell})
\end{aligned}
$$

$$\begin{aligned}
&+\sqrt{\frac{8\mathrm{Var}_{s'\sim\hat{q},\hat{p}}\left(\hat{V}^{k,\pi}_{h+1}(s')\right)\beta(n^k(s_\sigma,a_{\sigma\oplus\ell}),\delta')}{n^k(s_\sigma,a_{\sigma\oplus\ell})}}+12H^2\frac{\beta(n^k(s_\sigma,a_{\sigma\oplus\ell}),\delta')}{n^k(s_\sigma,a_{\sigma\oplus\ell})}\Bigg)\cdot\mathbb{1}\left\{s_\sigma\neq s_\perp\right\}\Bigg]\\
\leq& e\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\hat{w}^{k,\pi}_{\sigma\ell}(s,a)\left(5\sqrt{\frac{\mathrm{Var}_{s'\sim\hat{q},\hat{p}}\left(\hat{V}^{k,\pi}_{h+1}(s')\right)}{H^2}\frac{H^2\beta(n^k(s,a),\delta')}{n^k(s,a)}}+12H^2\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}\right)\\
\leq& 5e\sqrt{\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\hat{w}^{k,\pi}_{\sigma\ell}(s,a)\frac{\mathrm{Var}_{s'\sim\hat{q},\hat{p}}\left(\hat{V}^{k,\pi}_{h+1}(s')\right)}{H^2}}\sqrt{H^2\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\hat{w}^{k,\pi}_{\sigma\ell}(s,a)\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}}\\
&+12eH^2\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\hat{w}^{k,\pi}_{\sigma\ell}(s,a)\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}\\
=& 5e\sqrt{\frac{1}{H^2}\mathbb{E}_{\hat{q},\hat{p},\pi}\left[\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\mathrm{Var}_{s'\sim\hat{q},\hat{p}}\left(\hat{V}^{k,\pi}_{h+1}(s')\right)\right]}\sqrt{H^2\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\hat{w}^{k,\pi}_{\sigma\ell}(s,a)\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}}\\
&+12eH^2\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\hat{w}^{k,\pi}_{\sigma\ell}(s,a)\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}\\
\overset{\text{(a)}}{=}& 10e\sqrt{H^2\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\hat{w}^{k,\pi}_{\sigma\ell}(s,a)\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}}+12eH^2\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\hat{w}^{k,\pi}_{\sigma\ell}(s,a)\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)},
\end{aligned}\tag{31}$$

where (a) uses branching law of total variance (Lemma 4), which also holds for the estimated model $(\hat{q}^k,\hat{p}^k)$ if adding a clip operation $\hat{q}^k(s,a)\leftarrow\min\{\hat{q}^k(s,a),\frac{1}{m}\}$ in algorithm `BranchRFE` to guarantee $\hat{q}^k\leq\frac{1}{m}$.

Define

$$\begin{aligned}
B^{k,\pi}_h(s,A):=&\min\left\{\sum_{a\in A}\left(12H^2\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}+\left(1+\frac{1}{H}\right)\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top B^{k,\pi}_{h+1}\right),H\right\}\\
B^{k,\pi}_h(s):=&B^{k,\pi}_h(s,\pi(s)).\\
B^{k}_h(s,A):=&\min\left\{\sum_{a\in A}\left(12H^2\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}+\left(1+\frac{1}{H}\right)\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top B^{k}_{h}(s)\right),H\right\}\\
B^{k}_h(s):=&\max_{A\in\mathcal{A}}B^{k}_h(s,A)
\end{aligned}$$

In the following, we show

$$\left|\hat{Q}^{k,\pi}_1(s,A;r)-Q^{\pi}_1(s,A;r)\right|\leq 4e\sqrt{B^{k,\pi}_1(s,A)}+B^{k,\pi}_1(s,A)\tag{32}$$

If $B^{k,\pi}_1(s,A)=H$, Eq. (32) holds trivially. Otherwise, unfolding $B^{k,\pi}_1(s,A)$, we have

$$B^{k,\pi}_1(s,A)=12eH^2\sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}}\sum_{\ell=1}^{m}\sum_{(s,a),s\neq s_\perp}\hat{w}^{k,\pi}_{\sigma\ell}(s,a)\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)},$$

and using Eq. (31), we obtain

$$\left|\hat{Q}^{k,\pi}_1(s,A;r)-Q^{\pi}_1(s,A;r)\right|\leq 4e\sqrt{B^{k,\pi}_1(s,A)}+B^{k,\pi}_1(s,A).$$

Thus, by the definitions of $B_h^k(s, A)$ and $B_h^k(s)$, we have

$$\left|\hat{V}_1^{k,\pi}(s;r) - V_1^{\pi}(s;r)\right| \leq 4e\sqrt{B_1^{k,\pi}(s)} + B_1^{k,\pi}(s) \leq 4e\sqrt{B_1^k(s)} + B_1^k(s).$$

□

D.1.5. Proof of Theorem 8

Now, we prove the sample complexity upper bound for algorithm `BranchRFE` (Theorem 8).

Proof of Theorem 8. First, we prove the correctness.

Let K denote the number of episodes that algorithm `BranchRFE` costs. According to the stopping rule (Line 2 in Algorithm `BranchRFE`) and Lemma 23, when algorithm `BranchRFE` stops in episode K, we have that for any π, r,

$$\left|\hat{V}_1^{K,\pi}(s_1;r) - V_1^{\pi}(s_1;r)\right| \leq 4e\sqrt{B_1^K(s)} + B_1^K(s) \leq \frac{\varepsilon}{2}$$

Then, we have that for any r,

$$\begin{aligned}
V_1^*(s_1;r) - V_1^{\hat{\pi}^*}(s_1;r) =& V_1^*(s_1;r) - \hat{V}_1^{K,\pi^*}(s_1;r) + \hat{V}_1^{K,\pi^*}(s_1;r) - \hat{V}_1^{K,\hat{\pi}^*}(s_1;r) + \hat{V}_1^{K,\hat{\pi}^*}(s_1;r) - V_1^{\hat{\pi}^*}(s_1;r) \\
\leq& \left|V_1^*(s_1;r) - \hat{V}_1^{K,\pi^*}(s_1;r)\right| + \left|\hat{V}_1^{K,\hat{\pi}^*}(s_1;r) - V_1^{\hat{\pi}^*}(s_1;r)\right| \\
\leq& \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\
=& \varepsilon
\end{aligned}$$

Now, we prove the sample complexity.

$$\begin{aligned}
B_h^k(s, A) \leq& \sum_{a \in A} \left(12H^2 \frac{\beta(n^k(s,a), \delta')}{n^k(s,a)} + \left(1 + \frac{1}{H}\right) \hat{q}^k(s,a)\hat{p}^k(\cdot|s,a)^\top B_{h+1}^k \right) \\
=& \sum_{a \in A} \left(12H^2 \frac{\beta(n^k(s,a), \delta')}{n^k(s,a)} + \left(1 + \frac{1}{H}\right) q(s,a)p(\cdot|s,a)^\top B_{h+1}^k \right. \\
& \left. + \left(1 + \frac{1}{H}\right) \left(\hat{q}^k(s,a)\hat{p}^k(\cdot|s,a) - q(s,a)p(\cdot|s,a)\right)^\top B_{h+1}^k \right) \\
\overset{\text{(a)}}{\leq}& \sum_{a \in A} \left(12H^2 \frac{\beta(n^k(s,a), \delta')}{n^k(s,a)} + \left(1 + \frac{1}{H}\right) q(s,a)p(\cdot|s,a)^\top B_{h+1}^k \right. \\
& \left. + \left(1 + \frac{1}{H}\right) \left(\sqrt{2\text{Var}_{s' \sim q,p}\left(B_{h+1}^k(s')\right) \frac{\beta(n^k(s,a), \delta')}{n^k(s,a)}} + \frac{2}{3}H \frac{\beta(n^k(s,a), \delta')}{n^k(s,a)} \right) \right) \\
\leq& \sum_{a \in A} \left(12H^2 \frac{\beta(n^k(s,a), \delta')}{n^k(s,a)} + \left(1 + \frac{1}{H}\right) q(s,a)p(\cdot|s,a)^\top B_{h+1}^k \right. \\
& \left. + \left(1 + \frac{1}{H}\right) \left(\sqrt{2Hq(s,a)p(\cdot|s,a)^\top B_{h+1}^k \frac{\beta(n^k(s,a), \delta')}{n^k(s,a)}} + \frac{2}{3}H \frac{\beta(n^k(s,a), \delta')}{n^k(s,a)} \right) \right) \\
=& \sum_{a \in A} \left(12H^2 \frac{\beta(n^k(s,a), \delta')}{n^k(s,a)} + \left(1 + \frac{1}{H}\right) q(s,a)p(\cdot|s,a)^\top B_{h+1}^k \right. \\
& \left. + \left(1 + \frac{1}{H}\right) \left(\sqrt{\frac{1}{H}q(s,a)p(\cdot|s,a)^\top B_{h+1}^k 2H^2 \frac{\beta(n^k(s,a), \delta')}{n^k(s,a)}} + \frac{2}{3}H \frac{\beta(n^k(s,a), \delta')}{n^k(s,a)} \right) \right)
\end{aligned}$$

$$
\begin{aligned}
&\le \sum_{a \in A} \Bigg(12H^2 \frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} + \left(1+\frac{1}{H}\right) q(s,a)p(\cdot|s,a)^\top B^k_{h+1} \\
&\quad + \left(1+\frac{1}{H}\right)\left(\frac{1}{H} q(s,a)p(\cdot|s,a)^\top B^k_{h+1} + 2H^2 \frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} + \frac{2}{3}H\frac{\beta(n^k(s,a),\delta')}{n^k(s,a)}\right)\Bigg) \\
&\le \sum_{a \in A} \Bigg(12H^2 \frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} + \left(1+\frac{1}{H}\right) q(s,a)p(\cdot|s,a)^\top B^k_{h+1} + \frac{2}{H} q(s,a)p(\cdot|s,a)^\top B^k_{h+1} \\
&\quad + 6H^2 \frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} \Bigg) \\
&= \sum_{a \in A} \left(18H^2 \frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} + \left(1+\frac{3}{H}\right) q(s,a)p(\cdot|s,a)^\top B^k_{h+1} \right),
\end{aligned}
$$

where (a) uses Lemma 20.

Then, unfolding $B^k_1(s) = B^k_1(s, \pi^k(s))$ and summing over $k = 1, \dots, K-1$, we have

$$
\begin{aligned}
\sum_{k=1}^{K-1} B^k_1(s) \le & \sum_{k=1}^{K-1} \sum_{\sigma=\emptyset}^{m^{\oplus(H-1)}} \sum_{\ell=1}^{m} \mathbb{E}_{q,p,\pi^k}\left[18e^3H^2 \frac{\beta(n^k(s_\sigma, a_{\sigma\oplus\ell}),\delta')}{n^k(s_\sigma, a_{\sigma\oplus\ell})} \cdot \mathbb{1}\left\{ s_\sigma \neq s_\perp \right\} \right] \\
= & 18e^3H^2 \mathbb{E}_{q,p,\pi^k}\left[\sum_{k=1}^{K-1} \sum_{(s,a), s\neq s_\perp} X_k(s,a) \frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} \right] \\
\le & 18e^3H^2 \mathbb{E}_{q,p,\pi^k}\left[\sum_{(s,a), s\neq s_\perp} \sum_{k=1}^{K-1} X_k(s,a) \frac{\beta(n^k(s,a),\delta')}{n^k(s,a)} \right] \\
\le & 18e^3H^2 \cdot \beta\left((K-1)\frac{m^{H+1}-m}{m-1}, \delta' \right) \mathbb{E}_{q,p,\pi^k}\left[\sum_{(s,a), s\neq s_\perp} \log\left(n^{K-1}(s,a) \right) \right] \\
\le & 18e^3H^2 \cdot \beta\left((K-1)\frac{m^{H+1}-m}{m-1}, \delta' \right) \sum_{(s,a), s\neq s_\perp} \log\left(\mathbb{E}_{q,p,\pi^k}\left[n^{K-1}(s,a) \right] \right) \\
\le & 18e^3H^2SN \cdot \beta\left((K-1)\frac{m^{H+1}-m}{m-1}, \delta' \right) \log\left(H(K-1) \right) \\
\overset{\text{(a)}}{\le} & 18e^3H^2SN \cdot \left(\log\left(\frac{SN}{\delta'} \right) + S\log\left(8eHm^{H+1}(K-1) \right) \right) \log\left(H(K-1) \right),
\end{aligned} \tag{33}
$$

where (a) comes from $\beta(t,\kappa) := \log(SN/\kappa) + S\log(8e(t+1))$ and $\frac{m^{H+1}-m}{m-1} \le Hm^{H+1}$.

According to the stopping rule, we have $\varepsilon \le 4e\sqrt{B^k_1(s)} + B^k_1(s)$ for $k = 1, \dots, K-1$. Then, summing over $k = 1, \dots, K-1$ for both sides, we obtain

$$
\begin{aligned}
(K-1)\varepsilon \le & 4e \sum_{k=1}^{K-1} \sqrt{B^k_1(s)} + \sum_{k=1}^{K-1} B^k_1(s) \\
\le & 4e \sqrt{ (K-1) \sum_{k=1}^{K-1} B^k_1(s) } + \sum_{k=1}^{K-1} B^k_1(s)
\end{aligned}
$$

and thus

$$
(K-1) \le \frac{4e}{\varepsilon} \sqrt{ (K-1) \sum_{k=1}^{K-1} B^k_1(s) } + \frac{1}{\varepsilon} \sum_{k=1}^{K-1} B^k_1(s)
$$

$$
\begin{aligned}
\leq & \frac{4e}{\varepsilon}\sqrt{(K-1)\cdot 18e^3H^2SN\cdot\left(\log\left(\frac{SN}{\delta'}\right)+S\log\left(8eHm^{H+1}(K-1)\right)\right)\log\left(H(K-1)\right)} \\
& +\frac{18e^3H^2SN}{\varepsilon}\cdot\left(\log\left(\frac{SN}{\delta'}\right)+S\log\left(8eHm^{H+1}(K-1)\right)\right)\log\left(H(K-1)\right) \\
\leq & \frac{4e\sqrt{18e^3H^2SN}}{\varepsilon}\sqrt{(K-1)\cdot\log\left(\frac{SN}{\delta'}\right)\log\left(H(K-1)\right)+S\log\left(8em^{H+1}\right)\log^2\left(H(K-1)\right)} \\
& +\frac{18e^3H^2SN}{\varepsilon}\cdot\left(\log\left(\frac{SN}{\delta'}\right)\log\left(H(K-1)\right)+S\log\left(8em^{H+1}\right)\log^2\left(H(K-1)\right)\right)
\end{aligned}
$$

Using Lemma 13 in (Ménard et al., 2021) with $\tau = K-1$, $C = \frac{4e\sqrt{18e^3H^2SN}}{\varepsilon}$, $A = \log\left(\frac{SN}{\delta'}\right)$, $\alpha = H$, $B = E = S\log\left(8em^{H+1}\right)$ and $D = \frac{18e^3H^2SN}{\varepsilon}$, we obtain

$$
\begin{aligned}
K-1 =& \tilde{O}\left(C^2(A+B)C_1^2\right) \\
=& O\left(\frac{H^2SN}{\varepsilon^2}\left(\log\left(\frac{SN}{\delta}\right)+S\log\left(e\cdot m^H\right)\right)C_1^2\right),
\end{aligned}
$$

where $C_1 = \log(\alpha(A+E)(C+D))$.

Thus, we have

$$
K = O\left(\frac{H^2SN}{\varepsilon^2}\left(\log\left(\frac{SN}{\delta}\right)+S\log\left(e\cdot m^H\right)\right)C_1^2\right),
$$

where

$$
C_1 = \log\left(\left(\log\left(\frac{SN}{\delta}\right)+S\log\left(e\cdot m^H\right)\right)\cdot\frac{HSN}{\varepsilon}\right).
$$

□

D.2. Sample Complexity Lower Bound

In this subsection, we prove the sample complexity lower bound (Theorem 9) for branching RL-RFE.

Proof of Theorem 9. This lower bound analysis follows the proof procedure of Theorem 2 in (Dann & Brunskill, 2015).

We consider the same instance as the proof of regret minimization lower bound in Section C.2.

The optimal policy π_* is to take action $A_j(x_i)$ at state x_i, and we have

$$
V_1^*(s_1) = m(\alpha+\eta)+m^2(\alpha+\eta)^2+\cdots+m^H(\alpha+\eta)^H = H \tag{34}
$$

Fix a policy π. For each $i \in [S-3]$, let $G_i := \{\pi(x_i) = A_j(x_i)\}$ denotes the event that policy π chooses the optimal action $A_j(x_i)$ in state x_i. Then, we have

$$
\begin{aligned}
V_1^{\pi}(s_1) =& m(\alpha+\eta)+\frac{1}{(S-3)}\sum_{i=1}^{S-3}\mathbb{E}\Big[\mathbb{1}\left\{G_i\right\}\left(m^2(\alpha+\eta)^2+\cdots+m^H(\alpha+\eta)^H\right) \\
& +\left(1-\mathbb{1}\left\{G_i\right\}\right)\left(m^2(\alpha+\eta)\alpha+m^3(\alpha+\eta)\alpha(\alpha+\eta)+\cdots+m^H(\alpha+\eta)\alpha(\alpha+\eta)^{H-2}\right)\Big] \\
=& m(\alpha+\eta)+\frac{1}{(S-3)}\sum_{i=1}^{S-3}\mathbb{E}\Big[\mathbb{1}\left\{G_i\right\}\left(m^2(\alpha+\eta)^2+\cdots+m^H(\alpha+\eta)^H\right) \\
& +\left(1-\mathbb{1}\left\{G_i\right\}\right)\left(m^2(\alpha+\eta)\alpha+m^3(\alpha+\eta)^2\alpha+\cdots+m^H(\alpha+\eta)^{H-1}\alpha\right)\Big]
\end{aligned} \tag{35}
$$

Subtracting Eq. (35) by Eq. (34), we have

$$\begin{aligned}
V_1^*(s_1) - V_1^{\pi}(s_1) =& \frac{1}{(S-3)} \sum_{i=1}^{S-3} \mathbb{E}\Big[\left(1 - \mathbb{1}\left\{G_i\right\}\right) \left(m(\alpha+\eta)m\eta + m^2(\alpha+\eta)^2 m\eta + \cdots + m^{H-1}(\alpha+\eta)^{H-1} m\eta \right) \Big] \\
=& \frac{1}{(S-3)} \sum_{i=1}^{S-3} \mathbb{E}\Big[\left(1 - \mathbb{1}\left\{G_i\right\}\right) m\eta(H-1) \Big] \\
=& m\eta(H-1) \left(1 - \frac{1}{(S-3)} \sum_{i=1}^{S-3} \mathbb{E}\Big[\mathbb{1}\left\{G_i\right\} \Big] \right)
\end{aligned}$$

The following analysis follows the proof procedure of Theorem 2 in (Dann & Brunskill, 2015). For π to be ε-optimal, we need

$$\Pr\left[m\eta(H-1) \left(1 - \frac{1}{(S-3)} \sum_{i=1}^{S-3} \mathbb{E}\Big[\mathbb{1}\left\{G_i\right\} \Big] \right) \leq \varepsilon \right] \geq 1 - \delta$$

Let $\eta = \frac{8e^2\varepsilon}{cm(H-1)}$, where c is an absolute constant that we specify later. Then, we have

$$\Pr\left[\frac{1}{(S-3)} \sum_{i=1}^{S-3} \mathbb{E}\left[\mathbb{1}\left\{G_i\right\} \right] \leq 1 - \frac{c}{8e^4} \right] \geq 1 - \delta$$

Using Markov's inequality, we have

$$1 - \delta \leq \Pr\left[\frac{1}{(S-3)} \sum_{i=1}^{S-3} \mathbb{E}\left[\mathbb{1}\left\{G_i\right\} \right] \leq 1 - \frac{c}{8e^4} \right] \leq \frac{1}{(S-3)\left(1 - \frac{c}{8e^4}\right)} \sum_{i=1}^{S-3} \Pr\left[G_i\right]$$

Since all G_i ($i \in [S-3]$) are independent of each other, there exist $\{\delta_i\}_{i \in [S-3]}$ such that $\Pr\left[\bar{G}_i\right] \leq \delta_i$ and

$$\frac{1}{(S-3)\left(1 - \frac{c}{8e^4}\right)} \sum_{i=1}^{S-3} (1 - \delta_i) \geq 1 - \delta,$$

which is equivalent to

$$\sum_{i=1}^{S-3} \delta_i \leq (S-3) \left(1 + \delta\left(1 - \frac{c}{8e^4}\right) - \left(1 - \frac{c}{8e^4}\right) \right).$$

Let ε be small enough such that $\eta = \frac{8e^2\varepsilon}{cm(H-1)} \leq \frac{1}{4}$, and let δ to be small enough $\delta \leq \frac{c}{8e^4}$. Since all G_i are independent, we can use Theorem 1 in (Mannor & Tsitsiklis, 2004) to obtain

$$\begin{aligned}
\delta_i \leq& \frac{1}{c} \left(1 + \delta\left(1 - \frac{c}{8e^4}\right) - \left(1 - \frac{c}{8e^4}\right) \right) \\
\leq& \frac{1}{c} \left(1 + \delta - \left(1 - \frac{c}{8e^4}\right) \right) \\
=& \frac{\delta}{c} + \frac{1}{8e^4} \\
\leq& \frac{2}{8e^4}
\end{aligned}$$

Let n_i denote the number of observations on state x_i. The KL-divergence of the trigger distribution on state x_i between our constructed instance and the uniform instance is $m \cdot \mathrm{KL}\left(\texttt{Bernoulli}(\alpha) || \texttt{Bernoulli}(\alpha+\eta)\right) \leq m \cdot \frac{\eta^2}{(\alpha+\eta)(1-(\alpha+\eta))} = O(m\eta^2)$ with $(\alpha+\eta)(1-(\alpha+\eta)) \geq c_1$ for some absolute positive constant c_1. Then, to ensure $\Pr\left[\bar{G}_i\right] \leq \delta_i$, we need

$$\mathbb{E}\left[n_i\right] \geq \frac{c_1 d}{m\eta^2} \log\left(\frac{c_2}{\delta_i}\right) \cdot \mathbb{1}\left\{ c\delta_i \leq \left(1 + \delta - \left(1 - \frac{c}{8e^4}\right) \right) \right\}, \tag{36}$$

where c_1 and c_2 are appropriate absolute constant, e.g., $c_1 = 400$ and $c_2 = 4$.

In the following, we compute the worst bound over all $\delta_1, \dots, \delta_{S-3}$ to ensure that $\sum_{i=1}^{S-3} \delta_i \leq (S-3)\left(1 + \delta\left(1 - \frac{c}{8e^4}\right) - \left(1 - \frac{c}{8e^4}\right)\right)$.

$$\begin{aligned} \min_{\delta_1,\dots,\delta_{S-3}} & \sum_{i=1}^{S-3} \log\left(\frac{1}{\delta_i}\right) \cdot \mathbb{1}\left\{c\delta_i \leq \left(1 + \delta - \left(1 - \frac{c}{8e^4}\right)\right)\right\} \\ s.t. & \sum_{i=1}^{S-3} \delta_i \leq (S-3)\left(1 + \delta\left(1 - \frac{c}{8e^4}\right) - \left(1 - \frac{c}{8e^4}\right)\right) \end{aligned} \tag{37}$$

Using Lemma D.1 in (Dann & Brunskill, 2015), the optimal solution of this optimization is $\delta_1 = \dots = \delta_{S-3} = z$, if $c\,(1 - \log z) \leq 1$ with $z = 1 + \delta\left(1 - \frac{c}{8e^4}\right) - \left(1 - \frac{c}{8e^4}\right)$.

Since $z \geq 1 - \left(1 - \frac{c}{8e^4}\right) = \frac{c}{8e^4}$ and $c\,(1 - \log z)$ is decreasing wi respect to z, we can obtain a sufficient condition for $c\,(1 - \log z) \leq 1$ as

$$c\left(1 - \log\left(\frac{c}{8e^4}\right)\right) \leq 1$$

Let $c = \frac{1}{10}$, which satisfies this condition. Thus, $\delta_1 = \dots = \delta_{S-3} = z$ is the optimal solution to Eq. (37).

Since in each episode, we only observe a single state x_i, the number of required episodes is at least

$$\begin{aligned} K \geq & \sum_{i=1}^{S-3} \mathbb{E}\left[n_i\right] \\ \geq & \frac{c_1 d(S-3)}{m\eta^2} \log\left(\frac{c_2}{1 + \delta\left(1 - \frac{c}{8e^4}\right) - \left(1 - \frac{c}{8e^4}\right)}\right) \\ \geq & \frac{c_1 c^2 d(S-3)m(H-1)^2}{64e^4\varepsilon^2} \log\left(\frac{c_2}{\delta\left(1 - \frac{c}{8e^4}\right) + \frac{c}{8e^4}}\right) \\ = & \Omega\left(\frac{SNH^2}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right) \end{aligned}$$

□