An Analysis of Abstracted Model-Based Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

Many methods for Model-based Reinforcement learning (MBRL) in Markov decision processes (MDPs) provide guarantees for both the accuracy of the model they can deliver and the learning efficiency. At the same time, state abstraction techniques allow for a reduction of the size of an MDP while maintaining a bounded loss with respect to the original problem. Therefore, it may come as a surprise that no such guarantees are available when combining both techniques, i.e., where MBRL merely observes abstract states. Our theoretical analysis shows that abstraction can introduce a dependence between samples collected online (e.g., in the real world). That means that, without taking this dependence into account, results for MBRL do not directly extend to this setting. Our result shows that we can use concentration inequalities for martingales to overcome this problem. This result makes it possible to extend the guarantees of existing MBRL algorithms to the setting with abstraction. We illustrate this by combining R-MAX, a prototypical MBRL algorithm, with abstraction, thus producing the first performance guarantees for 'Abstracted RL': model-based reinforcement learning with an abstract model.

1 Introduction

Tabular Model-based Reinforcement Learning (MBRL) methods provide guarantees that show they can learn efficiently in Markov decision processs (MDPs) (Brafman & Tennenholtz, 2002; Strehl & Littman, 2008; Jaksch et al., 2010; Fruit et al., 2018; Talebi & Maillard, 2018; Zhang & Ji, 2019; Bourel et al., 2020). They do this by finding solutions to a fundamental problem for Reinforcement Learning (RL), the exploration-exploitation dilemma: when to take actions to obtain more information (explore) and when to take actions that maximize reward based on the current knowledge (exploit). However, MDPs can be huge, which can be problematic for tabular methods. One way to deal with large problems is by using abstractions, such as state abstractions (Li, 2009; Abel et al., 2016). State abstractions reduce the size of the problem by aggregating together states according to different criteria, depending on the specific type of abstraction. We can view state abstraction as a special case of function approximation, where every state maps to its abstract state (Mahadevan, 2010), and we can roughly divide them into *exact* and *approximate* abstractions (Li, 2009; Abel et al., 2016). Approximate abstractions relax the criteria of exact abstractions, and therefore allow for a larger reduction in the state space. In our setting, we assume that the environment returns states s, but the agent only observes abstract states $\phi(s)$ (see Figure 1). This setting, which has been considered before (Ortner et al., 2014a; Abel et al., 2018),[1] is what we call *Abstracted RL*.

Abstracted RL corresponds to RL in a Partially Observable MDP (POMDP) (Kaelbling et al., 1998), as previously described (Bai et al., 2016). Specifically, the abstraction function serves as an observation function. Rather than observing its true state s, the agent observes the abstract state $\phi(s)$ and its policy chooses an action based on this abstract state. It is well known that policies for POMDPs that only base their action on the last

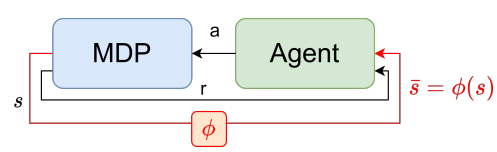

Figure 1: Abstracted RL, the agent receives the abstract state $\bar{s} = \phi(s)$ as an observation instead of the state s. Image based on Abel et al. (2018).

[1] We refer to Section 4 for a comparison with the related work.

observation can be arbitrarily bad (Singh et al., 1994). However, this worst-case does not apply when ϕ is an *exact model similarity abstraction* [2] (Li, 2009), because the resulting problem can be considered an MDP; this abstraction maps states to the same abstract state only when their reward and transition functions in the abstract space are the same (Li et al., 2006). We focus on the related *approximate model similarity abstraction* (Abel et al., 2016), which maps states to the same abstract state only when their reward and transition functions in the abstract space are close. Intuitively, because of its connection to the exact model similarity, one could expect that for this abstraction the worst-case also does not apply. However, MBRL methods typically (e.g., (Strehl & Littman, 2008; Jaksch et al., 2010; Fruit et al., 2018; Bourel et al., 2020)) use results that rely on the assumption of independent and identically distributed (i.i.d.) samples to prove efficient learning. With abstraction, the transitions between abstract states need not be Markov, and the samples may depend on the history.

We analyze collecting samples in Abstracted RL and prove that, with abstraction, samples are not guaranteed to be independent. This means that *most guarantees of existing MBRL methods do not hold in the online Abstracted RL setting.* [3] The primary technical result in this work shows that we can still learn an accurate model in Abstracted RL by replacing concentration inequalities that rely on independent samples with a concentration inequality for martingales. For approximate model similarity abstraction, this allows us to extend the guarantees of MBRL methods to Abstracted RL, which we illustrate for the prototypical R-MAX algorithm (Brafman & Tennenholtz, 2002).

The outline is as follows: Sections 2.1 and 2.2 introduce MBRL and state abstraction, respectively. Section 3.1 analyzes sample collection in Abstracted RL and shows that samples are not necessarily independent in this setting. Then Section 3.2 gives the main result that inequalities for martingales can be used to guarantee accurate model learning in Abstracted RL. Finally, Section 4 covers related work, Section 5 provides a discussion, and Section 6 closes with conclusions.

2 Background

Section 3 will cover the combination of MBRL and abstraction in MDPs, in this section we introduce the required background. In Section 2.1, we first briefly introduce the MDP and MBRL. Then, in Section 2.2, we cover state abstraction for MDPs when the model is known and give the definition of the approximate model similarity abstraction. We need state abstraction in general to show in Section 3.1 that the combination of MBRL and abstraction leads to a dependence of samples. Later, in Section 3.2, we use results for approximate state abstraction with known models to establish bounds for Abstracted RL.

2.1 Model-Based RL

As is typical for RL problems, we assume the environment the agent is acting in can be represented by an infinite horizon MDP $M \triangleq \langle S, A, T, R, \gamma \rangle$ (Puterman, 2014). Here S is a finite set of states $s \in S$, A a finite set of actions $a \in A$, T a transition function $T(s'|s,a) = \Pr(s'|s,a)$, R a reward function $R(s,a)$ which gives the reward received when the agent executes action a in state s, and γ is a discount factor with $0 \leq \gamma \leq 1$ that determines the importance of future rewards. The agent's goal is to find an optimal policy $\pi^* : S \to A$, i.e., a policy that maximizes the expectation of the cumulative reward in the MDP. $V^\pi(s)$ denotes the expected value of the cumulative reward under policy π starting from state s. Similarly, $Q^\pi(s,a)$ denotes the expected value of the cumulative reward when first taking action a from state s and then following policy π afterward.

MBRL methods learn a model from the experience that the agent gains by taking actions and observing the rewards it gets and the states it reaches. For a fixed state-action pair (s,a), we let $\tau_1, \tau_2, \cdots, \tau_{N(s,a)}$ be the first $N(s,a)$ time steps at which the agent took action a in state s. The first $N(s,a)$ states s' that the agent reached after taking action a in state s are stored as the sequence $Y_{s,a} \triangleq (s'^{(\tau_1+1)}, s'^{(\tau_2+1)}, \cdots, s'^{(\tau_{N(s,a)}+1)})$.

[2] Also known as stochastic bisimulation (Givan et al., 2003).

[3] Of course, certain guarantees on the combination of abstraction and RL are known. However, in most related work in abstraction settings (e.g., abstraction selection), the complication of samples not being independent does not occur due to particular assumptions (Paduraru et al., 2008; Hallak et al., 2013; Maillard et al., 2013; Majeed & Hutter, 2018; Ortner et al., 2019; Du et al., 2019). Section 4 gives details for individual papers.

We use Y to refer to the collection of all $Y_{s,a}$. Typically, in MBRL, the obtained experience is used to construct the empirical model T_Y (Brafman & Tennenholtz, 2002; Strehl & Littman, 2008; Jaksch et al., 2010; Fruit et al., 2018; Talebi & Maillard, 2018; Zhang & Ji, 2019; Bourel et al., 2020). This model is constructed simply by counting how often the agent reached a particular next state s' and normalizing the obtained quantity by the total count:

$$\forall s' \in S : T_Y(s'|s,a) \triangleq \frac{1}{N(s,a)} \sum_{i=1}^{N(s,a)} \mathbb{1}\{Y_{s,a}^{(\tau_i+1)} = s'\}. \tag{1}$$

Here $\mathbb{1}\{\cdot\}$ denotes the indicator function of the specified event, i.e., $\mathbb{1}\{Y_{s,a}^{(\tau_i+1)} = s'\}$ is 1 if $Y_{s,a}^{(\tau_i+1)} = s'$ and 0 otherwise.

The quality of the empirical model T_Y is crucial for performance guarantees, irrespective of the form of the guarantee, e.g., PAC-MDP (Strehl & Littman, 2008) or regret (Jaksch et al., 2010). The quality of the empirical model is high when the distance between $T_Y(\cdot|s,a)$ and $T(\cdot|s,a)$ is small. We can, for instance, measure this distance with the L_1 norm, defined as follows:

$$||T_Y(\cdot|s,a) - T(\cdot|s,a)||_1 \triangleq \sum_{s' \in S} |T_Y(s'|s,a) - T(s'|s,a)|. \tag{2}$$

Concentration inequalities are often used to guarantee that, with enough samples, this distance will be small, e.g.:

Lemma 1 (L_1 inequality (Weissman et al., 2003))**.** *Let $Y_{s,a} = Y_{s,a}^{(1)}, Y_{s,a}^{(2)}, \cdots, Y_{s,a}^{(N(s,a))}$ be i.i.d. random variables distributed according to $T(\cdot|s,a)$. Then, for all $\epsilon > 0$,*

$$\Pr(||T_Y(\cdot|s,a) - T(\cdot|s,a)||_1 \geq \epsilon) \leq (2^{|S|} - 2)e^{-\frac{1}{2}N(s,a)\epsilon^2}. \tag{3}$$

These inequalities typically make use of the fact that samples are i.i.d. In most MBRL settings (without abstraction), this is not a problem under some assumptions, e.g., when the MDP is communicating. [4] Due to the Markov property, the obtained samples are i.i.d. in this case. In this way, MBRL can upper bound the probability that the empirical model $T_Y(\cdot|s,a)$ will be far away ($\geq \epsilon$) from the actual model $T(\cdot|s,a)$. When the empirical model is accurate, a policy based on this model leads to near-optimal performance in the MDP M (Brafman & Tennenholtz, 2002; Strehl & Littman, 2008; Jaksch et al., 2010; Bourel et al., 2020).

2.2 State Abstraction for Known Models

We can formulate state abstraction as a mapping from states to abstract states (Li et al., 2006). This mapping is done with an abstraction function ϕ, a surjective function that maps from states $s \in S$ to abstract states $\bar{s} \in \bar{S}$: $\phi(s) : S \to \bar{S}$. We use the $^-$ notation to refer to the abstract space and define $\bar{S}$ as $\bar{S} = \{\phi(s)|s \in S\}$. We slightly overload the definition of $\bar{s}$ to be able to write $s \in \bar{s}$. In this case, $\bar{s}$ is the set of states that map to $\bar{s}$, i.e., $\bar{s} = \{s \in S \mid \phi(s) = \bar{s}\}$. This form of state abstraction is general, and clusters states with different dynamics into abstract states. We assume that the state abstraction deterministically maps states to an abstract state. Since each state maps to precisely one abstract state and multiple states can map to the same abstract state, the abstract state space is typically (much) smaller than the original state space, $|\bar{S}| \leq |S|$.

We focus on a type of abstraction *approximate model similarity abstraction* (Abel et al., 2016), also known as approximate stochastic bisimulation (Dean et al., 1997; Givan et al., 2003). In this abstraction, two states can map to the same abstract state only if their behavior is similar in the abstract space, i.e., when the reward function and the transitions to abstract states are close. We can determine the transition probability to an abstract state $P(\bar{s}'|s,a)$ as:

$$\Pr(\bar{s}'|s,a) = \sum_{s' \in \bar{s}'} T(s'|s,a). \tag{4}$$

[4] An MDP is communicating if, for all $s_1, s_2 \in S$, a deterministic policy exists that eventually leads from s_1 to s_2 (Puterman, 2014).

Then, we can use equation 4 to define approximate model similarity abstraction:

Definition 1. *An approximate model similarity abstraction, $\phi_{model,\eta_R,\eta_T}$, for fixed η_R, η_T, satisfies*

$$\phi_{model,\eta}(s_1) = \phi_{model,\eta}(s_2) \implies \forall a \in A : |R(s_1, a) - R(s_2, a)| \leq \eta_R,$$
$$\forall \bar{s}' \in \bar{S}, a \in A : |\Pr(\bar{s}'|s_1, a) - \Pr(\bar{s}'|s_2, a)| \leq \eta_T. \tag{5}$$

From now on, we will refer to $\phi_{\text{model},\eta_R,\eta_T}$ as ϕ. We note that this abstraction is still quite generic. It can cluster together states that have different transition and reward functions.

In the planning setting, where the model is known a priori, we can use the abstraction function ϕ to construct an abstract MDP. An abstract MDP can be helpful because it is smaller, making it easier to find a solution, and a solution for the abstract MDP can work well in the original MDP (Li et al., 2006; Abel et al., 2016). We construct an abstract MDP $\bar{M}_\omega$ from the model of an MDP M, an abstraction function ϕ, and an action-specific weighting function ω. [5] The weighting function ω gives a weight to every state-action pair: $\forall s \in S,\ a \in A : 0 \leq \omega(s, a) \leq 1$. The weights of the state-action pairs associated with an abstract state $\bar{s}$ sum up to 1: $\sum_{s' \in \phi(s)} \omega(s', a) = 1$. We can use the weighting function to create an abstract transition and reward function, which are weighted averages of the original transition and reward functions. In this way, from M, ϕ, and any ω, we can *construct* an abstract MDP $\bar{M}_\omega$:

Definition 2. *Given an MDP M, ϕ, and ω, $\bar{M}_\omega = \langle \bar{S}, A, \bar{T}_\omega, \bar{R}_\omega \rangle$ is constructed as: $\bar{S} = \{\phi(s) \mid s \in S\}, A = A$,*

$$\forall \bar{s} \in \bar{S},\ a \in A : \bar{R}_\omega(\bar{s}, a) = \sum_{s \in \bar{s}} \omega(s, a) R(s, a), \tag{6}$$

$$\forall \bar{s}, \bar{s}' \in \bar{S},\ a \in A : \bar{T}_\omega(\bar{s}'|\bar{s}, a) = \sum_{s \in \bar{s}} \sum_{s' \in \bar{s}'} \omega(s, a) T(s'|s, a). \tag{7}$$

Note that the abstract MDP $\bar{M}_\omega$ itself is an MDP. So we can use planning methods for MDPs to find an optimal policy $\bar{\pi}^*$ for $\bar{M}_\omega$. A desirable property of the approximate model similarity abstraction is that we can upper bound the difference between the optimal value V^* in M and the value $V^{\bar{\pi}^*}$ obtained when following the policy $\bar{\pi}^*$ in M. These bounds exists in different forms (Dearden & Boutilier, 1997; Abel et al., 2016; Taïga et al., 2018). For completeness, we give these bounds for both the undiscounted finite horizon and the discounted infinite horizon:

Theorem 1. *Let $M = \langle S, A, T, R \rangle$ be an MDP and $\bar{M} = \langle \bar{S}, A, \bar{T}, \bar{R} \rangle$ an abstract MDP, for some defined abstract transitions and rewards. We assume that*

$$\forall \bar{s}, \bar{s}' \in \bar{S},\ s \in \bar{s},\ a \in A : |\bar{T}(\bar{s}'|\bar{s}, a) - \Pr(\bar{s}'|s, a)| \leq \eta_T \tag{8}$$
$$\text{and } |\bar{R}(\bar{s}, a) - R(s, a)| \leq \eta_R. \tag{9}$$

Then, for a finite horizon problem with horizon h we have:

$$V^*(s) - V^{\bar{\pi}^*}(s) \leq 2h\eta_R + (h+1)h\eta_T|\bar{S}|R_{max}. \tag{10}$$

And for a discounted infinite horizon problem with discount γ we have:

$$V^*(s) - V^{\bar{\pi}^*}(s) \leq \frac{2\eta_R}{1-\gamma} + \frac{2\gamma\eta_T|\bar{S}|R_{max}}{(1-\gamma)^2}. \tag{11}$$

The proof of Theorem 1 is in Appendix A.3.

[5] The action-specific weighting function is more general than the typically used weighting function, which is not action-specific and only depends on the state s (Li et al., 2006). More formally, it is the case where $\forall a, a' \in A,\ s \in S : \omega(s, a) = \omega(s, a')$.

3 Abstracted MBRL

In Abstracted RL, we have an abstraction function ϕ and, in contrast to the planning setting in Section 2.2, act in an MDP M of which we do *not* know the transition and reward functions. We want to extend the results of MBRL to Abstracted RL. Section 3.1 first analyzes how abstraction leads to dependence between samples. Then, in Section 3.2, we show how concentration inequalities for martingales can be used to still learn an accurate model in Abstracted RL. To illustrate how this result can be used to extend the results of MBRL methods to Abstracted RL, we extend the results of the R-MAX algorithm (Brafman & Tennenholtz, 2002). R-MAX is a well-known and straightforward method that guarantees sample efficient learning.

In Abstracted RL, the agent collects data for every abstract state-action pair $(\bar{s}, a)$, stored as sequences $\bar{Y}_{\bar{s},a}$:

$$\bar{Y}_{\bar{s},a} : \{\bar{s}'^{(\tau_1+1)}, \bar{s}'^{(\tau_2+1)}, \cdots, \bar{s}'^{(\tau_{N(\bar{s},a)}+1)}\}. \tag{12}$$

Like in equation 1, we construct an empirical model $\bar{T}_Y$, now looking at the abstract next-states that the agent reached:

$$\bar{T}_Y(\bar{s}'|\bar{s}, a) \triangleq \frac{1}{N(\bar{s}, a)} \sum_{i=1}^{N(\bar{s},a)} \mathbb{1}\{\bar{Y}^{(i)}_{\bar{s},a} = \bar{s}'\}. \tag{13}$$

Suppose we could guarantee that the empirical model $\bar{T}_Y$ was equal, or close, to the transition function $\bar{T}_\omega$ of an abstract MDP $\bar{M}_\omega$ constructed from the true MDP with ϕ and a valid ω. In that case, we could bound the loss in performance due to applying the learned policy $\bar{\pi}^*$ to M instead of the optimal policy π^* (Abel et al., 2016; Taïga et al., 2018). Our main question is: do the finite-sample model learning guarantees of MBRL algorithms still hold in the Abstracted RL setting?

3.1 Abstracted RL Can Lead to Dependent Samples

We want to guarantee that the empirical model $\bar{T}_Y$ is close to the transition model of an abstract MDP $\bar{M}_\omega$. Before defining this transition model of $\bar{M}_\omega$, we examine the data collection. In the online data collection, the agent obtains a sample for $\bar{Y}_{\bar{s},a}$ when it is in a state $s \in \bar{s}$ and takes action a. Specifically, the agent obtains the i-th sample $\bar{Y}^{(i)}_{\bar{s},a} = \bar{s}'^{\tau_1+1}$ from state $X^{(i)}_{\bar{s},a} = s^{\tau_i} \in \bar{s}$:

$$\bar{Y}^{(i)}_{\bar{s},a} \sim \Pr(\cdot|X^{(i)}_{\bar{s},a} = s^{\tau_i}, a). \tag{14}$$

Let $X_{\bar{s},a} = (X^{(i)}_{\bar{s},a})_{i=1}^{N(\bar{s},a)}$ denote the sequence of states $s \in \bar{s}$ from which the agent took action a. Each state s gets a weight according to how often it appears in $X_{\bar{s},a}$, which we formalize with the weighting function ω_X: $\forall s \in \bar{s}, a \in A : \omega_X(s, a) \triangleq \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} \mathbb{1}\{X^{(i)}_{\bar{s},a} = s\}$. We use ω_X to define $\bar{T}_{\omega_X}$ analogous to equation 7:

$$\forall \bar{s}, \bar{s}' \in \bar{S}, a \in A : \bar{T}_{\omega_X}(\bar{s}'|\bar{s}, a) \triangleq \sum_{s \in \bar{s}} \omega_X(s, a) \sum_{s' \in \bar{s}'} T(s'|s, a). \tag{15}$$

We want a concentration inequality to provide bounds on the deviation of the empirical model $\bar{T}_Y$ from $\bar{T}_{\omega_X}$; we refer to this inequality as the abstract L1 inequality, similar in form to equation 3:

$$P(|\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_X}(\cdot|\bar{s}, a)|_1 \geq \epsilon) \leq \delta, \tag{16}$$

where $\bar{T}_Y(\cdot|\bar{s}, a)$ is defined according to equation 13 and $\bar{T}_{\omega_X}$ according to equation 15.

Suppose we could directly obtain i.i.d. samples from $\bar{T}_{\omega_X}$ and base our empirical model $\bar{T}_Y$ on the obtained samples. In that case, we could show that the abstract L1 inequality holds by applying Lemma 1. This lemma would be applicable because we could obtain a number $N(\bar{s}, a)$ of i.i.d. samples per abstract state-action pair, distributed according to $\bar{T}_{\omega_X}(\cdot|\bar{s}, a)$. However, the samples are not guaranteed to be i.i.d. in Abstracted RL. The samples may neither be identically distributed nor independent. We will first cover the distribution of the samples and show that samples not being identically distributed is not a problem. Then we prove that samples are not guaranteed to be independent. Afterward, Section 3.2 shows that we can still learn when the samples are dependent.

Non-Identically Distributed The samples are not necessarily identically distributed in Abstracted RL since the agent obtains a sample $\bar{Y}^{(i)}$ when taking action a from state $X^{(i)}_{\bar{s},a} = s \in \bar{s}$, as in equation 14. If $X^{(i)}_{\bar{s},a} \neq X^{(j)}_{\bar{s},a}$, these states can have different transition distributions. This implies that in general we might not be able to apply Lemma 1, because it assumes identically distributed random variables. However, different distributions by themselves need not be a problem; we show that the result also holds when the random variables are not identically distributed:

Lemma 2. *Let $X_{\bar{s},a} = s_1, \cdots, s_m$ be a sequence of states $s \in \bar{s}$ and let $\bar{\boldsymbol{Y}}_{\bar{s},a} = \bar{Y}^{(1)}, \bar{Y}^{(2)}, \cdots, \bar{Y}^{(m)}$ be independent random variables distributed according to $\Pr(\cdot|s_1, a), \cdots, \Pr(\cdot|s_m, a)$ (equation 4). Then, for all $\epsilon > 0$,*

$$\Pr(||\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_X}(\cdot|\bar{s}, a)||_1 \geq \epsilon) \leq (2^{|\bar{S}|} - 2)e^{-\frac{1}{2}m\epsilon^2}. \tag{17}$$

The proof can be found in Appendix B. Therefore, if the samples in Abstracted RL were independent, then we could apply Lemma 2 to guarantee an accurate model.

Independence One could be tempted to assume the samples are independent, i.e.,

$$\forall \bar{s}'_1, \cdots, \bar{s}'_m \in (\bar{S})^m : \Pr(\bar{Y}^{(1)}_{\bar{s},a} = \bar{s}'_1, \cdots, \bar{Y}^{(m)}_{\bar{s},a} = \bar{s}'_m) = \Pr(\bar{Y}^{(1)}_{\bar{s},a} = \bar{s}'_1) \cdots P(\bar{Y}^{(m)}_{\bar{s},a} = \bar{s}'_m). \tag{18}$$

However, this is not true in general in Abstracted RL:

Observation 1. *When collecting samples online using an abstraction function, such samples are not necessarily independent.*

Samples can be dependent when 1) samples are collected online in the real environment, of which we do not know the transitions, and 2) the samples are collected for abstract states $\bar{s}$. Observation 1 can be understood from the perspective that the abstracted RL problem corresponds to RL in a POMDP (Bai et al., 2016). The corresponding POMDP uses the abstraction function as the observation function and the abstract states as observations. Since the transitions between observations need not be Markov in POMDPs, the samples from abstract states can depend on the history. While this observation may be clear from the POMDP perspective, work in Abstracted RL regularly assumes (explicity or implicitly) that independent samples can somehow be obtained (Paduraru et al., 2008; Ortner et al., 2014a; Jiang et al., 2015; Ortner et al., 2019). In the following counterexample, we rigorously show that samples are not necessarily independent.

Counterexample We use the example MDP and abstraction in Figure 2, where we have four states, three abstract states, and only one action. Since the example MDP has only one action, we omit the action from the notation. We examine the transition function of abstract state A, $\bar{T}_Y(\cdot|A)$ and consider the first two times we transition from A. These two transition samples, $\bar{s}'_1$ and $\bar{s}'_2$, are the first two entries in $\bar{Y}_A$. We show that the samples are not independent for at least one combination of $\bar{s}'_1$ and $\bar{s}'_2$.

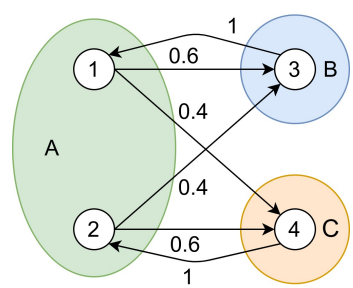

Figure 2: Simple MDP, with only 1 action, and abstraction. The small circles are states (1,2,3,4). A, B and C are the abstract states. The arrows show the transition probabilities, e.g. $P(3|1) = 0.6$.

Let $\bar{s}'_1 = \bar{s}'_2 = B$, i.e., the first two times we experience a transition from the abstract state A, we end up in B. We denote the i-th experienced transition from abstract state A as $\bar{Y}^{(i)}_A$. Let state 1 be the starting state.

We start with the product of the probabilities:

$$\Pr(\bar{Y}^{(1)}_A = B) \Pr(\bar{Y}^{(2)}_A = B). \tag{19}$$

We have $\Pr(\bar{Y}^{(1)}_A = B) = \Pr(B|1) = 0.6$ for the first term since state 1 is the starting state. The second term is more complex than the first since it includes the probability of

starting the transition from state 1 and state 2. We have:

$$\Pr(\bar{Y}_A^{(2)} = B) = \sum_{\bar{s} \in \bar{S}} \Pr(\bar{Y}_A^{(2)} = B | \bar{Y}_A^{(1)} = \bar{s}) \Pr(\bar{Y}_A^{(1)} = \bar{s}) \tag{20}$$

$$= \Pr(\bar{Y}_A^{(2)} = B | \bar{Y}_A^{(1)} = A) \Pr(\bar{Y}_A^{(1)} = A) + \Pr(\bar{Y}_A^{(2)} = B | \bar{Y}_A^{(1)} = B) \Pr(\bar{Y}_A^{(1)} = B)$$

$$+ \Pr(\bar{Y}_A^{(2)} = B | \bar{Y}_A^{(1)} = C) \Pr(\bar{Y}_A^{(1)} = C). \tag{21}$$

$$= \Pr(\bar{Y}_A^{(2)} = B | \bar{Y}_A^{(1)} = B) \Pr(\bar{Y}_A^{(1)} = B) + \Pr(\bar{Y}_A^{(2)} = B | \bar{Y}_A^{(1)} = C) \Pr(\bar{Y}_A^{(1)} = C). \tag{22}$$

$$= \Pr(Y_A^{(2)} = 3 | \bar{Y}_A^{(1)} = 3) \Pr(Y_A^{(1)} = 3) + \Pr(Y_A^{(2)} = 3 | Y_A^{(1)} = 4) \Pr(Y_A^{(1)} = 4) \tag{23}$$

$$= 0.6 \cdot 0.6 + 0.4 \cdot 0.4 = 0.52. \tag{24}$$

For the step from equation 21 to equation 22, $\Pr(\bar{Y}_A^{(1)} = A)$ is 0 because there is no transition from a state in A to a state in A. Then, from equation 22 to equation 23, we use that both abstract states B and C consist of exactly 1 state. So, e.g., $\Pr(\bar{Y}_A^{(2)} = B | \bar{Y}_A^{(1)} = B) = \Pr(Y_A^{(2)} = 3 | \bar{Y}_A^{(1)} = 3)$. So, for the product of the probabilities, we end up with: $\Pr(\bar{Y}_A^{(1)} = B) \Pr(\bar{Y}_A^{(2)} = B) = 0.6 \cdot 0.52 = 0.321$.

For the joint probability, we have:

$$Pr(\bar{Y}_A^{(1)} = B, \bar{Y}_A^{(2)} = B) = \Pr(\bar{Y}_A^{(1)} = B) \Pr(\bar{Y}_A^{(2)} = B | \bar{Y}_A^{(1)} = B) \tag{25}$$

$$= \Pr(B|1)(\Pr(B|1) \Pr(1|B)) \tag{26}$$

$$= 0.6 \cdot (0.6 \cdot 1) \tag{27}$$

$$= 0.6 \cdot 0.6 \tag{28}$$

$$= 0.36. \tag{29}$$

Here, $\Pr(\bar{Y}_A^{(2)} = B | \bar{Y}_A^{(1)} = B) = \Pr(B|1) \Pr(1|B)$ because the first transition ends in state B and we always transition to state 1 from state B. Hence, $\Pr(\bar{Y}_A^{(2)} = B | \bar{Y}_A^{(1)} = B) = \Pr(B|1) \Pr(1|B) = 0.6 \cdot 1$.

Combining the joint probability and the product of probabilities, we end up with:

$$0.36 = \Pr(\bar{Y}_A^{(1)} = B, \bar{Y}_A^{(2)} = B) \neq \Pr(\bar{Y}_A^{(1)} = B) \Pr(\bar{Y}_A^{(2)} = B) = 0.6 \cdot 0.52. \tag{30}$$

Thus, the samples are not independent. Leading us to the second observation.

Observation 2. *As independence cannot be guaranteed, Lemmas 1 and 2 cannot be readily applied to show that the abstract L1 inequality holds.*

This claim follows from the fact that Lemmas 1 and 2 both use the assumption of independence in their proofs. It would still be possible to obtain independent samples if we could, for example, have access to a simulator of the problem. In that case, it is still possible to give guarantees on the accuracy of the model, which we show in Appendix E. However, we consider the setting where a simulator is not available.

3.2 Guarantees for Abstract Model Learning Using Martingales

Now we want to give a guarantee in the form of the abstract L1 inequality from equation 16. In Section 3.1, we found this was not possible with concentration inequalities such as Hoeffding's inequality because the samples are not guaranteed to be independent. Here we consider a related bound for weakly dependent samples, the Azuma-Hoeffding inequality. This inequality makes use of the properties of a martingale difference sequence, which are slightly weaker than independence:

Definition 3 (Martingale difference sequence (Azuma, 1967))**.** *The sequence $Z_1, Z_2, \cdots$ is a martingale difference sequence if, $\forall i$, it satisfies the following conditions:*

$$E[Z_i | Z_1, Z_2, \cdots, Z_{i-1}] = 0,$$
$$|Z_i| < \infty.$$

The properties of the martingale difference sequence can be used to obtain the following concentration inequality:

Lemma 3 (Azuma-Hoeffding Inequality (Hoeffding, 1963; Azuma, 1967))**.** *If the random variables $Z_1, Z_2, \cdots$ form a martingale difference sequence (Def. 3), with $|Z_i| \leq b$, then*

$$\Pr(\sum_{i=1}^{n} Z_i > \epsilon) \leq e^{-\frac{\epsilon^2}{2b^2 n}}. \tag{31}$$

Our main result, Theorem 2, shows that we can use Lemma 3 to obtain a concentration inequality for the abstract transition function in Abstracted RL. Specifically, we show that, with high probability, the empirical abstract transition function $\bar{T}_Y$ will be close to the abstract transition function $\bar{T}_{\omega_X}$:

Theorem 2 (Abstract L1 inequality)**.** *If an agent has access to a state abstraction function ϕ and uses this to collect data for any abstract state-action pair $(\bar{s}, a)$ by acting in an MDP M according to a policy $\bar{\pi}$, we have that the following holds with a probability of at least $1 - \delta$ for a fixed value of $N(\bar{s}, a)$:*

$$||\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_X}(\cdot|\bar{s}, a)||_1 \leq \epsilon, \tag{32}$$

where $\delta = 2^{|\bar{S}|} e^{-\frac{1}{8} N(\bar{s},a)\epsilon^2}$.

Here we use the definitions of $\bar{T}_Y(\cdot|\bar{s}, a)$ (equation 13) and $\bar{T}_{\omega_X}(\cdot|\bar{s}, a)$ (equation 15). To apply Lemma 3, we define a suitable Z_i in the proof. The proof can be found in Appendix C.

3.3 Abstracted R-MAX

Finally, we show how we can use Theorem 2 to provide guarantees for MBRL methods in Abstracted RL with an *approximate model similarity abstraction*. We illustrate this using the R-MAX algorithm (Brafman & Tennenholtz, 2002). We start with a short description of R-MAX and how it operates with abstraction.

The R-MAX algorithm maintains a model of the environment. It uses this model to compute a policy periodically and then follows this policy for several steps. Initially, all the state-action pairs are *unknown* and the algorithm optimistically initializes their reward and transition functions: $R(s, a) = R_{\max}$ (the maximum reward), $T(s|s, a) = 1$, and $\forall s' \neq s : T(s'|s, a) = 0$. This initialization means that, in the model the algorithm maintains, these unknown (s, a) lead to the maximum reward, hence the name R-MAX. A state-action pair's transition and reward function are only updated once they have been visited sufficiently often, at which point the state-action pair is considered *known*. Together, this ensures that the algorithm explores sufficiently. During execution, the algorithm operates in episodes of n-steps. At the start of every episode it calculates an optimal n-step policy and follows this for n timesteps, or untill a state-action pair becomes known. Once all the state-action pairs are known it calculates the optimal policy for the final model and then runs this indefinitely. The algorithm has the following guarantee:

Theorem 3 (R-MAX in MDPs (Brafman & Tennenholtz, 2002))**.** *Given an MDP M, with $|S|$ states and $|A|$ actions, and inputs ϵ and δ. With probability of at least $1 - \delta$ the R-MAX algorithm will attain an expected average return of $Opt(\prod_M(\epsilon, T_\epsilon)) - 2\epsilon$ within a number of steps polynomial in $|S|, |A|, \frac{1}{\epsilon}\frac{1}{\delta}, T_\epsilon$. Where T_ϵ is the ϵ-return mixing time of the optimal policy, the policies for M whose ϵ-return mixing time is T_ϵ are denoted by $\prod_M(\epsilon, T_\epsilon)$, the optimal expected T_ϵ-step undiscounted average return achievable by such policies are denoted by $Opt(\prod_M(\epsilon, T_\epsilon))$.*

Here T_ϵ is the ϵ-return mixing time of a policy π, it is the minimum number of steps needed to guarantee that the expected average return is within ϵ of the optimal expected average return (Kearns & Singh, 2002).

For R-MAX with abstraction, We make the following assumptions that stem from the original analysis: we assume that the MDP is ergodic (Puterman, 2014), [6] that we know S and A, that the reward function is deterministic, and that we know the minimum and maximum reward. W.l.o.g., we assume the rewards are between 0 and $R_{\max}$, with $0 < R_{\max} < \infty$. We add the assumption that the agent has access to an approximate model similarity abstraction function ϕ.

[6] An ergodic, or recurrent, MDP is an MDP where every state is recurrent under every stationary policy, i.e., asymptotically, every state will be visited infinitely often (Puterman, 2014).

Algorithm 1 Procedure: Abstracted R-MAX

Input: $\phi, \delta, \epsilon, T_\epsilon$
for all $(\bar{s}, a) \in \bar{S} \times A$ **do**
  $\bar{T}_Y(\bar{s}|\bar{s}, a) = 1$
  $\bar{R}_Y(\bar{s}, a) = 1$
  $\bar{Y}_{\bar{s},a} = [\ ]$
end for
$\bar{M}_Y = \langle \bar{S}, A, \bar{T}_Y, \bar{R}_Y \rangle$
Select m, the number of samples required per (abstract) state-action pair to make them known.
// While there is still an unknown state-action pair.
while $\min_{(\bar{s},a)} |Y_{\bar{s},a}| < m$ **do**
  Compute optimal T_ϵ-step policy $\bar{\pi}$ in $\bar{M}_Y$ for the current abstract state.
  for T_ϵ timesteps **do**
    $\bar{s} = \phi(s)$
    $a = \bar{\pi}(\bar{s})$
    $s', r = \text{Step}(s, a)$
    $s = s'$
    if $|\bar{Y}_{\bar{s},a}| < m$ **then**
      $\bar{Y}_{\bar{s},a}$.append$(\phi(s'))$
      if $|\bar{Y}_{\bar{s},a}| = m$ **then**
        // State-action pair has become known.
        for all $\bar{s}' \in \bar{S}$ **do**
          $\bar{T}_Y(\bar{s}'|\bar{s}, a) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{\bar{Y}^{(i)}_{\bar{s},a} = \bar{s}'\}$
        end for
        $\bar{R}_Y(\bar{s}, a) = r$
        break
      end if
    end if
  end for
end while
Compute optimal policy $\bar{\pi}^*$ for $\bar{M}$ and run indefinitely.

Algorithm 1 shows the procedure for Abstracted R-MAX. It follows the same steps as the orignal algorithm, except that it makes use of an abstraction function ϕ and maintains an abstract model. As in the original, the input to the algorithm is the allowed failure probability δ, the error bound ϵ, and the ϵ-return mixing time T_ϵ of an optimal policy. We add the abstraction function ϕ as a new input. The algorithm uses this function to observe $\phi(s)$, as in Figure 1, and it builds an empirical (abstract) model from the observations it obtains.

Because the algorithm uses an abstraction function ϕ, we cannot guarantee the ϵ error bound. However, with Theorem 5 we can still guarantee an error bound that is a function of ϵ and the error η of the abstraction, thus providing the first finite-sample guarantees for Abstracted RL:

Theorem 4. *Given an MDP M, an approximate model similarity abstraction ϕ, with η_R and η_T, and inputs $|\bar{S}|, |A|, \epsilon, \delta, T_\epsilon$. With probability of at least $1-\delta$ the R-MAX algorithm adapted to abstraction (Algorithm 1) will attain an expected average return of $Opt(\prod_M(\epsilon, T_\epsilon)) - 3g(\eta_T, \eta_R) - 2\epsilon$ within a number of steps polynomial in $|\bar{S}|, |A|, \frac{1}{\epsilon}\frac{1}{\delta}, T_\epsilon$. Where T_ϵ is the ϵ-return mixing time of the optimal policy, the policies for M whose ϵ-return mixing time is T_ϵ are denoted by $\prod_M(\epsilon, T_\epsilon)$, the optimal expected T_ϵ-step undiscounted average return achievable by such policies are denoted by $Opt(\prod_M(\epsilon, T_\epsilon))$, and*

$$g(\eta_T, \eta_R) = T_\epsilon \eta_R + \frac{(T_\epsilon - 1)T_\epsilon}{2} \eta_T |\bar{S}|.$$

The proof can be found in Appendix D.2 and follows the line of the original R-MAX proof, using the assumptions mentioned at the start of Section 3. To translate the results to the Abstracted RL setting, we

use the Abstract L1 inequality (Theorem 2) to show that the empirical abstract model is accurate with high probability.

As is typical with abstraction, there is a trade-off between the performance and the required number of steps. Without abstraction, Theorem 3 shows that the algorithm will attain an expected return of $\text{Opt}(\prod_M(\epsilon, T_\epsilon)) - 2\epsilon$ within a number of steps polynomial in $|S|, |A|, \frac{1}{\epsilon}\frac{1}{\delta}, T_\epsilon$. With the abstraction, the additional penalty of $3g(\eta_T, \eta_R)$ appears because we use an approximate model similarity abstraction. The advantage is that the number of steps within which this is achieved is polynomial in the size of the abstract space $|\bar{S}|$ rather than the (larger) original state space $|S|$.

We can consider the martingale property an approximate version of independence (Chow & Teicher, 2003). This martingale property allowed us to give a guarantee on the quality of the learned model (Theorem 2), and we used this result to give finite-sample guarantees for Abstracted RL (Theorem 4). These results show that the performance is not arbitrarily bad with approximate model similarity abstraction but can be close to optimal when η_T and η_R are small.

4 Related Work

The problem we resolved in this paper may seem intuitive, but as we will make clear here, it is a fundamental problem in rich literature. Many studies have considered the combination of abstraction with either planning or RL. Some studies avoid, or ignore, the issue of dependency by simply assuming that samples are independent (Paduraru et al., 2008; Ortner et al., 2014b; Jiang et al., 2015; Badings et al., 2022). Others avoid it by looking at convergence in the limit (Singh et al., 1995; Hutter, 2016; Majeed & Hutter, 2018) or by assuming access to an MDP model (Hallak et al., 2013; Maillard et al., 2013; Ortner et al., 2019).

A negative result has been provided in the Abstracted RL setting, showing that R-MAX (Brafman & Tennenholtz, 2002) no longer maintains its guarantees when paired with any state abstraction function (Abel et al., 2018). For this negative result, they give an example that uses approximate Q^* similarity abstractions (Abel et al., 2016). Our counterexample is more powerful: indicating problems with the analysis even for approximate model similarity abstractions. Nevertheless, our second result shows that it is still possible to give guarantees in Abstracted RL for R-MAX-like algorithms when we use an approximate model similarity abstraction and take the η_R and η_T inaccuracies into account.

Another study considered a setting related to abstraction, where the transition and reward functions can change over time, either abruptly or gradually (Ortner et al., 2020). The reward and transition probabilities depend on the timestep t, so $T(s'|s, a, t)$ instead of $T(s'|s, a)$. They bound the variation in the reward and transition functions over time. By taking the variation over time into account they are able to give results. In their setting, the MDP is fixed given the timestep. However, in Abstracted RL this is not fixed. Each time the transition function at a timestep t could be different.

Recently regret guarantees have been found for the episodic (continuous) RL setting (Sinclair et al., 2022). Their abstraction method learns an abstraction adaptively. The model-based version of their algorithm requires that the state and actions spaces are embedded in compact metric spaces. In this way, they can define a measure of the difference between state-action pairs in these metric spaces. However, they require oracle access to this metric. In our setting this would require knowing the transition and reward functions, which we assume we do not.

There are quite a few studies in the area of abstraction selection, where the agent has access to a set of abstraction functions (state representations)(Hallak et al., 2013; Maillard et al., 2013; Lattimore et al., 2013; Ortner et al., 2014a; 2019). Several studies assume that the given set of state representations contains a Markov model (Hallak et al., 2013; Maillard et al., 2013; Ortner et al., 2019). One study gives asymptotic guarantees for selecting the correct model and building an exact MDP model (Hallak et al., 2013). The assumption that an MDP model exists in the given set of representations is crucial in their analysis since the samples are i.i.d. for this MDP model. Similarly, other studies also assume that the given set of state representations contains a Markov model (Maillard et al., 2013; Ortner et al., 2019). They create an algorithm for which they obtain regret bounds, and their analysis also uses the Markov representation.

Some studies in abstraction selection do not assume the given abstraction functions contain a Markov model (Lattimore et al., 2013; Ortner et al., 2014a). One study assumes access to a set of environments, including the true environment, rather than a set of abstractions (Lattimore et al., 2013). Because they have access to environments, they do not need to learn a transition model, making it different from our setting. The other study uses Theorem 2.1 from Weissman et al. (2003), which requires i.i.d. samples (Ortner et al., 2014a). We have shown that independent samples *cannot* be guaranteed in Abstracted RL.

Other related work is in MDPs with rich observations or block structure (Azizzadenesheli et al., 2016; Du et al., 2019; Allen et al., 2021). In that setting, each observation can be generated only from a *single* hidden state, which means that the issue of non-i.i.d. data due to abstraction does not arise. We can view the rich observation setting as an aggregation problem, where the observations can be aggregated to form a small (latent) MDP (Azizzadenesheli et al., 2016). Their setting is related to exact model similarity (or bisimulation) (Du et al., 2019). In contrast, in Abstracted RL, each observation can be generated from multiple hidden states, and we do not try to learn the MDP, as it is not small. Furthermore, we focus on approximate model similarity, which introduces the problems as described in Section 3.1.

One way to avoid the issue of dependent samples is by assuming that samples are obtained independently (Paduraru et al., 2008; Ortner et al., 2014b; Jiang et al., 2015; Badings et al., 2022). One study considers the setting with a continuous domain where we are given a data set with i.i.d. samples (Paduraru et al., 2008). They use discretization to aggregate states into abstract states and give a guarantee that, with a high probability, the model will be ϵ-accurate given a fixed data set. While they assume that the data has been gathered i.i.d., our results show that martingale concentration inequalities could be used to extend their results to the online data collection in the Abstracted RL setting. Discretization has been used in another study in a continuous space (Badings et al., 2022). They search for a solution for a linear dynamical system, where the transitions are deterministic, except for a noise component. They assume this noise is distributed i.i.d. and try to learn the resulting abstract transition functions by iteratively sampling N samples per abstract-state action pair until a threshold is reached. They assume that samples can be obtained cheaply, e.g., through a simulator, whereas we focus on the exploration problem where data has to be collected online and is expensive. Another study operates in the abstraction selection setting (Jiang et al., 2015). While they do not assume that a Markov model exists in the given set of abstraction functions, they assume a given data set, with i.i.d. data. They give a bound on how accurate the Q-values based on the (implicitly) learned model will be rather than on the accuracy of the model itself. As we showed, the assumption that the data is i.i.d. is not trivial since it means the data cannot just be collected online. Another study's primary focus is on bandits but also gives results for MDPs with a coloring function (Ortner et al., 2014b). We can view state aggregation as a special case of this coloring. They extend the results from UCRL2 (Jaksch et al., 2010) to the setting with a coloring function. They use the Azuma-Hoeffding inequality for the transition function, which holds for weakly dependent samples. However, they assume the samples are independent and do not show the martingale difference sequence property for the (actually dependent) samples.

Another way to deal with dependence between samples is by looking at convergence in the limit (Singh et al., 1995; Hutter, 2016; Majeed & Hutter, 2018). One study gives an asymptotic result for convergence of Q-learning and TD(0) in MDPs with soft state aggregation (Singh et al., 1995). In soft state aggregation, a state s belongs to a cluster x with some probability $P(x|s)$, which means a state s can belong to several clusters. Their result requires an ergodic MDP and a stationary policy that assigns a non-zero probability to every action. Together these imply a limiting state distribution, and they use this to show convergence asymptotically. Another study gives multiple results focusing on approximate and exact abstractions in environments without MDP assumptions (Hutter, 2016). Several of these results are in the planning setting, similar to other planning results for approximate abstractions (Abel et al., 2016). Most relevant for us is their Theorem 12, which for online RL shows convergence in the limit of the empirical transition function under weak conditions, e.g., when the abstract process is an MDP. Under this condition, however, the problem reduces to RL in an (abstract) MDP rather than Abstracted RL. Follow-up work builds on some of these results and focuses on the combination of model-free RL and exact abstraction (Majeed & Hutter, 2018). They show that, under state uniformity, Q-learning converges in the limit to the optimal solution. State uniformity means that aggregated histories have the same optimal Q-values. In contrast to our setting, they used an exact abstraction and left extending the results to approximate abstraction as an open question.

For planning in abstract MDPs, there are results for exact state abstractions (Li et al., 2006) and approximate state abstractions (Abel et al., 2016). The results for approximate state abstractions allow for quantifying an upper bound on performance for the optimal policy of an abstract MDP, e.g., as in Theorem 1 for approximate model similarity in Section 2.2. A study built on these results by giving a result for performing RL interacting with an explicitly constructed abstract MDP (Taïga et al., 2018); since the abstract MDP is still an MDP, this is different from Abstracted RL.

In certain cases, dependence can also arise for RL in MDPs. For instance, a study shows that dependence can occur in a non-communicating MDP (Strehl & Littman, 2008). They show that this specific case of dependence is not a problem because it is still possible to use a concentration inequality for independent samples, e.g., Hoeffding's inequality, as an upper bound. However, their proof uses that transitions and rewards are identically distributed, which is not guaranteed in Abstracted RL.

5 Discussion and Future Work

Some assumptions we made, i.e., that the reward function is deterministic and each state in an abstract state has the same reward function, can be relaxed. To accurately learn an abstract reward function, one should define a suitable martingale difference sequence, after which Lemma 3 can be used. We considered approximate model similarity abstraction and used the properties of this abstraction to establish an upper bound on the difference in value between the original MDP and an abstract MDP under any abstract policy. This bound was imperative for our results. We established this bound by proof of induction on the difference in value for a horizon n. This technique could be used to establish similar bounds and extend our results for other abstractions, e.g., approximate Q^* similarity abstractions (Abel et al., 2016).

Our analysis showed how to extend the results of R-MAX (Brafman & Tennenholtz, 2002) to Abstracted RL. Extending results of other algorithms, e.g., MBIE (Strehl & Littman, 2008) and UCRL2 (Jaksch et al., 2010), requires adapting to slightly different assumptions. For instance, R-MAX assumes ergodicity, while UCRL2 and MBIE assume the problem is communicating and non-communicating, respectively. Other algorithms sometimes use concentration inequalities other than Hoeffding's Inequality, e.g., the empirical Bernstein inequality (Audibert et al., 2007; Maurer & Pontil, 2009) or the Chernoff bound. To adapt these, we could, for instance, use Bernstein-type inequalities for martingales (Dzhaparidze & Van Zanten, 2001).

Theorem 4 shows that, despite problems with dependence, we can give finite-sample guarantees when combining approximate model similarity abstractions with model-based RL. For good abstraction functions, i.e., when η_R and η_T are small and $|S| \gg |\bar{S}|$, this leads to near-optimal solutions while needing fewer samples, compared to learning without abstraction. In deep model-based RL, several recent empirical works have shown promising results by focusing on learning exact abstractions (van der Pol et al., 2020; Biza & Platt, 2019). An interesting direction is adapting these methods to learn approximate abstractions instead of exact abstractions. Since, compared to exact model similarity abstractions, approximate model similarity abstraction generally results in a smaller (abstract) state space; this could lead to faster learning.

6 Conclusion

We analyzed Abstracted RL: online MBRL combined with state abstraction when the model of the MDP is unavailable. Via a counterexample, we showed that it cannot be guaranteed that samples obtained online in Abstracted RL are independent. Many current guarantees from MBRL methods use concentration results that assume i.i.d. samples, e.g., Theorem 2.1 from Weissman et al. (2003), the empirical Bernstein inequality (Audibert et al., 2007; Maurer & Pontil, 2009), or the Chernoff bound. Because they use these concentration inequalities, their guarantees do not hold in Abstracted RL. In fact, none of the existing analyses of MBRL apply to Abstracted RL. We showed that samples in Abstracted RL are only weakly dependent and that concentration inequalities for (weakly) dependent variables, such as Lemma 3, are a viable alternative through which we can come to guarantees on the empirical model. We used this result to present the first sample efficient learning results for Abstracted RL, thus showing it is possible to combine the benefits of abstraction and MBRL.

References

David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In *International Conference on Machine Learning*, 2016.

David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. State abstractions for lifelong reinforcement learning. In *International Conference on Machine Learning*, 2018.

Cameron Allen, Neev Parikh, Omer Gottesman, and George Konidaris. Learning markov state abstractions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 2021.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *International conference on algorithmic learning theory*, 2007.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning in rich-observation mdps using spectral methods. *arXiv preprint arXiv:1611.03907*, 2016.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 1967.

Thom S Badings, Alessandro Abate, Nils Jansen, David Parker, Hasan A Poonawala, and Marielle Stoelinga. Sampling-based robust control of autonomous systems with non-gaussian noise. *AAAI*, 2022.

Aijun Bai, Siddharth Srivastava, and Stuart J Russell. Markovian state and action abstractions for mdps via hierarchical mcts. In *IJCAI*, 2016.

Ondrej Biza and Robert Platt. Online abstraction with mdp homomorphisms for deep learning. In *AAMAS*, 2019.

George Boole. *An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities*. Dover Publications, 1854.

Hippolyte Bourel, Odalric Maillard, and Mohammad Sadegh Talebi. Tightening exploration in upper confidence reinforcement learning. In *International Conference on Machine Learning*, 2020.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 2002.

Yuan Shih Chow and Henry Teicher. *Probability theory: independence, interchangeability, martingales*. Springer Science & Business Media, 2003.

Thomas Dean, Robert Givan, and Sonia Leach. Model reduction techniques for computing approximately optimal solutions for markov decision processes. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, 1997.

Richard Dearden and Craig Boutilier. Abstraction and approximate decision-theoretic planning. *Artificial Intelligence*, 1997.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, 2019.

Kacha Dzhaparidze and JH Van Zanten. On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 2001.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, 2018.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 2003.

Assaf Hallak, Dotan Di-Castro, and Shie Mannor. Model selection in markovian processes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 1963.

Marcus Hutter. Extreme state aggregation beyond markov decision processes. *Theoretical Computer Science*, 2016.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 2010.

Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *International Conference on Machine Learning*, 2015.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 1998.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 2002.

Tor Lattimore, Marcus Hutter, Peter Sunehag, et al. The sample-complexity of general reinforcement learning. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

David A Levin and Yuval Peres. *Markov chains and mixing times*. American Mathematical Soc., 2017.

Lihong Li. *A unifying framework for computational reinforcement learning theory*. PhD thesis, Rutgers University-Graduate School-New Brunswick, 2009.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. In *ISAIM*, 2006.

Sridhar Mahadevan. Representation discovery in sequential decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2010.

Odalric-Ambrym Maillard, Phuong Nguyen, Ronald Ortner, and Daniil Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. In *International Conference on Machine Learning*, 2013.

Sultan Javed Majeed and Marcus Hutter. On Q-learning convergence for non-markov decision processes. In *IJCAI*, 2018.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Ronald Ortner, Odalric-Ambrym Maillard, and Daniil Ryabko. Selecting near-optimal approximate state representations in reinforcement learning. In *International Conference on Algorithmic Learning Theory*, 2014a.

Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless markov bandits. *Theoretical Computer Science*, 2014b.

Ronald Ortner, Matteo Pirotta, Alessandro Lazaric, Ronan Fruit, and Odalric-Ambrym Maillard. Regret bounds for learning state representations in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2019.

Ronald Ortner, Pratik Gajane, and Peter Auer. Variational regret bounds for reinforcement learning. In *Uncertainty in Artificial Intelligence*, 2020.

Cosmin Paduraru, Robert Kaplow, Doina Precup, and Joelle Pineau. Model-based reinforcement learning with state aggregation. In *8th European Workshop on Reinforcement Learning*, 2008.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

Sean R Sinclair, Siddhartha Banerjee, and Christina Lee Yu. Adaptive discretization in online reinforcement learning. *Operations Research*, 2022.

Satinder P Singh, Tommi Jaakkola, and Michael I Jordan. Learning without state-estimation in partially observable markovian decision processes. In *Machine Learning Proceedings*, 1994.

Satinder P Singh, Tommi Jaakkola, and Michael I Jordan. Reinforcement learning with soft state aggregation. In *Advances in neural information processing systems*, 1995.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 2008.

Adrien Ali Taïga, Aaron Courville, and Marc G Bellemare. Approximate exploration through state abstraction. *arXiv preprint arXiv:1808.09819*, 2018.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, 2018.

Elise van der Pol, Thomas Kipf, Frans A Oliehoek, and Max Welling. Plannable approximations to mdp homomorphisms: Equivariance under actions. In *AAMAS*, 2020.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. *Advances in Neural Information Processing Systems*, 2019.

A Well Known Results

We restate some well-known results that we use in the proofs in the other sections.

A.1 Hoeffding's Inequality

Hoeffding's inequality can tell us the probability that the average of m random independent (but not necessarily identically distributed) samples deviates more than ϵ from its expectation.

Let $Z^{(1)}, Z^{(2)}, \cdots, Z^{(m)}$ be bounded independent random variables, and let $\bar{Z}$ and μ be defined as

$$\bar{Z} \triangleq \frac{Z^{(1)} + \cdots + Z^{(m)}}{m}, \tag{33}$$

$$\mu \triangleq E[\bar{Z}] = \frac{E[Z^{(1)} + \cdots + Z^{(m)}]}{m}. \tag{34}$$

Then Hoeffding's inequality states:

Lemma 4 (Hoeffding's inequality (Hoeffding, 1963))**.** *If $Z^{(1)}, Z^{(2)}, \cdots, Z^{(m)}$ are independent and $0 \leq Z^{(i)} \leq 1$ for $i = 1, \cdots, m$, then for $0 < \epsilon < 1 - \mu$ we have the following inequalities*

$$\Pr(\bar{Z} - \mu \geq \epsilon) \leq e^{-2m\epsilon^2}, \tag{35}$$

$$\Pr(|\bar{Z} - \mu| \geq \epsilon) \leq 2e^{-2m\epsilon^2}, \tag{36}$$

$$\Pr(\sum_{i=1}^{m}(Z^{(i)} - \mu) \geq \epsilon) \leq e^{-2\frac{\epsilon^2}{m}}, \tag{37}$$

$$\Pr(|\sum_{i=1}^{m}(Z^{(i)} - \mu)| \geq \epsilon) \leq 2e^{-2\frac{\epsilon^2}{m}}. \tag{38}$$

A.2 Union Bound

Given that we have a set of events, the union bound allows us to upper bound the probability that at least one of the events happens, even when these events are not independent.

Lemma 5 (Union Bound (Boole, 1854))**.** *For a countable set of events $A_1, A_2, A_3, \cdots$, we have*

$$\Pr(\cup_i A_i) \leq \sum_i \Pr(A_i). \tag{39}$$

I.e., the probability that at least one of the events happens is, at most, the sum of the probabilities of the individual events.

A.3 Value Bounds for Abstract and True Models

Here we give upper bounds on the difference in value between the real MDP and an abstract MDP under various policies. We will use these bounds in Appendix D to adapt the results of R-MAX (Brafman & Tennenholtz, 2002) to Abstracted RL. These bounds and proofs are very similar to existing bounds (Brafman & Tennenholtz, 2002; Strehl & Littman, 2008; Abel et al., 2016; Taïga et al., 2018). Here we repeat these for abstract models in the undiscounted finite horizon and in the discounted infinite horizon.

We define the finite horizon value function $\forall s \in S$:

$$V^{\pi,n}(s) = R(s, \pi(s)) + \sum_{s' \in S} T(s'|s, \pi(s)) V^{\pi,n-1}(s), \tag{40}$$

$$V^{\pi,1}(s) = R(s, \pi(s)). \tag{41}$$

We use $V^{\bar{\pi},n}$ to denote the value in M under policy $\bar{\pi}$ and $\bar{V}^{\bar{\pi},n}$ to denote the value in $\bar{M}$ under policy $\bar{\pi}$.

Theorem 1. *Let $M = \langle S, A, T, R\rangle$ be an MDP and $\bar{M} = \langle \bar{S}, A, \bar{T}, \bar{R}\rangle$ an abstract MDP, for some defined abstract transitions and rewards. We assume that*

$$\forall \bar{s}, \bar{s}' \in \bar{S},\ s \in \bar{s},\ a \in A : |\bar{T}(\bar{s}'|\bar{s}, a) - \Pr(\bar{s}'|s, a)| \leq \eta_T \quad \text{and } |\bar{R}(\bar{s}, a) - R(s, a)| \leq \eta_R. \tag{42}$$

Then, for a finite horizon problem with horizon h we have:

$$V^*(s) - V^{\bar{\pi}^*}(s) \leq 2h\eta_R + (h+1)h\eta_T|\bar{S}|R_{max}. \tag{43}$$

And for a discounted infinite horizon problem with discount γ we have:

$$V^*(s) - V^{\bar{\pi}^*}(s) \leq \frac{2\eta_R}{1-\gamma} + \frac{2\gamma\eta_T|\bar{S}|R_{max}}{(1-\gamma)^2}. \tag{44}$$

We will use the following two Lemmas to proof the Theorem.

Lemma 6. *Under the assumption of equation 42 and for every abstract policy $\bar{\pi}$ and for every state $s \in \bar{s}$, we have: for a finite horizon problem with horizon h:*

$$|V^{\bar{\pi},h}(s) - \bar{V}^{\bar{\pi},h}(s)| \leq h\eta_R + \frac{(h-1)h}{2}\eta_T|\bar{S}|R_{max}, \tag{45}$$

and for a discounted infinite horizon problem with discount γ:

$$|V^{\bar{\pi}}(s) - \bar{V}^{\bar{\pi}}(s)| \leq \frac{\eta_R}{1-\gamma} + \frac{\gamma\eta_T|\bar{S}|R_{max}}{(1-\gamma)^2}. \tag{46}$$

Proof. The proof follows the same steps for both the discounted infinite horizon and the undiscounted finite horizon. For completeness, we show them both here.

First, for the undiscounted finite horizon. By induction, we will show that for $n \geq 1$

$$\forall \bar{s} \in \bar{S}, s \in \bar{s} : |V^{\bar{\pi},n}(s) - \bar{V}^{\bar{\pi},n}(s)| \leq n\eta_R + \frac{(n-1)n}{2}\eta_T|\bar{S}|R_{\max}. \tag{47}$$

For $n = 1$, we have

$$|V^{\bar{\pi},1}(s) - \bar{V}^{\bar{\pi},1}(s)| = |R(s, \pi(\bar{s})) - \bar{R}(\bar{s}, \bar{\pi}(\bar{s}))| \leq \eta_R. \tag{48}$$

Now assume that the induction hypothesis, equation 47, holds for $n-1$, then

$$|V^{\bar{\pi},n}(s) - \bar{V}^{\bar{\pi},n}(s)| = |R(s, \bar{\pi}(\bar{s})) - \bar{R}(\bar{s}, \bar{\pi}(\bar{s})) + \sum_{s' \in S} T(s'|s, \bar{\pi}(s))V^{\bar{\pi},n-1}(s') - \sum_{\bar{s}' \in \bar{S}} \bar{T}(\bar{s}'|\bar{s}, \bar{\pi}(\bar{s}))\bar{V}^{\bar{\pi},n-1}(\bar{s}')| \tag{49}$$

$$\leq |R(s, \bar{\pi}(\bar{s})) - \bar{R}(\bar{s}, \bar{\pi}(\bar{s}))| + |\sum_{\bar{s}' \in \bar{S}} \sum_{s' \in \bar{s}'} T(s'|s, \bar{\pi}(s))V^{\bar{\pi},n-1}(s') - \sum_{\bar{s}' \in \bar{S}} \bar{T}(\bar{s}'|\bar{s}, \bar{\pi}(\bar{s}))\bar{V}^{\bar{\pi},n-1}(\bar{s}')| \tag{50}$$

$$\leq \eta_R + |\sum_{\bar{s}' \in \bar{S}} \sum_{s' \in \bar{s}'} T(s'|s, \bar{\pi}(s))V^{\bar{\pi},n-1}(s') - \sum_{\bar{s}' \in \bar{S}} \bar{V}^{\bar{\pi},n-1}(\bar{s}') \sum_{s' \in \bar{s}'} T(s'|s, \bar{\pi}(\bar{s}))|$$
$$+ |\sum_{\bar{s}' \in \bar{S}} \bar{V}^{\bar{\pi},n-1}(\bar{s}') \sum_{s' \in \bar{s}'} T(s'|s, \bar{\pi}(\bar{s})) - \sum_{\bar{s}' \in \bar{S}} \bar{T}(\bar{s}'|\bar{s}, \bar{\pi}(\bar{s}))\bar{V}^{\bar{\pi},n-1}(\bar{s}')| \tag{51}$$

$$\leq \eta_R + |\sum_{\bar{s}' \in \bar{S}} \sum_{s' \in \bar{s}'} T(s'|s, \bar{\pi}(s))\big[V^{\bar{\pi},n-1}(s') - \bar{V}^{\bar{\pi},n-1}(\bar{s}')\big]| + |\sum_{\bar{s}' \in \bar{S}} \big[\bar{T}(\bar{s}'|\bar{s}, \bar{\pi}(\bar{s})) - \sum_{s' \in \bar{s}'} T(s'|s, \bar{\pi}(\bar{s}))\big]\bar{V}^{\bar{\pi},n-1}(\bar{s}')| \tag{52}$$

$$\leq \eta_R + (n-1)\eta_R + \frac{(n-1-1)(n-1)}{2}\eta_T|\bar{S}|R_{\max} + \eta_T|\bar{S}|(n-1)R_{\max} \tag{53}$$

$$= n\eta_R + \frac{(n-2)(n-1)}{2}\eta_T|\bar{S}|R_{\max} + \eta_T|\bar{S}|(n-1)R_{\max} \tag{54}$$

$$= n\eta_R + \frac{(n-1)n}{2}\eta_T|\bar{S}|R_{\max}. \tag{55}$$

For the step from equation 50 to equation 51, we subtract and add $\sum_{\bar{s}'\in\bar{S}} \bar{V}^{\bar{\pi},n-1}(\bar{s}') \sum_{s'\in\bar{s}'} T(s'|s,\bar{\pi}(\bar{s}))$, and from equation 52 to equation 53, we use the inductive hypothesis and the fact that $(n-1)R_{\max}$ is an upper bound on $\bar{V}^{\bar{\pi},n-1}(\bar{s}')$ since the maximum reward per timestep is $R_{\max}$.

Now, for the discounted infinite horizon. By induction, we will show that for $n \geq 1$

$$\forall \bar{s} \in \bar{S}, s \in \bar{s} : |V^{\bar{\pi},n}(s) - \bar{V}^{\bar{\pi},n}(s)| \leq \eta_R \gamma^{n-1} + \sum_{i=0}^{n-2} \gamma^i (\eta_R + \frac{\gamma \eta_T |\bar{S}| R_{\max}}{1-\gamma}). \tag{56}$$

For $n = 1$, we have

$$|V^{\bar{\pi},1}(s) - \bar{V}^{\bar{\pi},1}(s)| = |R(s,\pi(\bar{s})) - \bar{R}(\bar{s},\bar{\pi}(\bar{s}))| \leq \eta_R. \tag{57}$$

Now assume that the induction hypothesis, equation 56, holds for $n-1$, then

$$|V^{\bar{\pi},n}(s) - \bar{V}^{\bar{\pi},n}(s)| = |R(s,\bar{\pi}(\bar{s})) - \bar{R}(\bar{s},\bar{\pi}(\bar{s})) + \gamma(\sum_{s'\in S} T(s'|s,\bar{\pi}(s)) V^{\bar{\pi},n-1}(s') - \sum_{\bar{s}'\in\bar{S}} \bar{T}(\bar{s}'|\bar{s},\bar{\pi}(\bar{s})) \bar{V}^{\bar{\pi},n-1}(\bar{s}'))| \tag{58}$$

$$\leq |R(s,\bar{\pi}(\bar{s})) - \bar{R}(\bar{s},\bar{\pi}(\bar{s}))| + \gamma| \sum_{\bar{s}'\in\bar{S}} \sum_{s'\in\bar{s}'} T(s'|s,\bar{\pi}(s)) V^{\bar{\pi},n-1}(s') - \sum_{\bar{s}'\in\bar{S}} \bar{T}(\bar{s}'|\bar{s},\bar{\pi}(\bar{s})) \bar{V}^{\bar{\pi},n-1}(\bar{s}')| \tag{59}$$

$$\leq \eta_R + \gamma| \sum_{\bar{s}'\in\bar{S}} \sum_{s'\in\bar{s}'} T(s'|s,\bar{\pi}(s)) V^{\bar{\pi},n-1}(s') - \sum_{\bar{s}'\in\bar{S}} \bar{V}^{\bar{\pi},n-1}(\bar{s}') \sum_{s'\in\bar{s}'} T(s'|s,\bar{\pi}(\bar{s}))|$$
$$+ \gamma| \sum_{\bar{s}'\in\bar{S}} \bar{V}^{\bar{\pi},n-1}(\bar{s}') \sum_{s'\in\bar{s}'} T(s'|s,\bar{\pi}(\bar{s})) - \sum_{\bar{s}'\in\bar{S}} \bar{T}(\bar{s}'|\bar{s},\bar{\pi}(\bar{s})) \bar{V}^{\bar{\pi},n-1}(\bar{s}')| \tag{60}$$

$$\leq \eta_R + \gamma| \sum_{\bar{s}'\in\bar{S}} \sum_{s'\in\bar{s}'} T(s'|s,\bar{\pi}(s)) \big[V^{\bar{\pi},n-1}(s') - \bar{V}^{\bar{\pi},n-1}(\bar{s}')\big]| + \gamma| \sum_{\bar{s}'\in\bar{S}} \big[\bar{T}(\bar{s}'|\bar{s},\bar{\pi}(\bar{s})) - \sum_{s'\in\bar{s}'} T(s'|s,\bar{\pi}(\bar{s}))\big] \bar{V}^{\bar{\pi},n-1}(\bar{s}')| \tag{61}$$

$$\leq \eta_R + \gamma(\eta_R \gamma^{n-2} + \sum_{i=0}^{n-3} \gamma^i (\eta_R + \frac{\gamma \eta_T |\bar{S}| R_{\max}}{1-\gamma})) + \gamma \eta_T |\bar{S}| \frac{R_{\max}}{1-\gamma} \tag{62}$$

$$= \eta_R + \eta_R \gamma^{n-1} + \sum_{i=1}^{n-2} \gamma^i (\eta_R + \frac{\gamma \eta_T |\bar{S}| R_{\max}}{1-\gamma}) + \gamma \eta_T |\bar{S}| \frac{R_{\max}}{1-\gamma} \tag{63}$$

$$= \eta_R \gamma^{n-1} + \sum_{i=0}^{n-2} \gamma^i (\eta_R + \frac{\gamma \eta_T |\bar{S}| R_{\max}}{1-\gamma}). \tag{64}$$

For the step from equation 59 to equation 60, we subtract and add $\sum_{\bar{s}'\in\bar{S}} \bar{V}^{\bar{\pi},n-1}(\bar{s}') \sum_{s'\in\bar{s}'} T(s'|s,\bar{\pi}(\bar{s}))$, and from equation 61 to equation 62, we use the inductive hypothesis and the fact that $\frac{R_{\max}}{1-\gamma}$ is an upper bound on $\bar{V}^{\bar{\pi},n-1}(\bar{s}')$.

Finally, taking the limit for $n \to \infty$, we get:

$$|V^{\bar{\pi}}(s) - \bar{V}^{\bar{\pi}}(s)| \leq \eta_R \times 0 + \frac{1}{1-\gamma}(\eta_R + \frac{\gamma \eta_T |\bar{S}| R_{\max}}{1-\gamma})$$
$$= \frac{\eta_R}{1-\gamma} + \frac{\gamma \eta_T |\bar{S}| R_{\max}}{(1-\gamma)^2}.$$

□

Lemma 7. *Under the assumption of equation 42 and for every state $s \in \bar{s}$, we have: for a finite horizon problem with horizon h:*

$$|V^{*,h}(s) - \bar{V}^{*,h}(s)| \leq h\eta_R + \frac{(h-1)h}{2} \eta_T |\bar{S}| R_{max}, \tag{65}$$

and for a discounted infinite horizon problem with discount γ:

$$|V^*(s) - \bar{V}^*(s)| \leq \frac{\eta_R}{1-\gamma} + \frac{\gamma \eta_T |\bar{S}| R_{max}}{(1-\gamma)^2}. \tag{66}$$

Proof. First, we define

$$\forall \bar{s} \in \bar{S}, s \in S : V^{*,n}(s) = \max_{a \in A} \big[R(s,a) + \gamma \sum_{s' \in S} T(s'|s,a) V^{*,n-1}(s')\big], \tag{67}$$

$$\bar{V}^{*,n}(\bar{s}) = \max_{a \in A} \big[\bar{R}(\bar{s},a) + \gamma \sum_{\bar{s}' \in \bar{S}} \bar{T}(\bar{s}'|s,a) \bar{V}^{*,n-1}(\bar{s}')\big]. \tag{68}$$

For the undiscounted case $\gamma = 1$, so we can drop γ from the notation.

The proof follows the same steps as the proof of Lemma 6. We start again with the undiscounted finite horizon.

By induction, we will show that for $n \geq 1$

$$\forall \bar{s} \in \bar{S}, s \in \bar{s} : |V^{*,n}(s) - \bar{V}^{*,n}(\bar{s})| \leq n\eta_R + \frac{(n-1)n}{2} \eta_T |\bar{S}| R_{\max}. \tag{69}$$

Making use of the fact that $|\max f - \max g| \leq \max |f - g|$, we have for $n = 1$

$$|V^{*,1}(s) - \bar{V}^{*,1}(s)| = |\max_{a \in A} R(s,a) - \max_{a \in A} \bar{R}(\bar{s},a)| \leq \max_{a \in A} |R(s,a) - \bar{R}(\bar{s},a)| \leq \eta_R. \tag{70}$$

Now assume that the induction hypothesis, equation 69, holds for $n-1$, then

$$|V^{*,n}(s) - \bar{V}^{*,n}(\bar{s})| = \max_{a \in A} |R(s,a) - \bar{R}(\bar{s},a) + \sum_{s' \in S} T(s'|s,a) V^{*,n-1}(s') - \sum_{\bar{s}' \in \bar{S}} \bar{T}(\bar{s}'|\bar{s},a) \bar{V}^{*,n-1}(\bar{s}')| \tag{71}$$

$$\leq \max_{a \in A} |R(s,a) - \bar{R}(\bar{s},a)| + \max_{a \in A} |\sum_{\bar{s}' \in \bar{S}} \sum_{s' \in \bar{s}'} T(s'|s,a) V^{*,n-1}(s') - \sum_{\bar{s}' \in \bar{S}} \bar{T}(\bar{s}'|\bar{s},a) \bar{V}^{*,n-1}(\bar{s}')| \tag{72}$$

$$\leq \eta_R + \max_{a \in A} |\sum_{\bar{s}' \in \bar{S}} \sum_{s' \in \bar{s}'} T(s'|s,a) V^{*,n-1}(s') - \sum_{\bar{s}' \in \bar{S}} \bar{V}^{*,n-1}(\bar{s}') \sum_{s' \in \bar{s}'} T(s'|s,a)|$$

$$+ \max_{a \in A} |\sum_{\bar{s}' \in \bar{S}} \bar{V}^{*,n-1}(\bar{s}') \sum_{s' \in \bar{s}'} T(s'|s,a) - \sum_{\bar{s}' \in \bar{S}} \bar{T}(\bar{s}'|\bar{s},a) \bar{V}^{*,n-1}(\bar{s}')| \tag{73}$$

$$\leq \eta_R + \max_{a \in A} |\sum_{\bar{s}' \in \bar{S}} \sum_{s' \in \bar{s}'} T(s'|s,a) \big[V^{*,n-1}(s') - \bar{V}^{*,n-1}(\bar{s}')\big]| + \max_{a \in A} |\sum_{\bar{s}' \in \bar{S}} \big[\bar{T}(\bar{s}'|\bar{s},a) - \sum_{s' \in \bar{s}'} T(s'|s,a)\big] \bar{V}^{*,n-1}(\bar{s}')| \tag{74}$$

$$\leq \eta_R + (n-1)\eta_R + \frac{(n-1-1)(n-1)}{2} \eta_T |\bar{S}| R_{\max} + \eta_T (n-1) |\bar{S}| R_{\max} \tag{75}$$

$$= n\eta_R + \frac{(n-1)n}{2} \eta_T |\bar{S}| R_{\max}. \tag{76}$$

For the step from equation 72 to equation 73, we subtract and add $\sum_{\bar{s}' \in \bar{S}} \bar{V}^{*,n-1}(\bar{s}') \sum_{s' \in \bar{s}'} T(s'|s,a)$, and from equation 74 to equation 75, we use the inductive hypothesis and again the fact that $(n-1)R_{\max}$ is an upper bound on $\bar{V}^{*,n-1}(\bar{s}')$ since the maximum reward per timestep is $R_{\max}$.

Now, for the discounted infinite horizon. By induction, we will show that for $n \geq 1$

$$\forall \bar{s} \in \bar{S}, s \in \bar{s} : |V^{\bar{\pi},n}(s) - \bar{V}^{\bar{\pi},n}(s)| \leq \eta_R \gamma^{n-1} + \sum_{i=0}^{n-2} \gamma^i (\eta_R + \frac{\gamma \eta_T |\bar{S}| R_{\max}}{1-\gamma}). \tag{77}$$

For $n = 1$, we have

$$|V^{\bar{\pi},1}(s) - \bar{V}^{\bar{\pi},1}(s)| = |\max_{a \in A} R(s,a) - \max_{a \in A} \bar{R}(\bar{s},a)| \leq \max_{a \in A} |R(s,a) - \bar{R}(\bar{s},a)| \leq \eta_R. \tag{78}$$

Now assume that the induction hypothesis, equation 77, holds for $n-1$, then

$$|V^{*,n}(s) - \bar{V}^{*,n}(\bar{s})| = \max_{a\in A} |R(s,a) - \bar{R}(\bar{s},a) + \gamma \sum_{s'\in S} T(s'|s,a)V^{*,n-1}(s') - \gamma \sum_{\bar{s}'\in\bar{S}} \bar{T}(\bar{s}'|\bar{s},a)\bar{V}^{*,n-1}(\bar{s}')| \tag{79}$$

$$\leq \max_{a\in A} |R(s,a) - \bar{R}(\bar{s},a)| + \max_{a\in A} \gamma| \sum_{\bar{s}'\in\bar{S}} \sum_{s'\in\bar{s}'} T(s'|s,a)V^{*,n-1}(s') - \sum_{\bar{s}'\in\bar{S}} \bar{T}(\bar{s}'|\bar{s},a)\bar{V}^{*,n-1}(\bar{s}')| \tag{80}$$

$$\leq \eta_R + \max_{a\in A} \gamma| \sum_{\bar{s}'\in\bar{S}} \sum_{s'\in\bar{s}'} T(s'|s,a)V^{*,n-1}(s') - \sum_{\bar{s}'\in\bar{S}} \bar{V}^{*,n-1}(\bar{s}') \sum_{s'\in\bar{s}'} T(s'|s,a)|$$

$$+ \max_{a\in A} \gamma| \sum_{\bar{s}'\in\bar{S}} \bar{V}^{*,n-1}(\bar{s}') \sum_{s'\in\bar{s}'} T(s'|s,a) - \sum_{\bar{s}'\in\bar{S}} \bar{T}(\bar{s}'|\bar{s},a)\bar{V}^{*,n-1}(\bar{s}')| \tag{81}$$

$$\leq \eta_R + \max_{a\in A} \gamma| \sum_{\bar{s}'\in\bar{S}} \sum_{s'\in\bar{s}'} T(s'|s,a)\big[V^{*,n-1}(s') - \bar{V}^{*,n-1}(\bar{s}')\big]| + \max_{a\in A} \gamma| \sum_{\bar{s}'\in\bar{S}} \big[\bar{T}(\bar{s}'|\bar{s},a) - \sum_{s'\in\bar{s}'} T(s'|s,a)\big]\bar{V}^{*,n-1}(\bar{s}')| \tag{82}$$

$$\leq \eta_R + \gamma(\eta_R\gamma^{n-2} + \sum_{i=0}^{n-3} \gamma^i(\eta_R + \frac{\gamma\eta_T|\bar{S}|R_{\max}}{1-\gamma})) + \gamma\eta_T|\bar{S}|\frac{R_{\max}}{1-\gamma} \tag{83}$$

$$= \eta_R + \eta_R\gamma^{n-1} + \sum_{i=1}^{n-2} \gamma^i(\eta_R + \frac{\gamma\eta_T|\bar{S}|R_{\max}}{1-\gamma}) + \gamma\eta_T|\bar{S}|\frac{R_{\max}}{1-\gamma} \tag{84}$$

$$= \eta_R\gamma^{n-1} + \sum_{i=0}^{n-2} \gamma^i(\eta_R + \frac{\gamma\eta_T|\bar{S}|R_{\max}}{1-\gamma}). \tag{85}$$

For the step from equation 80 to equation 81, we subtract and add $\sum_{\bar{s}'\in\bar{S}} \bar{V}^{\bar{\pi},n-1}(\bar{s}') \sum_{s'\in\bar{s}'} T(s'|s,\bar{\pi}(\bar{s}))$, and from equation 82 to equation 83, we use the inductive hypothesis and the fact that $\frac{R_{\max}}{1-\gamma}$ is an upper bound on $\bar{V}^{*,n-1}(\bar{s}')$.

Finally, taking the limit for $n \to \infty$, we get:

$$\begin{aligned}|V^*(s) - \bar{V}^*(s)| &\leq \eta_R \times 0 + \frac{1}{1-\gamma}(\eta_R + \frac{\gamma\eta_T|\bar{S}|R_{\max}}{1-\gamma}) \\ &= \frac{\eta_R}{1-\gamma} + \frac{\gamma\eta_T|\bar{S}|R_{\max}}{(1-\gamma)^2}.\end{aligned}$$

□

Proof of Theorem 1. We can now proof Theorem 1, by using the triangle inequality and the results of Lemmas 6 and 7. For the undiscounted finite horizon:

$$\begin{aligned}|V^{*,h}(s) - V^{\bar{\pi}^*,h}(s)| &\leq |V^{*,h}(s) - \bar{V}^{*,h}(s)| + |\bar{V}^{*,h}(s) - V^{\bar{\pi}^*,h}(s)| \\ &= |V^{*,h}(s) - \bar{V}^{*,h}(s)| + |\bar{V}^{\bar{\pi}^*,h}(s) - V^{\bar{\pi}^*,h}(s)| \\ &\leq h\eta_R + \frac{(h-1)h}{2}\eta_T|\bar{S}|R_{\max} + h\eta_R + \frac{(h-1)h}{2}\eta_T|\bar{S}|R_{\max} \\ &= 2h\eta_R + (h-1)h\eta_T|\bar{S}|R_{\max}.\end{aligned}$$

For the discounted infinite horizon:

$$\begin{aligned}|V^*(s) - V^{\bar{\pi}^*}(s)| &\leq |V^*(s) - \bar{V}^*(s)| + |\bar{V}^*(s) - V^{\bar{\pi}^*}(s)| \\ &= |V^*(s) - \bar{V}^*(s)| + |\bar{V}^{\bar{\pi}^*}(s) - V^{\bar{\pi}^*}(s)| \\ &\leq \frac{\eta_R}{1-\gamma} + \frac{\gamma\eta_T|\bar{S}|R_{\max}}{(1-\gamma)^2} + \frac{\eta_R}{1-\gamma} + \frac{\gamma\eta_T|\bar{S}|R_{\max}}{(1-\gamma)^2} \\ &= \frac{2\eta_R}{1-\gamma} + \frac{2\gamma\eta_T|\bar{S}|R_{\max}}{(1-\gamma)^2}.\end{aligned}$$

□

A.3.1 Value difference for similar MDPs

Finally, we give a simulation lemma for two MDP on the same state-action space.

Lemma 8. *Let M and M' be two MDP on the same state-action space, with*

$$\forall s, a \in S \times A: \ |R_M(s,a) - R_{M'}(s,a)| \leq \eta_R, \tag{86}$$

$$\forall s, a, s' \in S \times A \times S: \ |T_M(s'|s,a) - T_{M'}(s'|s,a)| \leq \eta_T. \tag{87}$$

Then, for every policy π and for every state $s \in S$ we have:

$$|V_M^{\pi,n}(s) - V_{M'}^{\pi,n}(s)| \leq n\eta_R + \frac{(n-1)n}{2}\eta_T|S|R_{max}. \tag{88}$$

Proof. By induction, we will show that for $n \geq 1$

$$\forall s \in S: \ |V_M^{\pi,n}(s) - V_{M'}^{\pi,n}(s)| \leq n\eta_R + \frac{(n-1)n}{2}\eta_T|S|R_{\max}. \tag{89}$$

For $n = 1$, we have

$$|V_M^{\pi,1}(s) - V_{M'}^{\pi,1}(s)| = |R_M(s,\pi(s)) - R_{M'}(s,\pi(s))| \leq \eta_R. \tag{90}$$

Now assume that the induction hypothesis, equation 89, holds for $n-1$, then

$$|V_M^{\pi,n}(s) - V_{M'}^{\pi,n}(s)| = |R_M(s,\pi(\bar{s})) - R_{M'}(s,\pi(s)) + \sum_{s'\in S} T_M(s'|s,\pi(s))V_M^{\pi,n-1}(s') - \sum_{s'\in S} T_{M'}(s'|s,\pi(s)))V_{M'}^{\pi,n-1}(s')| \tag{91}$$

$$\leq |R_M(s,\pi(s)) - R_{M'}(s,\pi(s))| + |\sum_{s'\in S} T_M(s'|s,\pi(s))V_M^{\pi,n-1}(s') - \sum_{s'\in S} T_{M'}(s'|s,\pi(s))V_{M'}^{\pi,n-1}(s')| \tag{92}$$

$$\leq \eta_R + |\sum_{s'\in S} T_M(s'|s,\pi(s))V_M^{\pi,n-1}(s') - \sum_{s'\in S} T_M(s'|s,\pi(s))V_{M'}^{\pi,n-1}(s')|$$

$$+ |\sum_{s'\in S} T_M(s'|s,\pi(s))V_{M'}^{\pi,n-1}(s') - \sum_{s'\in S} T_{M'}(s'|s,\pi(s))V_{M'}^{\pi,n-1}(s')| \tag{93}$$

$$\leq \eta_R + |\sum_{s'\in S} T_M(s'|s,\pi(s))\big[V_M^{\pi,n-1}(s') - V_{M'}^{\pi,n-1}(s')\big]| + |\sum_{s'\in S} \big[T_M(s'|s,\pi(s)) - T_{M'}(s'|s,\pi(s))\big]V_{M'}^{\pi,n-1}(s')| \tag{94}$$

$$\leq \eta_R + (n-1)\eta_R + \frac{(n-1-1)(n-1)}{2}\eta_T|S|R_{\max} + \eta_T(n-1)|S|R_{\max} \tag{95}$$

$$= n\eta_R + \frac{(n-2)(n-1)}{2}\eta_T|S|R_{\max} + \eta_T(n-1)|S|R_{\max} \tag{96}$$

$$= n\eta_R + \frac{(n-1)n}{2}\eta_T|S|R_{\max}. \tag{97}$$

For the step from equation 92 to equation 93, we add and subtract $\sum_{\bar{s}'\in S} T_M(s'|s,\pi(s))V_{M'}^{\pi,n-1}(s')$, and from equation 94 to equation 95, we use the inductive hypothesis and the fact that $(n-1)R_{\max}$ is an upper bound on $V_{M'}^{\pi,n-1}(s')$ since the maximum reward per timestep is $R_{\max}$. □

This shows that the values under any policy are similar for similar MDPs.

B L1 Inequality for Independent but not Identically Distributed Variables

We show that we can adapt the proof of Weissman et al. (2003) for independent, but not identically distributed, samples to obtain the following result:

Lemma 2. *Let $X_{\bar{s},a} = s_1, \cdots, s_m$ be a sequence of states $s \in \bar{s}$ and let $\bar{\boldsymbol{Y}}_{\bar{s},a} = \bar{Y}^{(1)}, \bar{Y}^{(2)}, \cdots, \bar{Y}^{(m)}$ be independent random variables distributed according to $\Pr(\cdot|s_1,a), \cdots, \Pr(\cdot|s_m,a)$. Then, $\forall \epsilon > 0$,*

$$\Pr(||\bar{T}_Y(\cdot|\bar{s},a) - \bar{T}_{\omega_X}(\cdot|\bar{s},a)||_1 \geq \epsilon) \leq (2^{|\bar{S}|} - 2)e^{-\frac{1}{2}m\epsilon^2}. \tag{98}$$

Proof of Lemma 2. The proof mostly follows the steps by Weissman et al. (2003).

To shorten the notation, we define $P_Y \triangleq \bar{T}_Y(\cdot|\bar{s}, a)$ and $P_{\omega_X} \triangleq \bar{T}_{\omega_X}(\cdot|\bar{s}, a)$.

We will make use of the following result (Proposition 4.2 in (Levin & Peres, 2017)), that for any distribution Q on $\bar{S}$

$$||Q - P_{\omega_X}||_1 = 2 \max_{\bar{\mathcal{S}} \subseteq \bar{S}} (Q(\bar{\mathcal{S}}) - P_{\omega_X}(\bar{\mathcal{S}})),$$

where $\bar{\mathcal{S}}$ is a subset of $\bar{S}$, and $P_{\omega_X}(\bar{\mathcal{S}}) = \sum_{\bar{s}' \in \bar{\mathcal{S}}} P_{\omega_X}(\bar{s}')$. Thus, we have that

$$||P_Y - P_{\omega_X}||_1 = 2 \max_{\bar{\mathcal{S}} \subseteq \bar{S}} (P_Y(\bar{\mathcal{S}}) - P_{\omega_X}(\bar{\mathcal{S}})). \tag{99}$$

Using this, we can write

$$\Pr(||P_Y - P_{\omega_X}||_1 \geq \epsilon) = \Pr\left[2 \max_{\bar{\mathcal{S}} \subseteq \bar{S}} \left[P_Y(\bar{\mathcal{S}}) - P_{\omega_X}(\bar{\mathcal{S}})\right] \geq \epsilon\right] \tag{100}$$

$$= \Pr\left[\max_{\bar{\mathcal{S}} \subseteq \bar{S}} \left[P_Y(\bar{\mathcal{S}}) - P_{\omega_X}(\bar{\mathcal{S}})\right] \geq \frac{\epsilon}{2}\right] \tag{101}$$

$$= \Pr\left[\cup_{\bar{\mathcal{S}} \subseteq \bar{S}} \left[P_Y(\bar{\mathcal{S}}) - P_{\omega_X}(\bar{\mathcal{S}}) \geq \frac{\epsilon}{2}\right]\right] \tag{102}$$

$$\leq \sum_{\bar{\mathcal{S}} \subseteq \bar{S}} \Pr\left[P_Y(\bar{\mathcal{S}}) - P_{\omega_X}(\bar{\mathcal{S}}) \geq \frac{\epsilon}{2}\right], \tag{103}$$

where the last step follows from the union bound (Lemma 5).

Assuming $\epsilon > 0$, we have that $\Pr(P_Y(\bar{\mathcal{S}}) - P_{\omega_X}(\bar{\mathcal{S}}) \geq \frac{\epsilon}{2}) = 0$ when $\bar{\mathcal{S}} = \bar{S}$ or $\bar{\mathcal{S}} = \emptyset$. For every other subset $\bar{\mathcal{S}}$, we can define a random binary variable that is 1 when $Y^{(i)} \in \bar{\mathcal{S}}$ and 0 otherwise. Here $P_{\omega_X}(\bar{\mathcal{S}})$ acts as μ (equation 34) from Lemma 4 and $P_Y(\bar{\mathcal{S}})$ as $\bar{Z}$ (equation 33). Thus, by applying Lemma 4 to this random variable, we have

$$\Pr(P_Y(\bar{\mathcal{S}}) - P_{\omega_X}(\bar{\mathcal{S}}) \geq \frac{\epsilon}{2}) \leq e^{-2m\frac{\epsilon}{2}^2} = e^{-\frac{1}{2}m\epsilon^2}. \tag{104}$$

Then it follows that

$$\Pr(||P_Y - P_{\omega_X}||_1 \geq \epsilon) \leq \sum_{\bar{\mathcal{S}} \subseteq \bar{S}} \Pr(P_Y(\bar{\mathcal{S}}) - P_{\omega_X}(\bar{\mathcal{S}}) \geq \frac{\epsilon}{2}) \tag{105}$$

$$\leq \sum_{\bar{\mathcal{S}} \subset \bar{S}: \bar{\mathcal{S}} \neq \bar{S}, \emptyset} \Pr(P_Y(\bar{\mathcal{S}}) - P_{\omega_X}(\bar{\mathcal{S}}) \geq \frac{\epsilon}{2}) \tag{106}$$

$$\leq (2^{|\bar{S}|} - 2) e^{-\frac{1}{2}m\epsilon^2}, \tag{107}$$

where $\bar{S} \subset \bar{\mathcal{S}} : \bar{S} \neq \bar{\mathcal{S}}, \emptyset$ denotes that the empty set $\emptyset$ and the complete set $\bar{\mathcal{S}}$ are excluded. □

C Proof of Main Result

Here we show how we can use a concentration inequality for martingales to learn an accurate transition model in Abstracted RL. Specifically, the following result shows that, with a high probability, the empirical abstract transition function $\bar{T}_Y$ will be close to the abstract transition function $\bar{T}_{\omega_X}$. In the proof, follows the general approach of Ortner et al. (2020), we define a suitable martingale difference sequence for the abstract transition function and use this to obtain the following result for learning a transition function in Abstracted RL:

Theorem 2 (Abstract L1 inequality)**.** *If an agent has access to a state abstraction function ϕ and uses this to collect data for any abstract state-action pair $(\bar{s}, a)$ by acting in an MDP M according to a policy $\bar{\pi}$, we have that the following holds with a probability of at least $1 - \delta$ for a fixed value of $N(\bar{s}, a)$:*

$$||\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_X}(\cdot|\bar{s}, a)||_1 \leq \epsilon, \tag{108}$$

where $\delta = 2^{|\bar{S}|} e^{-\frac{1}{8}N(\bar{s},a)\epsilon^2}$.

Proof of Theorem 2. We first define an abstract transition function based on $X_{\bar{s},a}$ as

$$\forall(\bar{s},a),\bar{s}' : \bar{T}_{\omega_X}(\bar{s}'|\bar{s},a) \triangleq \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} T(\bar{s}'|X^{(i)}_{\bar{s},a},a), \tag{109}$$

where $T(\bar{s}'|X^{(i)}_{\bar{s},a},a) \triangleq \sum_{s'\in\bar{s}'} T(s'|X^{(i)}_{\bar{s},a},a,a)$. We write $\bar{T}_{\omega_X}$ because this definition is equivalent to using a weighting function as in equation 15:

$$\begin{aligned}
\forall(\bar{s},a),\bar{s}' : \bar{T}_{\omega_X}(\bar{s}'|\bar{s},a) &\triangleq \sum_{s\in\bar{s}} \omega_X(s,a) \sum_{s'\in\bar{s}'} T(s'|s,a) && \text{(Eq. 15)} && (110)\\
&= \sum_{s\in\bar{s}} \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} \mathbb{1}\{X^{(i)}_{\bar{s},a}=s\} \sum_{s'\in\bar{s}'} T(s'|s,a) && && (111)\\
&= \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} \sum_{s\in\bar{s}} \mathbb{1}\{X^{(i)}_{\bar{s},a}=s\} \sum_{s'\in\bar{s}'} T(s'|s,a) && && (112)\\
&= \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} \sum_{s'\in\bar{s}'} T(s'|X^{(i)}_{\bar{s},a},a) && && (113)\\
&= \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} T(\bar{s}'|X^{(i)}_{\bar{s},a},a). && \text{(Eq. 109)} && (114)
\end{aligned}$$

Now we use $\boldsymbol{z}$ to denote a vector of size $|\bar{S}|$ with entries ±1, and we write $z(\bar{s})$ for the entry in $\boldsymbol{z}$ with index $\bar{s}$. Then we have

$$\begin{aligned}
||\bar{T}_Y(\cdot|\bar{s},a) - \bar{T}_{\omega_X}(\cdot|\bar{s},a)||_1 &= \sum_{\bar{s}'} |\bar{T}_Y(\bar{s}'|\bar{s},a) - \bar{T}_{\omega_X}(\bar{s}'|\bar{s},a)| && (115)\\
&= \max_{\boldsymbol{z}\in\{-1,1\}^{\bar{S}}} \sum_{\bar{s}'} \Big(\bar{T}_Y(\bar{s}'|\bar{s},a) - \bar{T}_{\omega_X}(\bar{s}'|\bar{s},a)\Big) z(\bar{s}') && (116)\\
&= \max_{\boldsymbol{z}\in\{-1,1\}^{\bar{S}}} \sum_{\bar{s}'} \Big(\frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} \mathbb{1}\{\bar{Y}^{(i)}_{\bar{s},a}=\bar{s}'\} - \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} T(\bar{s}'|X^{(i)}_{\bar{s},a},a)\Big) z(\bar{s}') && (117)\\
&= \max_{\boldsymbol{z}\in\{-1,1\}^{\bar{S}}} \sum_{\bar{s}'} \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} \mathbb{1}\{\bar{Y}^{(i)}_{\bar{s},a}=\bar{s}'\} z(\bar{s}') - \sum_{\bar{s}'} \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} T(\bar{s}'|X^{(i)}_{\bar{s},a},a) z(\bar{s}') && (118)\\
&= \max_{\boldsymbol{z}\in\{-1,1\}^{\bar{S}}} \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} z(\bar{Y}^{(i)}_{\bar{s},a}) - \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} \sum_{\bar{s}'} T(\bar{s}'|X^{(i)}_{\bar{s},a},a) z(\bar{s}') && (119)\\
&= \max_{\boldsymbol{z}\in\{-1,1\}^{\bar{S}}} \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} \Big(z(\bar{Y}^{(i)}_{\bar{s},a}) - \sum_{\bar{s}'} T(\bar{s}'|X^{(i)}_{\bar{s},a},a) z(\bar{s}')\Big) && (120)\\
&= \max_{\boldsymbol{z}\in\{-1,1\}^{\bar{S}}} \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a, \bar{Y}^{(i)}_{\bar{s},a}), && (121)
\end{aligned}$$

where we set

$$Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a}) \triangleq z(\bar{Y}^{(i)}_{\bar{s},a}) - \sum_{\bar{s}'} T(\bar{s}'|X^{(i)}_{\bar{s},a}, a_{\tau_i}) z(\bar{s}').$$

To show that $\sum_i^{N(\bar{s},a)} Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a})$ is a martingale difference sequence, we should follow Definition 3 and show that $\forall i : E[Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a})|Z_{\tau_1}, Z_{\tau_2}, \cdots, Z_{\tau_{i-1}}] = 0$ and $|Z_i| < \infty$. For the second part, we

have that $\forall i : |Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a})| \leq 2$, since $|z(\bar{Y}^{(i)}_{\bar{s},a})| \leq 1$ and $|\sum_{\bar{s}'} T(\bar{s}'|X^{(i)}_{\bar{s},a}, a_{\tau_i})z(\bar{s}')| \leq 1$. For the first part, we use the following Lemma, the proof of which follows after the current proof.

Lemma 9. *Let π be a policy, and suppose the sequence $s_1, a_1 \cdots, s_{t-1}, a_{t-1}, s_t$ is to be generated by π. If $1 \leq \tau_1 < \tau_2 < \cdots < \tau_{i-1} < \tau_i \leq t$, then $E[Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a})|Z_{\tau_1}, Z_{\tau_2}, \cdots, Z_{\tau_{i-1}}] = 0$.*

Lemma 9 shows that $\sum_i^{N(\bar{s},a)} Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a})$ is a martingale difference sequence with $\forall i : |Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a})| \leq 2$ for any fixed $\boldsymbol{z}$ and fixed $N(\bar{s}, a) = n$ so that by the Azuma-Hoeffding inequality (Lemma 3):

$$\Pr(\sum_{i=1}^{N(\bar{s},a)} Z_{\tau_i} > \epsilon) \leq e^{-\frac{\epsilon^2}{8N(\bar{s},a)}}. \tag{122}$$

Similarly, $\sum_i^{N(\bar{s},a)} \frac{1}{N(\bar{s},a)} Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a})$ is a martingale difference sequence with $\forall i : |\frac{1}{N(\bar{s},a)} Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a})| \leq \frac{2}{N(\bar{s},a)}$ for any fixed $\boldsymbol{z}$ and $N(\bar{s}, a) = n$ so that, by the Azuma-Hoeffding inequality (Lemma 3), the following holds:

$$\Pr(\frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} Z_{\tau_i} > \epsilon) \leq e^{-\frac{\epsilon^2}{2\frac{4}{N(\bar{s},a)^2} N(\bar{s},a)}} \tag{123}$$

$$= e^{-\frac{\epsilon^2}{\frac{8}{N(\bar{s},a)}}} \tag{124}$$

$$= e^{-\frac{1}{8}N(\bar{s},a)\epsilon^2}. \tag{125}$$

From equation 115 and equation 121, we then obtain

$$\Pr(||\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_X}(\cdot|\bar{s}, a)||_1 > \epsilon) = \Pr(\max_{\boldsymbol{z}\in\{-1,1\}^S} \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} Z_{\tau_i} > \epsilon). \tag{126}$$

A union bound (Lemma 5) over all $2^{|\bar{S}|}$ vectors $\boldsymbol{z}$ for a fixed value of $N(s, a)$ shows

$$\Pr(\max_{\boldsymbol{z}\in\{-1,1\}^S} \frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} Z_{\tau_i} > \epsilon) \leq \sum_{\boldsymbol{z}\in\{-1,1\}^S} \Pr(\frac{1}{N(\bar{s},a)} \sum_{i=1}^{N(\bar{s},a)} Z_{\tau_i} > \epsilon). \tag{127}$$

So, using equation 125, we have that the following holds with probability $1 - 2^{|\bar{S}|}e^{-\frac{1}{8}N(\bar{s},a)\epsilon^2}$:

$$||\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_X}(\cdot|\bar{s}, a)||_1 \leq \epsilon. \tag{128}$$

□

Now we give the proof of Lemma 9:

Proof of Lemma 9. We follow the general structure of the proof of Lemma 8 by Strehl & Littman (2008). We have

$$E[Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a})] = \sum_{c_{\tau_i+1}} \Pr(c_{\tau_i+1}) Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a}) \tag{129}$$

$$= \sum_{c_{\tau_i}} \Pr(c_{\tau_i}) \sum_{\bar{Y}^{(i)}_{\bar{s},a}} \Pr(\bar{Y}^{(i)}_{\bar{s},a}|c_{\tau_i}, a_{\tau_i}) Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a}) \tag{130}$$

$$= \sum_{c_{\tau_i}} \Pr(c_{\tau_i}) \sum_{\bar{Y}^{(i)}_{\bar{s},a}} \Pr(\bar{Y}^{(i)}_{\bar{s},a}|X^{(i)}_{\bar{s},a}, a_{\tau_i}) Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a}). \tag{131}$$

The sum $\sum_{c_{\tau_i+1}}$ is over all possible sequences c_{τ_i+1} that end in a state $\bar{s}_{\tau_i+1}$, resulting from τ_i actions chosen by an agent following policy π. Conditioning on the sequence of random variables $Z_{\tau_1}, Z_{\tau_2}, \cdots, Z_{\tau_{i-1}}$ can make some sequences c_{τ_i} more likely and others less likely, that is

$$E[Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a})|Z_{\tau_1}, Z_{\tau_2}, \cdots, Z_{\tau_{i-1}}] \tag{132}$$

$$= \sum_{c_{\tau_i}} \Pr(c_{\tau_i}|Z_{\tau_1}, Z_{\tau_2}, \cdots, Z_{\tau_{i-1}}) \sum_{\bar{Y}^{(i)}_{\bar{s},a}} \Pr(\bar{Y}^{(i)}_{\bar{s},a}|X^{(i)}_{\bar{s},a}, a_{\tau_i}) Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a}). \tag{133}$$

Significantly, since $P(\bar{Y}^{(i)}_{\bar{s},a}|\bar{s}_{\tau_i}, a_{\tau_i}, Z_{\tau_1}, \cdots, Z_{\tau_i-1}) = P(\bar{Y}^{(i)}_{\bar{s},a}|\bar{s}_{\tau_i}, a_{\tau_i})$, fixed values of $Z_{\tau_1}, Z_{\tau_2}, \cdots, Z_{\tau_{i-1}}$ do not influence the innermost sum of equation 133. For this innermost sum, we have

$$\sum_{\bar{Y}^{(i)}_{\bar{s},a}} \Pr(\bar{Y}^{(i)}_{\bar{s},a}|X^{(i)}_{\bar{s},a}, a_{\tau_i}) Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a}) \tag{134}$$

$$= \sum_{\bar{Y}^{(i)}_{\bar{s},a}} \Pr(\bar{Y}^{(i)}_{\bar{s},a}|X^{(i)}_{\bar{s},a}, a_{\tau_i}) \left[z(\bar{Y}^{(i)}_{\bar{s},a}) - \sum_{\bar{s}'} T(\bar{s}'|X^{(i)}_{\bar{s},a}, a_{\tau_i}) z(\bar{s}') \right] \tag{135}$$

$$= \sum_{\bar{Y}^{(i)}_{\bar{s},a}} \Pr(\bar{Y}^{(i)}_{\bar{s},a}|X^{(i)}_{\bar{s},a}, a_{\tau_i}) z(\bar{Y}^{(i)}_{\bar{s},a}) - \sum_{\bar{Y}^{(i)}_{\bar{s},a}} \Pr(\bar{Y}^{(i)}_{\bar{s},a}|X^{(i)}_{\bar{s},a}, a_{\tau_i}) \sum_{\bar{s}'} T(\bar{s}'|X^{(i)}_{\bar{s},a}, a_{\tau_i}) z(\bar{s}') \tag{136}$$

$$= \sum_{\bar{Y}^{(i)}_{\bar{s},a}} \Pr(\bar{Y}^{(i)}_{\bar{s},a}|X^{(i)}_{\bar{s},a}, a_{\tau_i}) z(\bar{Y}^{(i)}_{\bar{s},a}) - \sum_{\bar{s}'} T(\bar{s}'|X^{(i)}_{\bar{s},a}, a_{\tau_i}) z(\bar{s}') \tag{137}$$

$$= 0. \tag{138}$$

So we conclude

$$E[Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a})|Z_{\tau_1}, Z_{\tau_2}, \cdots, Z_{\tau_{i-1}}] \tag{139}$$

$$= \sum_{c_{\tau_i}} \Pr(c_{\tau_i}|Z_{\tau_1}, Z_{\tau_2}, \cdots, Z_{\tau_{i-1}}) \sum_{\bar{s}_{\tau_i+1}} \Pr(\bar{s}_{\tau_i+1}|X^{(i)}_{\bar{s},a}, a_{\tau_i}) Z_{\tau_i}(\boldsymbol{z}, X^{(i)}_{\bar{s},a}, a_{\tau_i}, \bar{Y}^{(i)}_{\bar{s},a}) \tag{140}$$

$$= \sum_{c_{\tau_i}} \Pr(c_{\tau_i}|Z_{\tau_1}, Z_{\tau_2}, \cdots, Z_{\tau_{i-1}}) \times 0 \tag{141}$$

$$= 0. \tag{142}$$

□

D R-MAX Adapted to Abstracted RL

Here we use the result of Theorem 2 to show that we can provide efficient learning guarantees for R-MAX (Brafman & Tennenholtz, 2002) in Abstracted RL. In Appendix D.1, we use Theorem 2 and the value bounds in Appendix A.3 to establish two supporting Lemmas. Then, in Appendix D.2, we adapt one lemma and the guarantees of R-MAX to Abstracted RL.

D.1 Supporting Lemmas

We can use Theorem 2 to determine the number of samples required to guarantee that the distance $||\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_x}(\cdot|\bar{s}, a)||_1$ will be smaller than ϵ with probability $1 - \delta$:

Lemma 10. *For inputs κ and ϵ ($0 < \kappa < 1, 0 < \epsilon < 2$), the following holds for a number of samples $m \geq \frac{2[\ln(2^{|\bar{S}|}-2)-\ln(\kappa)]}{\epsilon^2}$:*

$$\Pr(||\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_x}(\cdot|\bar{s}, a)||_1 \geq \epsilon) \leq \kappa. \tag{143}$$

Proof. To shorten the notation, we use the definitions $P_Y \triangleq \bar{T}_Y(\cdot|\bar{s}, a)$ and $P_{\omega_x} \triangleq \bar{T}_{\omega_x}(\cdot|\bar{s}, a)$. It follows from Theorem 2 that

$$\Pr(||P_Y - P_{\omega_x}||_1 \geq \epsilon) \leq 2^{|\bar{S}|} e^{-\frac{1}{8}m\epsilon^2}. \tag{144}$$

We need to select m such that $\kappa \geq 2^{|\bar{S}|} e^{-\frac{1}{8}m\epsilon^2}$:

$$\kappa \geq 2^{|\bar{S}|} e^{-\frac{1}{8}m\epsilon^2} \tag{145}$$

$$\frac{\kappa}{2^{|\bar{S}|}} \geq e^{-\frac{1}{8}m\epsilon^2} \tag{146}$$

$$\ln(\kappa) - \ln(2^{|\bar{S}|}) \geq -\frac{m\epsilon^2}{8} \tag{147}$$

$$\frac{m\epsilon^2}{8} \geq \ln(2^{|\bar{S}|}) - \ln(\kappa) \tag{148}$$

$$m \geq \frac{8[\ln(2^{|\bar{S}|}) - \ln(\kappa)]}{\epsilon^2}. \tag{149}$$

Thus if $m \geq \frac{8[\ln(2^{|\bar{S}|}) - \ln(\kappa)]}{\epsilon^2}$, we have

$$\Pr(||P_Y - P_{\omega_x}||_1 \geq \epsilon) \leq \kappa.$$

□

We want to give results for an empirical abstract model $\hat{\bar{M}}$ in the abstract space from ϕ, whose transition probabilities and rewards are within η_T and η_R, respectively, from those of an abstract MDP $\bar{M}$. We use $V^{*,n}$ to denote the value in M under the n-step optimal policy and $V^{\hat{\bar{\pi}}^*,n}$ to denote the value in M under the n-step optimal policy $\hat{\bar{\pi}}^*$ for $\hat{\bar{M}}$. The following lemma shows that we can upper bound the loss in value when applying $\hat{\bar{\pi}}^*$ to M:

Lemma 11. *Let M be an MDP, $\bar{M}$ an abstract MDP constructed using an approximate model similarity abstraction ϕ, with η_R and η_T, and $\hat{\bar{M}}$ an MDP in the abstract space from ϕ with*

$$|\bar{T}(\bar{s}'|\bar{s}, a) - \hat{\bar{T}}(\bar{s}'|\bar{s}, a)| \leq \epsilon, |\bar{R}(\bar{s}, a) - \hat{\bar{R}}(\bar{s}, a)| = 0. \tag{150}$$

Then

$$V^{*,n}(s) - V^{\hat{\bar{\pi}}^*,n}(s) \leq 2n\eta_R + (n-1)n(\eta_T + \epsilon)|\bar{S}|R_{max}. \tag{151}$$

Proof. Note that we assume that $|\bar{R}(\bar{s}, a) - \hat{\bar{R}}(\bar{s}, a)| = 0$ because we assume a deterministic reward. Then, we have

$$\forall \bar{s}, a \in \bar{S} \times A, s \in \bar{s}: \ |R(s, a) - \hat{\bar{R}}(\bar{s}, a)| \leq \eta_R, \tag{152}$$

$$\forall \bar{s}, a, \bar{s}' \in \bar{S} \times A \times \bar{S}, s \in \bar{s}: \ |\sum_{s' \in \bar{s}'} T(s'|s, a) - \hat{\bar{T}}(\bar{s}'|\bar{s}, a)| \leq \eta_T + \epsilon. \tag{153}$$

We use $\hat{\bar{V}}^{\hat{\bar{\pi}}^{*,n}}(\bar{s})$ to denote the n-step value under the n-step optimal policy $\hat{\bar{\pi}}^{*,n}$ for the empirical abstract MDP $\hat{\bar{M}}$. Then, by Theorem 1, we have $\forall s \in \bar{s}, \bar{s} \in \bar{S}$:

$$V^{*,n}(s) - V^{\hat{\bar{\pi}}^*,n}(s) = 2n\eta_R + (n-1)n(\eta_T + \epsilon)|\bar{S}|R_{\max}. \tag{154}$$

□

D.2 Proof of Theorem 4

First, we restate an Implicit Explore or Exploit Lemma that is used in the proof of R-MAX. We are interested in the event A_M, the event that we encounter an unknown state-action pair during an n-step trail in M. For two MDPs with different dynamics only in the unknown state-action pairs, the probability that we encounter an unknown state-action pair in an n-step trial is tiny if the difference in the n-step value between the two MDPs is slight. The proof follows the steps the proof of Lemma 3 from Strehl & Littman (2008).

Lemma 12 (Implicit Explore or Exploit)**.** *Let M be an MDP. Let L be the set of known abstract state-action pairs, and let M_L be an MDP that is the same as M on the known pairs $(\bar{s}, a) \in L$, but different on the unknown pairs $(\bar{s}, a) \notin L$. Let s be some state, and A_M the event that an unknown abstract state-action pair is encountered in a trial generated by starting from state s_1 and following π for n steps in M. Then,*

$$V_M^{\pi,n}(s_1) \geq V_{M_L}^{\pi,n}(s_1) - nR_{max}\Pr(A_M). \tag{155}$$

Proof. For a fixed path $p_t = s_1, a_1, r_1, \cdots, s_t, a_t, r_t$, we define $\Pr_M(p_t)$ as the probability that p_t occurs when running policy π in M starting from state s_1. We let L_t be the set of paths p_t such that there is at least one unknown state s_i in p_t ($(\phi(s_i), a) \notin L$). We further let $r_M(t)$ be the reward received at time t and $r_M(p_t, t)$ the reward at time t in the path p_t. We have the following:

$$E\big[r_{M_L}(t)\big] - E\big[r_M(t)\big] = \sum_{p_t \in L_t} \big(\Pr_{M_L}(p_t) r_{M_L}(p_t, t) - \Pr_M(p_t) r_M(p_t, t)\big) \tag{156}$$

$$+ \sum_{p_t \notin L_t} \big(\Pr_{M_L}(p_t) r_{M_L}(p_t, t) - \Pr_M(p_t) r_M(p_t, t)\big) \tag{157}$$

$$= \sum_{p_t \notin L_t} \big(\Pr_{M_L}(p_t) r_{M_L}(p_t, t) - \Pr_M(p_t) r_M(p_t, t)\big) \tag{158}$$

$$\leq \sum_{p_t \notin L_t} \Pr_{M_L}(p_t) r_{M_L}(p_t, t) \leq R_{\max}\Pr(A_M). \tag{159}$$

Here $\sum_{p_t \in L_t} \big(\Pr_{M_L}(p_t) r_{M_L}(p_t, t) - \Pr_M(p_t) r_M(p_t, t)\big) = 0$ because, by definition, M and M_L behave identically on the known state-action pairs, and $\sum_{p_t \notin L_t} \Pr_{M_L}(p_t) r_{M_L}(p_t, t) \leq R_{\max}\Pr(A_M)$ is true because $r_{M_L}(p_t, t)$ is at most $R_{\max}$. Finally we can write

$$V_{M_L}^{\pi,n}(s_1) - V_M^{\pi,n}(s_1) = \sum_{t=0}^{n} (E\big[r_{M_L}(t)\big] - E\big[r_M(t)\big]) \tag{160}$$

$$\leq nR_{\max}\Pr(A_M). \tag{161}$$

Thus, $V_M^{\pi,n}(s_1) \geq V_{M_L}^{\pi,n}(s_1) - nR_{\max}\Pr(A_M)$. □

Now we are ready to prove the theorem.

Theorem 4. *Given an MDP M, an approximate model similarity abstraction ϕ, with η_R and η_T, and inputs $|\bar{S}|, |A|, \epsilon, \delta, T_\epsilon$. With probability of at least $1 - \delta$ the R-MAX algorithm adapted to abstraction (Algorithm 1) will attain an expected return of $Opt(\prod_M(\epsilon, T_\epsilon)) - 3\frac{g(\eta_T, \eta_R)}{T_\epsilon} - 2\epsilon$ within a number of steps polynomial in $|\bar{S}|, |A|, \frac{1}{\epsilon}\frac{1}{\delta}, T_\epsilon$. Here T_ϵ is the ϵ-return mixing time of the optimal policy, the policies for M whose ϵ-return mixing time is T_ϵ are denoted by $\prod_M(\epsilon, T_\epsilon)$, the optimal expected T_ϵ-step undiscounted average return achievable by such policies are denoted by $Opt(\prod_M(\epsilon, T_\epsilon))$, and*

$$g(\eta_T, \eta_R) = T_\epsilon \eta_R + \frac{(T_\epsilon - 1)T_\epsilon}{2}\eta_T |\bar{S}| R_{max}.$$

Proof of Theorem 4. The proof uses elements of the Theorem from Brafman & Tennenholtz (2002). The proof follows the following steps:

1. We show that the expected average reward of the algorithm is at least as stated if the algorithm does not fail.
2. The probability of failing is at most δ. We can decompose this probability into three elements.
   (a) Probability that the transition function estimates are not within the desired bounds.
   (b) The probability that we do not attain the number of required visits in polynomial time.
   (c) The probability that the actual return is lower than the expected return.

Now we first assume the algorithm does not fail. We define an abstract MDP $\bar{M}_{\omega_X}$ constructed from ϕ with η_T and η_R. Similar to M_L, $\bar{M}_{\omega_X,L}$ is the same as $\bar{M}_{\omega_X}$ on the known abstract state-action pairs, but with a self-loop and the maximum reward ($R_{\max}$) on the unknown abstract state-action pairs, i.e., $\forall(\bar{s},a) \notin L : \bar{T}_{\omega_X,L}(\bar{s}|\bar{s},a) = 1, \bar{R}_{\omega_X,L}(\bar{s},a) = R_{\max}$. We also define an empirical abstract MDP $\bar{M}_Y$, of which the transition probabilities $\bar{T}_Y(\bar{s}'|\bar{s},a)$ are within some ϵ_2 (defined later) of those in $\bar{M}_{\omega_X}$ and with $\bar{R}_{\omega_X}(\bar{s},a) = \bar{R}_Y(\bar{s},a)$ because of the assumption that the rewards are deterministic. Then, $\bar{M}_{Y,L}$ is the abstract MDP that is the same as $\bar{M}_Y$ on the known abstract state-action pairs and the same as $\bar{M}_{\omega_X,L}$ on the unknown abstract state-action pairs. We denote the R-MAX policy with $\bar{\pi}$.

Let A_M be the event that, following $\bar{\pi}$, we encounter an unknown abstract state-action pair $(\phi(s),a) \notin L$ in T_ϵ steps. From Lemma 12, we have that:

$$\forall s \in S : V_M^{\bar{\pi},n}(s) \geq V_{M_L}^{\bar{\pi},n}(s) - T_\epsilon R_{\max} \Pr(A_M). \tag{162}$$

Now suppose that $R_{\max} \Pr(A_M) < \epsilon_1$, for some ϵ_1 (defined later), then we have

$$\begin{aligned}
V_M^{\bar{\pi},T_\epsilon}(s) &\geq V_{M_L}^{\bar{\pi},T_\epsilon}(s) - T_\epsilon R_{\max} \Pr(A_M) && (163)\\
&\geq V_{M_L}^{\bar{\pi},T_\epsilon}(s) - T_\epsilon \epsilon_1 && (164)\\
&\geq V_{\bar{M}_{\omega_X,L}}^{\bar{\pi},T_\epsilon}(s) - T_\epsilon \epsilon_1 - g(\eta_T,\eta_R) && (165)\\
&\geq V_{\bar{M}_{Y,L}}^{\bar{\pi},T_\epsilon}(s) - T_\epsilon \epsilon_1 - g(\epsilon_2) - g(\eta_T,\eta_R) && (166)\\
&\geq V_{\bar{M}_Y}^{*,T_\epsilon}(s) - T_\epsilon \epsilon_1 - g(\epsilon_2) - g(\eta_T,\eta_R) && (167)\\
&\geq V_M^{*,T_\epsilon}(s) - T_\epsilon \epsilon_1 - g(\epsilon_2) - g(\eta_T,\eta_R) - 2g(\eta_T+\epsilon_2,\eta_R). && (168)
\end{aligned}$$

Here the step from equation 163 to equation 164 follows because of the assumption that $R_{\max} \Pr(A_M) < \epsilon_1$. The step from equation 164 to equation 165 follows from Lemma 6, where $g(\eta_T,\eta_R) = T_\epsilon \eta_R + \frac{(T_\epsilon-1)T_\epsilon}{2}\eta_T|\bar{S}|R_{\max}$. The step from equation 165 to equation 166 follows from Lemma 8, where $g(\epsilon_2) = \frac{(T_\epsilon-1)T_\epsilon}{2}\epsilon_2|\bar{S}|R_{\max}$. The step from equation 166 to equation 167 follows because the R-MAX policy $\bar{\pi}$ is the optimal policy for $\bar{M}_{Y,L}$, and $\bar{M}_{Y,L}$ is the same as $\bar{M}_Y$ on the known state-action pairs and overestimates the value of the unknown state-action pairs (to the maximum value). Finally, the step from equation 167 to equation 168 follows from Lemma 11.

In equation 168 the results are for the undiscounted T_ϵ-step sum of rewards, so to obtain the result for the average reward per step, we have to divide equation 168 by T_ϵ. We get

$$\text{Opt}(\prod_M(\epsilon, T_\epsilon)) - T_\epsilon\epsilon_1/T_\epsilon - g(\epsilon_2)/T_\epsilon - g(\eta_T, \eta_R)/T_\epsilon - 2g(\eta_T + \epsilon_2, \eta_R)/T_\epsilon \tag{169}$$

$$= \text{Opt}(\prod_M(\epsilon, T_\epsilon)) - \epsilon_1 - \frac{(T_\epsilon - 1)T_\epsilon}{2}\epsilon_2|\bar{S}|R_{\max}/T_\epsilon$$

$$- (T_\epsilon\eta_R + \frac{(T_\epsilon - 1)T_\epsilon}{2}\eta_T|\bar{S}|R_{\max})/T_\epsilon - 2(T_\epsilon\eta_R + \frac{(T_\epsilon - 1)T_\epsilon}{2}(\eta_T + \epsilon_2)|\bar{S}|R_{\max})/T_\epsilon \tag{170}$$

$$= \text{Opt}(\prod_M(\epsilon, T_\epsilon)) - \epsilon_1 - \frac{(T_\epsilon - 1)}{2}\epsilon_2|\bar{S}|R_{\max}$$

$$- \eta_R - \frac{(T_\epsilon - 1)}{2}\eta_T|\bar{S}|R_{\max} - 2\eta_R - (T_\epsilon - 1)(\eta_T + \epsilon_2)|\bar{S}|R_{\max} \tag{171}$$

$$= \text{Opt}(\prod_M(\epsilon, T_\epsilon)) - \epsilon_1 - \frac{(T_\epsilon - 1)}{2}\epsilon_2|\bar{S}|R_{\max}$$

$$- 3\eta_R - \frac{(T_\epsilon - 1)}{2}\eta_T|\bar{S}|R_{\max} - (T_\epsilon - 1)\epsilon_2|\bar{S}|R_{\max} - (T_\epsilon - 1)\eta_T|\bar{S}|R_{\max} \tag{172}$$

$$= \text{Opt}(\prod_M(\epsilon, T_\epsilon)) - \epsilon_1 - 3\frac{(T_\epsilon - 1)}{2}\epsilon_2|\bar{S}|R_{\max} - 3\eta_R - 3\frac{(T_\epsilon - 1)}{2}\eta_T|\bar{S}|R_{\max} \tag{173}$$

$$= \text{Opt}(\prod_M(\epsilon, T_\epsilon)) - \epsilon_1 - 3\frac{(T_\epsilon - 1)}{2}\epsilon_2|\bar{S}|R_{\max} - 3\frac{g(\eta_T, \eta_R)}{T_\epsilon} \tag{174}$$

$$= \text{Opt}(\prod_M(\epsilon, T_\epsilon)) - \frac{3}{8}\epsilon - 3\frac{(T_\epsilon - 1)}{2}\epsilon_2|\bar{S}|R_{\max} - 3\frac{g(\eta_T, \eta_R)}{T_\epsilon} \tag{175}$$

$$= \text{Opt}(\prod_M(\epsilon, T_\epsilon)) - \frac{3}{8}\epsilon - 3\frac{(T_\epsilon - 1)}{2}\frac{3\epsilon}{4|\bar{S}|R_{\max}(T_\epsilon - 1)}|\bar{S}|R_{\max} - 3\frac{g(\eta_T, \eta_R)}{T_\epsilon}. \tag{176}$$

$$= \text{Opt}(\prod_M(\epsilon, T_\epsilon)) - \frac{3}{8}\epsilon - \frac{9}{8}\epsilon - 3\frac{g(\eta_T, \eta_R)}{T_\epsilon}. \tag{177}$$

$$= \text{Opt}(\prod_M(\epsilon, T_\epsilon)) - \frac{3}{2}\epsilon - 3\frac{g(\eta_T, \eta_R)}{T_\epsilon}. \tag{178}$$

In the step from equation 173 to equation 174 we use that $g(\eta_T, \eta_R) = T_\epsilon\eta_R + \frac{(T_\epsilon-1)T_\epsilon}{2}\eta_T|\bar{S}|R_{\max}$. Then, in the last steps, we define ϵ_1 and ϵ_2. In the step from equation 174 to equation 175 we set $\epsilon_1 = \frac{3}{8}\epsilon$. And in the step from equation 175 to equation 176 we set $\epsilon_2 = 3\epsilon/(4|\bar{S}|R_{\max}(T_\epsilon - 1))$.

The above assumed that the algorithm did not fail, but we cannot guarantee this with probability 1 within a number of steps that is polynomial in the input. We will show that we can upper bound the probability of failure by δ. There are three reasons why the algorithm could fail.

1. First, we need to show that the transition functions of $\bar{M}_Y$ are within $\eta_T + \epsilon_2$ of the transition functions of $\bar{M}_{\omega_X}$, with high probability. This is to ensure that, once all the abstract state-action pairs are known, the loss of value because of an inaccurate transition model, $V^{*,T_\epsilon}_{\bar{M}_Y} - V^{*,T_\epsilon}_M$ is within $2g(\eta_T + \epsilon_2, \eta_R) = 2T_\epsilon\eta_R + (T_\epsilon - 1)T_\epsilon(\eta_T + \epsilon_2)|\bar{S}|R_{\max}$ by Lemma 11. We can use the martingale concentration inequality to choose a number of samples K_1 so that the probability that our transition estimate is outside the desired bound is less than $\frac{\delta}{3|S||A|}$ for every abstract state-action pair if we sample each pair K_1 times. By Lemma 10, we can guarantee this by using $K_1 \geq \frac{2[\ln(2^{|\bar{S}|}-2)-\ln(\delta/(3|\bar{S}||A|))]}{(\frac{3\epsilon}{4|\bar{S}|R_{\max}(T_\epsilon-1)})^2} = \frac{32|\bar{S}|^2R^2_{\max}(T_\epsilon-1)^2[\ln(2^{|\bar{S}|}-2)-\ln(\delta/(3|\bar{S}||A|))]}{9\epsilon^2}$. Then, by applying the Union Bound on all $|\bar{S}|A$ pairs, we have that the total probability that any transition function is outisde of the desired bound is less than $\delta/3$.

2. Before we assumed that $R_{\max}\Pr(A_M) < \epsilon_1 (= \frac{3\epsilon}{8})$. Here we can show that after K_2 T_ϵ-step trials where $R_{\max}\Pr(A_M) \geq \frac{3\epsilon}{8}$, all the abstract state-action pairs are visited at least K_1 times (become known) with a probability of at least $1-\delta/3$ by using Hoeffding's Inequality. Let X_i be the indicator variable that is 1 if we visit an unknown abstract state-action pair in a trial, and 0 otherwise. For the trials where

$$R_{\max}\Pr(X_i = 1) \geq \frac{3\epsilon}{8}$$
$$\Pr(X_i = 1) \geq \frac{3\epsilon}{8}/R_{\max},$$

we can use Hoeffding's Inequality to establish an upper bound, we have:

$$\Pr(\sum_{i=1}^{K_2}((\frac{3\epsilon}{8}/R_{\max}) - X_i) \geq K_2^{2/3}) = \tag{179}$$

$$\Pr(\frac{3\epsilon}{8}\frac{K_2}{R_{\max}} - \sum_{i=1}^{K_2} X_i \geq K_2^{2/3}) \leq e^{-\frac{2(K_2^{2/3})^2}{K_2}} \leq e^{-\frac{2(K_2^{2/3})^2}{K_2}} = e^{-2K_2^{1/3}}, \tag{180}$$

$$\Pr(\frac{3\epsilon}{8}\frac{K_2}{R_{\max}} - K_2^{2/3} \geq \sum_{i=1}^{K_2} X_i) \leq e^{-2K_2^{1/3}}. \tag{181}$$

After K_2 exploring episodes we want $\sum_{i=1}^{K_2} X_i$, the number of visits to unknown state-action pairs, to be $K_1|\bar{S}||A|$. So we can choose K_2 such that $\frac{3\epsilon}{8}\frac{K_2}{R_{\max}} - K_2^{2/3} \geq K_1|\bar{S}||A|$, and $e^{-2K_2^{1/3}} \leq \delta/3$ to guarantee that the probability of failing to explore enough is at most $\delta/3$.

3. Finally, the actual return may be lower than the expected return when we perform a T_ϵ-step trial where we do not explore. We use Hoeffding's Inequality to determine the number of steps K_3 needed to ensure that the actual average return is within $\epsilon/2$ of $\mathrm{Opt}(\prod_M(\epsilon, T_\epsilon)) - \frac{3}{2}\epsilon - 3\frac{g(\eta_T,\eta_R)}{T_\epsilon}$. We need to choose K_3 so that the probability of obtaining an actual return that is smaller than the desired $\mathrm{Opt}(\prod_M(\epsilon, T_\epsilon)) - 2\epsilon - 3\frac{g(\eta_T,\eta_R)}{T_\epsilon}$ is at most $\delta/3$ within $K_3 = Z|\bar{S}|T_\epsilon$ exploitation steps, with some number $Z > 0$. Let X_i denote the average return in the i-th exploitation step and μ the average expected return in an exploitation step so that μ is at least $\mathrm{Opt}(\prod_M(\epsilon, T_\epsilon)) - \frac{3}{2}\epsilon - 3\frac{g(\eta_T,\eta_R)}{T_\epsilon}$. Then

$$\Pr(\sum_{i=1}^{K_3}(\frac{\mu - X_i}{R_{\max}}) \geq K_3^{2/3}) \leq e^{-2\frac{(K_3^{2/3})^2}{K_3}} = e^{-2K_3^{1/3}}. \tag{182}$$

This means that, with a probability of at most $e^{-2K_3^{1/3}}$, the average return for K_3 exploitation steps is more than $\frac{R_{\max}}{K_3^{1/3}}$ lower than μ:

$$\Pr(\sum_{i=1}^{K_3}(\frac{\mu - X_i}{R_{\max}}) \geq K_3^{2/3}) \leq e^{-2K_3^{1/3}}, \tag{183}$$

$$\Pr(K_3\frac{\mu}{R_{\max}} - \sum_{i=1}^{K_3}\frac{X_i}{R_{\max}} \geq K_3^{2/3}) \leq e^{-2K_3^{1/3}}, \tag{184}$$

$$\Pr(K_3\mu - \sum_{i=1}^{K_3} X_i \geq R_{\max}K_3^{2/3}) \leq e^{-2K_3^{1/3}}, \tag{185}$$

$$\Pr(\mu - \sum_{i=1}^{K_3}\frac{X_i}{K_3} \geq \frac{R_{\max}}{K_3^{1/3}}) \leq e^{-2K_3^{1/3}}. \tag{186}$$

We can now choose Z, so that $\epsilon/2 \leq \frac{R_{\max}}{(Z|\bar{S}|T_\epsilon)^{\frac{1}{3}}}$ and $e^{-2(Z|\bar{S}|T_\epsilon)^{1/3}} \leq \delta/3$, to get the desired result: with probability at most $\delta/3$ the obtained value will be more than $\epsilon/2$ lower than the expected value.

The probability of failure is thus at most $3 * \delta/3 = \delta$, and an average return at most $2\epsilon + 3\frac{g(\eta_T,\eta_R)}{T_\epsilon}$ lower than $\text{Opt}(\prod_M(\epsilon, T_\epsilon))$ will be obtained with a probability of at least $1-\delta$. □

E Simulator Data Collection

Here we assume that we have access to a simulator and use this in our procedure to give a guarantee in the form of the abstract L1 inequality from equation 16. To some extent, this is not surprising, but to the best of our knowledge, this is the first work that explicitly shows how to combine MBRL and abstraction using a simulator. We assume that the simulator allows us to select (or move to) any state and draw a sample from its transition function, which we call the independent samples assumption:

Assumption 1 (Independent samples)**.** *We assume we can obtain independent samples, e.g., for any state-action pair (s, a), we can draw samples directly from its transition function $T(\cdot|s, a)$.*

Algorithm 2 Procedure: Abstracted MBRL

Input: $M, \phi, \delta, \epsilon, \pi$
$\bar{Y} = \textsc{CollectSamples}(M, \phi, \delta, \epsilon, \pi)$
The sampling results in sequences $\bar{Y}_{\bar{s},a}$, one for every pair $(\bar{s}, a)$:
$\bar{Y}_{\bar{s},a} = \phi(s'^{(1)}), \cdots, \phi(s'^{(m)})$
$= \bar{s}'^{(1)}, \cdots, \bar{s}'^{(m)}$
for all $(\bar{s}, a, \bar{s}') \in \bar{S} \times A \times \bar{S}$ **do**
  $\bar{T}_Y(\bar{s}'|\bar{s}, a) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{\bar{Y}^{(i)}_{\bar{s},a} = \bar{s}'\}$
end for
$\bar{M}_Y \triangleq \langle \bar{S}, A, \bar{T}_Y, \bar{R}, \gamma\rangle$
$\bar{\pi}^*_Y = \text{Value Iteration}(\bar{M}_Y)$
Apply $\bar{\pi}^*_Y$ to M

Algorithm 3 CollectSamples with Simulator

Input: $M, \phi, \delta, \epsilon$
$\kappa = \frac{\delta}{|\bar{S}||A|}$
$m = \lceil \frac{2[\ln(2^{|\bar{S}|}-2)-\ln(\kappa)]}{\epsilon^2} \rceil$
for all $(\bar{s}, a) \in \bar{S} \times A$ **do**
  $\bar{Y}_{\bar{s},a} = [\ ]$
  $x_{\bar{s},a}$ = select a prototype state $s \in \bar{s}$
  for $i = 1 : m$ **do**
    $s' = \text{Sample}(T(\cdot|x_{\bar{s},a}, a))$
    $\bar{Y}_{\bar{s},a}.\text{append}(\phi(s'))$
  end for
end for
Return: all $\bar{Y}_{\bar{s},a}$

If a simulator of the MDP is available, this is a reasonable assumption. For every pair $(\bar{s}, a)$, the simulator sampling procedure (Algorithm 3) selects a prototype $x_{\bar{s},a} \in \bar{s}$ from which to sample. We define a weighting function $\omega_x(s, a)$ that has a weight of 1 if s is the prototype $x_{\bar{s},a}$ and 0 otherwise:

$$\forall_{(\bar{s},a),s\in\bar{s}}\ \omega_x(s, a) \triangleq \mathbb{1}\{s = x_{\bar{s},a}\}. \tag{187}$$

Then we use this ω_x to define the abstract transition function $\bar{T}_{\omega_X}$ according to equation 7. $\bar{T}_{\omega_x}(\bar{s}'|\bar{s}, a) = \sum_{s'\in\bar{s}'} T(s'|s = x_{\bar{s},a}, a)$. This way, the samples we collect for one pair $(\bar{s}, a)$ are i.i.d. They are independent because of Assumption 1 and identically distributed because we sample from the prototype. Because the samples are i.i.d., we can use Lemma 1. We show that, with the simulator we can combine MBRL with abstraction and still learn an accurate model. We can guarantee that $\bar{T}_Y$ will be close to $\bar{T}_{\omega_x}$, with a high probability:

Theorem 5. *Under assumption 1, following the procedure in Algorithm 1, with the data collection from Algorithm 3 and inputs $|\bar{S}|, A, \epsilon$, and δ. For $\bar{T}_Y$ constructed by the algorithm, we have that with probability $1-\delta$, the following holds:*

$$\forall_{(\bar{s},a)}\ ||\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_x}(\cdot|\bar{s}, a)||_1 \leq \epsilon. \tag{188}$$

E.1 Proof of Theorem 5

Before starting with the actual proof, we first go over Algorithm 3 and give two lemmas the proof uses.

The agent will draw samples using the simulator as described in Algorithm 3. Since we assume we can sample directly from the transition functions $T(\cdot|s, a)$, this algorithm loops over all pairs $(\bar{s}, a)$ and samples m times[7]

[7]The value of m in Algorithm 3 is chosen based on the results further along in this section.

from each transition function. More formally, for every pair $(\bar{s}, a)$, the algorithm selects one prototype state $x_{\bar{s},a} = s \in \bar{s}$. Then, it loops over every pair $(\bar{s}, a)$ and samples m transitions from $T(\cdot|x_{\bar{s},a}, a)$. The set of collected experiences for each abstract state-action pair $(\bar{s}, a)$ is represented by $\bar{Y}_{\bar{s},a}$, as defined by equation 12.

Given $\bar{Y}_{\bar{s},a}$, we define the learned model $\bar{T}_Y(\cdot|\bar{s}, a)$ according to equation 13, $\bar{T}_{\omega_x}$ according to equation 15, and ω_x according to equation 187. It follows from Lemma 1 that we can derive a number of samples that we require to guarantee that $\Pr(||\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_x}(\cdot|\bar{s}, a)||_1 \geq \epsilon) \leq \kappa$ is true for inputs κ and ϵ:

Lemma 4. *For inputs κ and ϵ ($0 < \kappa < 1, 0 < \epsilon < 2$), we have that the following holds for $m \geq \frac{2[\ln(2^{|\bar{S}|}-2)-\ln(\kappa)]}{\epsilon^2}$:*

$$\Pr(||\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_x}(\cdot|\bar{s}, a)||_1 \geq \epsilon) \leq \kappa. \tag{189}$$

Proof. To shorten the notation, we use the definitions $P_Y \triangleq \bar{T}_Y(\cdot|\bar{s}, a)$ and $P_{\omega_x} \triangleq \bar{T}_{\omega_x}(\cdot|\bar{s}, a)$. From Lemma 1, we have that

$$\Pr(||P_Y - P_{\omega_x}||_1 \geq \epsilon) \leq (2^{|\bar{S}|} - 2)e^{-\frac{1}{2}m\epsilon^2}. \tag{190}$$

We need to select m such that $\kappa \geq (2^{|\bar{S}|} - 2)e^{-\frac{1}{2}m\epsilon^2}$:

$$\kappa \geq (2^{|\bar{S}|} - 2)e^{-\frac{1}{2}m\epsilon^2} \tag{191}$$

$$\frac{\kappa}{2^{|\bar{S}|} - 2} \geq e^{-\frac{1}{2}m\epsilon^2} \tag{192}$$

$$\ln(\kappa) - \ln(2^{|\bar{S}|} - 2) \geq -\frac{m\epsilon^2}{2} \tag{193}$$

$$\frac{m\epsilon^2}{2} \geq \ln(2^{|\bar{S}|} - 2) - \ln(\kappa) \tag{194}$$

$$m \geq \frac{2[\ln(2^{|\bar{S}|} - 2) - \ln(\kappa)]}{\epsilon^2} \tag{195}$$

Thus, if $m \geq \frac{2[\ln(2^{|\bar{S}|}-2)-\ln(\kappa)]}{\epsilon^2}$ we have

$$\Pr(||P_Y - P_{\omega_x}||_1 \geq \epsilon) \leq \kappa.$$

□

Using the Union bound, we can give a lower bound on the probability that $\bar{T}_Y(\cdot|\bar{s}, a)$ and $\bar{T}_{\omega_x}(\cdot|\bar{s}, a)$ are ϵ close for every $(\bar{s}, a)$:

Lemma 5. *If*

$$\forall_{(\bar{s},a)} \left[\Pr(||\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_x}(\cdot|\bar{s}, a)||_1 \geq \epsilon)\right] \leq \frac{\delta}{|\bar{S}||A|} \tag{196}$$

then the following holds with a probability of at least $1 - \delta$:

$$\max_{(\bar{s},a)} \left[||\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_x}(\cdot|\bar{s}, a)||_1\right] \leq \epsilon. \tag{197}$$

Proof. We define

$$\Delta_{\bar{s},a} \triangleq ||\bar{T}_Y(\cdot|\bar{s}, a) - \bar{T}_{\omega_x}(\cdot|\bar{s}, a)||_1. \tag{198}$$

Then $\Pr(\max_{(\bar{s},a)}\{\Delta_{\bar{s},a} \geq \epsilon\})$ is the probability that $\Delta_{\bar{s},a} \geq \epsilon$ for at least one abstract state-action pair. From the union bound, it follows that $\Pr(\max_{(\bar{s},a)}\{\Delta_{\bar{s},a} \geq \epsilon\}) \leq \delta$:

$$\Pr(\max_{(\bar{s},a)}\{\Delta_{\bar{s},a} \geq \epsilon\}) \leq \sum_{\bar{s},a} \Pr(\Delta_{\bar{s},a} \geq \epsilon) \tag{199}$$

$$\leq \sum_{\bar{s},a} \frac{\delta}{|\bar{S}||A|} \tag{200}$$

$$= \delta. \tag{201}$$

It follows that $\Pr(\max_{(\bar{s},a)}\{\Delta_{\bar{s},a} \leq \epsilon\}) \geq 1-\delta$ since $\Pr(\max_{(\bar{s},a)}\{\Delta_{\bar{s},a} \leq \epsilon\}) = 1 - \Pr(\max_{(\bar{s},a)}\{\Delta_{\bar{s},a} \geq \epsilon\})$. Thus the probability that equation 197 holds is at least $1-\delta$. □

Now we are ready to proof Theorem 5:

Proof of Theorem 5. By Assumption 1, and the earlier assumption that $|S|$ and $|A|$ are finite, we have that we can obtain m samples in finite time for every abstract state-action pair and any $m > 0$. Given the inputs $|\bar{S}|, A, \epsilon$, and δ, Algorithm 3 sets $m = \lceil \frac{2[\ln(2^{|\bar{S}|}-2)-\ln(\kappa)]}{\epsilon^2} \rceil$, where $\kappa = \frac{\delta}{|\bar{S}||A|}$. Then, for every $(\bar{s}, a)$, a prototype state $x_{\bar{s},a} = s \in \bar{s}$ is selected. We use equation 187 to define ω_x and equation 15 to define $\bar{T}_{\omega_x}$.

For all $(\bar{s}, a)$, Algorithm 3 obtains a sequence $\bar{Y}_{\bar{s},a}$ by sampling from the transition function from the prototype state $x_{\bar{s},a}$ and Algorithm 2 constructs the empirical transition functions as in equation 13.

Given our choice of m and the inputs $\kappa = \frac{\delta}{|\bar{S}||A|}$ and ϵ, it follows from Lemma 10 that

$$\forall_{(\bar{s},a)} \ \Pr(||\bar{T}_Y(\cdot|\bar{s},a) - \bar{T}_{\omega_x}(\cdot|\bar{s},a)||_1 \geq \epsilon) \leq \frac{\delta}{|\bar{S}||A|}. \tag{202}$$

By the union bound, we have that the following holds with a probability of at least $1-\delta$:

$$\forall_{(\bar{s},a)} \ ||\bar{T}_Y(\cdot|\bar{s},a) - \bar{T}_{\omega_x}(\cdot|\bar{s},a)||_1 \leq \epsilon. \tag{203}$$

□