Efficient Reinforcement Learning in Parameterized Models: Discrete Parameter Case.

Kirill Dyagilev
kirilld@tx.technion.ac.il
Department of EE
Technion
Haifa, Israel, 32000

Shie Mannor
shie.mannor@mcgill.ca
Department of ECE
McGill University
Montreal, Canada, H3A-2A7

Nahum Shimkin
shimkin@ee.technion.ac.il
Department of EE
Technion
Haifa, Israel, 32000

Abstract

We consider reinforcement learning in the parameterized setup, where the model is known to belong to a parameterized family of Markov Decision Processes (MDPs). We further impose here the assumption that set of possible parameters is finite, and consider the discounted return. We propose an on-line algorithm for learning in such parameterized models, dubbed the Parameter Elimination (PEL) algorithm, and analyze its performance in terms of the the total mistake bound criterion (also known as the sample complexity of exploration). The algorithm relies on Wald's Sequential Probability Ratio Test to eliminate unlikely parameters, and uses an optimistic policy for effective exploration. We establish that, with high probability, the total mistake bound for the algorithm is linear (up to a logarithmic term) in the size of the parameter space, independently of the cardinality of the state and action spaces.

1 Introduction

In the Reinforcement Learning (RL) framework, an agent interacts with a partially known environment with purpose of maximizing some numerical utility measure based on observations of the environment state and reward signals. ([SB98]). The environment is often modeled as a Markov Decision Process (MDP) with finite state and action spaces. One possible goal for the agent is to learn an (almost-)optimal control policy as quickly as possible. Alternatively, in an on-line setting where learning is performed during normal system operation, the agent's goal may be to maximize the actually obtained reward or to minimize of suboptimal moves relative to the optimal (non-learning) policy.

A fundamental issue that greatly affects the convergence rate of RL algorithms is the efficient exploration of the state and action spaces. In an on-line setting, in particular, the agent faces the well-known exploration-exploitation trade off: whether to keep trying to acquire new information (explore), or to act consistently with accumulated information to maximize reward (exploit). An efficient solution of this trade off is essential to obtain acceptable convergence guarantees.

Several different measures of the convergence rate were proposed for the on-line RL problem. These include the number of exploratory episodes [KS02, BT02], the number of time steps the agent follows a sub-optimal action, and the number of times the agent spends following a non-optimal policy [Kak03]. We shall consider here the latter two, and refer to them as the action-mistake count and policy-mistake count (also known as the *sample complexity of exploration*), respectively. A learning algorithm is said to be PAC (Probable Approximately Correct) if its relevant convergence-rate metric is polynomial in the model size parameters with high probability.

In recent years several PAC algorithms were introduced. These include the R-max algorithm [BT02], further analyzed in [Kak03], the MBIE algorithm [SL05] and the Delayed Q-learning algorithm [SLW$^+$06]. These algorithms do not use any prior knowledge on the model parameters, but rather estimate empirically either the transition probabilities or directly the Q-function for all state-action pairs. As a result, their convergence-rate metrics are at least proportional to the cardinality of the state and action spaces, which may not be acceptable for large problems. Possible approaches to handle such problems include various approximation schemes, and the use of prior knowledge about the system to enhance learning performance.

An effective use of *structural* knowledge about the system has been demonstrated for factored MDPs in [KK99]. Here we consider the case where a parameterized model of the system in question is available. The potential problem simplification offered by such models can be demonstrated through a simple queueing example. Consider a single-server queue with buffer and server capacity of 100 customers. Assume that the arrival and service processes are Poisson processes with rate parameters λ and μ, respectively. In this model, all transition probabilities are determined by two parameters only. Therefore, although the cardinality of the state space is 100, it is enough to estimate only two parameters in order to find an optimal policy. This observation turns out to be even more acute in case of a queuing system that contains several such queues, say $N > 1$. While the cardinality of the state space grows exponentially to 100^N, which makes learning on a per-state basis infeasible, the number of parameters associated with arrival/service processes grows linearly to $2N$.

Parameterized control models, in which all model parameters are defined in terms of a smaller parameter vector, have been extensively studied in the context adaptive control, and in particular stochastic adaptive control [KV98]. Several im-

portant issues were raised and formalized in this context, including the closed-loop identifiability problem [Man74] and the principle of "optimism in face of uncertainty"[KB82]. However, the results of this research are focused mostly on asymptotic convergence results, rather than on polynomial convergence bounds.

Our focus in this paper is on parameterized system models with a *finite* parameter space. We further consider the discounted reward problem. We present an efficient RL algorithm for this case, called the Parameter Elimination (PEL) algorithm, and show that its total mistake bound grows linearly (up to logarithmic terms) in the size of parameter space, and independently of the size of the state and action spaces. Essentially, the PEL algorithm operates as follows. It maintains a list of plausible parameters J, which is initially equal to the entire parameter set. Parameters are then eliminated one-by-one from the plausible set using the Sequential Probability Ratio Test (SPRT) [Wal52] based on the observed history. As for action selection, at every step t an "optimistic" parameter (relative to the current state) is selected from the set J. This parameter is the one that maximized the (discounted) value function from the current state. The current action is then selected as the optimal one for the optimistic parameter.

While the finite parameter case may be considered a simplified abstraction, it can serve as an approximation to the continuous parameter case through discretization. A detailed treatment of this approach falls beyond the scope of the present paper.

The rest of the paper is organized as follows. In Section 2 we present the model along with some definitions and notations. Section 3 defines the main performance metrics considered in this paper. In Section 4 we present the PEL algorithm and provide our main performance bounds for this algorithm. Section 5 is devoted to the proof of these results. In Section 6 we summarize the results obtained so far and discuss future work.

2 Model Formulation

An MDP M is specified by a five-tuple $\langle S, A, R, p, \eta \rangle$, where S is a finite state space, A is a finite action space, R is a finite reward set, $p : S \times A \to \Delta(S)$ is the transition probability function and $\eta : S \times A \to \Delta(R)$ is the reward probability function. Here $\Delta(S)$ denotes the set of probability vectors over the set S, and similarly for $\Delta(R)$. Given that at the time step t the state is $s_t \in S$ and the action is $a_t \in A$, the agent receives a random reward $r_t \in R$ with probability $\eta(r_t|s_t, a_t)$ and moves to state $s_{t+1} \in S$ with probability $p(s_{t+1}|s_t, a_t)$.

The observed history until time t is the sequence $h_t \stackrel{\triangle}{=} \{s_0, a_0, r_0, ..., s_{t-1}, a_{t-1}, r_{t-1}, s_t\}$. A (deterministic) decision rule is a mapping from history to action, namely $\pi_t : H_t \to A$, where $H_t = (S \times A \times R)^t \times S$. A policy $\mathcal{A}$ is a collection of decision rules $\{\pi_t\}_{t=0}^{\infty}$ so that $a_t = \pi_t(h_t)$. Note that a (deterministic) learning algorithm is such a policy. Given an initial state s, the policy $\mathcal{A}$ induces a stochastic process $(s_t, a_t, r_t)_{t=0}^{\infty}$ with probability measure $\mathbb{P}^{\mathcal{A},s}\cdot$. The expectation operator corresponding to this measure is denoted by $\mathbb{E}^{\mathcal{A},s}$.

Let $V^{\mathcal{A}}(s) \stackrel{\triangle}{=} \mathbb{E}^{\mathcal{A},s}\{\sum_{t=0}^{\infty} \gamma^t r_t\}$ denote the discounted return for policy $\mathcal{A}$ from state s. Here $0 < \gamma < 1$ is the discount factor, which we fix from now on. We refer to $V^{\mathcal{A}}(s)$ as the value function for policy $\mathcal{A}$. A policy $\mathcal{A} = \{\pi_t\}_{t=0}^{\infty}$ is called stationary if $\pi_t = \pi$ for all t, and, $\pi : S \to A$ is a function of the current state only. We hence use π to denote both the mapping $\pi : S \to A$ and the corresponding stationary policy. It is well known (e.g., [Put94]) that there exists a deterministic stationary policy π^* which is optimal in sense that $V^{\pi^*}(s) \geq V^{\mathcal{A}}(s)$ for any state s and any policy $\mathcal{A}$. Denote the corresponding optimal value function as $V^*(\cdot)$. Further define the action-value function (or Q-function) for state-action pair (s, a) as

$$Q^*(s, a) = \bar{r}(s, a) + \sum_{s' \in S} p(s'|s, a) V^*(s'),$$

where $\bar{r}(s, a) \stackrel{\triangle}{=} \sum_{r \in R} r\eta(r|s, a)$ denotes an expected reward for the state-action pair (s, a). The following equality, known as Bellman equation, holds for any stationary policy π and state $s \in S$:

$$V^{\pi}(s) = \bar{r}(s, \pi(s)) + \gamma \sum_{s' \in S} V^{\pi}(s') p(s'|s, \pi(s)), \quad (2.1)$$

while the optimal value function of policy $\pi^*(s)$ satisfies

$$V^*(s) = \max_a Q^*(s, a) \equiv Q^*(s, \pi^*(s)).$$

In this paper we assume that the true MDP belongs to a known family $\{M_\theta\}_{\theta \in \Theta}$ of parameterized models, where Θ is a finite parameter set. All models in the given family share the same action, reward and state spaces, while their transition and reward probabilities depend on the parameter $\theta \in \Theta$, i.e., $M_\theta =< S, A, R, p_\theta, \eta_\theta >$. For each MDP M_θ we denote by π^*_θ, V^*_θ and Q_θ an optimal stationary policy, the optimal value function and the Q-function, respectively. In case the optimal policy is not unique, we henceforth fix one (arbitrary) selection. The actual model M is thus corresponds to some parameter $\theta_0 \in \Theta$, namely $M = M_{\theta_0}$. We refer to θ_0 as the *true parameter*.

3 Performance Metrics

An effective measure of on-line learning efficiency in an RL problem is the number of time steps the algorithm prescribes sub-optimal control. We consider two possible criteria for sub-optimality at a given step. The first criterion examines the expected discounted return from the present step onward relative to the optimal one. The second criterion concerns sub-optimality of the action taken at that step. We next introduce these two criteria and show that they are closely related.

To define the first criterion, we introduce the notion of the algorithm's discounted return from the current step on. Denote by $\mathcal{A}$ the policy of the learning algorithm. Let h_τ be the observed history up to time τ, and denote by

$$V^{\mathcal{A}}(h_\tau) \stackrel{\triangle}{=} \mathbb{E}^{\mathcal{A},s_0}\left\{ \sum_{j=\tau}^{\infty} \gamma^{j-\tau} r_j \,\middle|\, h_\tau \right\}$$

the value of the policy $\mathcal{A}$ starting from time τ. The policy mistake count is defined as follows:

Definition 1 *Let ϵ be a positive number. The time step t in said to be an ϵ-**suboptimal step** if $V^{\mathcal{A}}(h_t) < V^*(s_t) - \epsilon$. Equivalently, we say that the learning agent follows an ϵ-**suboptimal policy** at time t. The **policy-mistake count** (PMC) of a learning algorithm $\mathcal{A}$ is defined as*

$$PMC(\epsilon) \stackrel{\triangle}{=} \sum_{t=0}^{\infty} \mathbb{I}\left\{V^{\mathcal{A}}(h_t) < V^*(s_t) - \epsilon\right\}.$$

The PMC counts the total number of ϵ-optimal time steps in the sense of Definition 1. This criterion was introduced in [Kak03], where it is also called the sample complexity of exploration, and further studied in [SL05, SLW$^+$06].

We now proceed to define the second criterion - the action-mistake count (AMC). Recall that an optimal action $a^* = \pi^*(s)$ in state s satisfies $Q^*(s, a^*) = V^*(s)$. Hence the difference $V^*(s) - Q^*(s, a)$ quantifies the effect of taking a single suboptimal action a, and thereafter proceeding optimally. The AMC measures the total number of ϵ-suboptimal state-action pairs visited by algorithm during its operation.

Definition 2 *For any $\epsilon \geq 0$, a state-action pair (s, a) is called ϵ-**suboptimal** if $Q^*(s, a) < V^*(s) - \epsilon$. The **action-mistake count** of a learning algorithm is defined as*

$$AMC(\epsilon) \stackrel{\triangle}{=} \sum_{t=0}^{\infty} \mathbb{I}\left\{Q^*(s_t, a_t) < V^*(s_t) - \epsilon\right\}.$$

Note that for ϵ small enough $AMC(\epsilon) = AMC(0)$ (due to the finiteness of the state and action spaces), so that only non-optimal actions are counted.

The AMC is dominated by the PMC as the next lemma indicates.

Lemma 3 *For any $\epsilon > 0$ and learning algorithm $\mathcal{A}$ the following inequality holds almost surely:*

$$AMC(\epsilon) \leq PMC(\epsilon).$$

Proof: Since a_t is the action chosen by $\mathcal{A}$ at time t, it follows by definition of $Q^*(s, a)$ that $V^{\mathcal{A}}(h_t) \leq Q^*(s_t, a_t)$. ■

It follows that any upper bound for the PMC also applies to AMC. For this reason we shall focus in the following on the PMC alone. We can now define the corresponding notion of a PAC algorithm in the following way:

Definition 4 *A learning algorithm A is called **PMC-PAC** (or just PAC) if, for any positive ϵ and δ, its policy-mistake count (action-mistake count) is polynomial in $\left(|\Theta|, \epsilon^{-1}, \delta^{-1}, (1-\gamma)^{-1}\right)$ with probability of at least $(1-\delta)$.*

4 The PEL algorithm

In discrete parameterized models, the learning problem may be reduced to the identification of the true parameter or, at least, a parameter that leads to a near-optimal control policy for the true model. Equivalently, one may try to eliminate all other parameters from the set of optional parameters.

Let R_{max} denote an upper bound on the one-step expected reward, namely

$$R_{max} \geq \max_{(s,a)\in S\times A} \max_{\theta\in\Theta}\{\bar{r}_\theta(s, a)\}.$$

Define the log-likelihood function of the observation $(s_{t-1}, a_{t-1}, r_{t-1}, s_t)$ at time step t as

$$l_t(\theta) = \log p_\theta(s_t|s_{t-1}, a_{t-1}) + \log \eta_\theta(r_{t-1}|s_{t-1}, a_{t-1}). \tag{4.1}$$

The *cumulative* log-likelihood is then $G_t(\theta) = \sum_{i=1}^{t} l_t(\theta)$.

The PEL algorithm proceeds as follows (see Algorithm 1 for details). As an input, the algorithm requires the finite family of possible MDPs $\{M_\theta\}_{\theta\in\Theta}$, with common state, reward and action spaces. The value function $V_\theta^*(\cdot)$ and the optimal policy $\pi_\theta^*(\cdot)$ for each model can be calculated using one of the standard algorithms, i.e., value iteration, policy iteration or linear programming (see [Put94]). An accuracy parameter ϵ and an allowed probability of error δ are also provided as input.

The algorithm maintains a list of plausible parameters J_t throughout its execution. Initially, all parameter values are considered plausible and then they are eliminated one by one. The elimination step is based on the Sequential Probability Ratio Test (SPRT), namely, comparing the log-likelihood ratio $G_t(\theta_i) - G_t(\theta_j)$ to a given threshold $G_{th} > 0$. If at time step t there exist parameters $\theta_i, \theta_j \in J_t$ so that $G(\theta_i) - G(\theta_j) > G_{th}$ then θ_j is eliminated. Equivalently, we first find $\hat{\theta}$, the most likely parameter in the set J_t, and then compare the likelihood of all other plausible parameters to $G(\hat{\theta})$. As the error probability of each elimination can be upper bounded by $e^{-G_{th}}$, the selection of $G_{th} = \log\left[\frac{3(|\Theta|-1)}{\delta}\right]$ yields cumulative error probability of all eliminations less than $\frac{\delta}{3}$ (see subsection 5.4 for details).

The exploration-exploitation tradeoff is addressed using the so-called "optimism in face of uncertainty" principle. At each time step t, the PEL algorithm selects an "optimistic" action in the following sense. First, the algorithm selects the parameter $\theta(t)$ that maximizes the value function $V_\theta^*(s_t)$ for the current state s_t, among all parameters in the plausible set J_t. The selected action is then the optimal action given $\theta(t)$, i.e., $a_t = \pi_{\theta(t)}^*(s_t)$. We note that unlike some PAC algorithms such as E^3 and R-max [KS02, BT02], PEL does not freeze its policy over long pre-determined intervals, but rather updates it each time some parameter is eliminated. Furthermore, the selected action generally corresponds to a different parameter θ at each state.

The main result of the paper is the following one.

Theorem 5 *Consider the PEL algorithm with parameter $0 < \epsilon < \frac{R_{max}}{(1-\gamma)}$ and $0 < \delta < 1$. With probability of at least $1-\delta$, PEL's policy-mistake count is upper bounded by*

$$PMC(\epsilon) \leq \kappa\,|\Theta|\,\frac{R_{max}^3}{\epsilon^3(1-\gamma)^6} \tag{4.2}$$

time steps, where $\kappa \stackrel{\triangle}{=} 801 \log\left(\frac{3|\Theta|}{\delta}\right) \log\frac{4R_{max}}{\epsilon(1-\gamma)}$.

A slightly tighter bound is given in (5.14). This theorem implies that the PEL algorithm is PAC in terms of the total mistake bound, and its PMC is linear (up to a logarithmic term) in the size of the parameter set. Note that the bound is independent of the cardinality of the state and action spaces. In the following section we establish the proof of Theorem 5.

Algorithm 1 Parameter ELimination

Input: $\{M_\theta\}_{\theta\in\Theta}$ – the finite family of possible MDPs, ϵ – a required accuracy of policy estimation, δ – an allowed probability of error.

Initialize: Initialize the list of plausible parameter values to $J_0 = \Theta$. Initialize the array of cumulative log-likelihood to $G_0(\theta) = 0$ for all $\theta \in \Theta$.

For $t = 0, 1, ...$ **do**

1. **Stopping condition**: If J_t is a singleton, namely $J_t = \{\theta\}$, then use the corresponding policy π^*_θ indefinitely and skip items (2)-(5) below.
2. **Find an optimistic parameter**: Select a parameter value that maximizes the value function among plausible parameter values: $\theta(t) := \arg\max_{\theta\in J_t} V^*_\theta(s_t)$.
3. **Act**: Execute the action according to the optimal policy for the optimistic parameter: $a_t := \pi^*_{\theta(t)}(s_t)$.
4. **Update**: Observe the reward r_t and the next state s_{t+1}. Update for all $\theta \in J_t$: $G_{t+1}(\theta) := G_t(\theta) + l_{t+1}(\theta)$ where l_{t+1} is defined in (4.1).
5. **Eliminate**: Set $J_{t+1} := J_t$ and do:
   a. For all $\theta \in J_{t+1}$ so that $G_{t+1}(\theta) = -\infty$, let $J_{t+1} := J_{t+1} \setminus \{\theta\}$.
   b. Find the most likely parameter in the plausible set $\hat\theta := \arg\max_{\theta\in J_{t+1}} G(\theta)$.
   c. For all $\theta \in J_{t+1}$ so that $G_{t+1}(\hat\theta) - G_{t+1}(\theta) > \log\left[\frac{3(|\Theta|-1)}{\delta}\right]$, let $J_{t+1} := J_{t+1} \setminus \{\theta\}$.

5 Proof of the Main Result

An outline of the proof of Theorem 5 is as follows. We begin in Section 5.1 by introducing an optimistic auxiliary model that will prove useful later on. In Section 5.2 we define *informative state-action pairs* (Definition 9) that are roughly state-action pairs that distinguish the true MDP and the auxiliary model. We next show in Theorem 10 that within a finite time interval following an ϵ-suboptimal time step (Definition 1), there is a positive probability to reach an informative state-action pair. Moreover, Theorem 12 (Section 5.3) implies that the number of ϵ-suboptimal steps encountered is bounded with high probability in terms of number of actual visits to informative state-action pairs. Hence, once we show that the number of visits to informative state-action pairs is bounded, we can conclude that the policy-mistake count is bounded as well. To show the former, we bound in Section 5.4 the stopping time of the SPRT test (for any fixed parameter $\theta \neq \theta_0$) using a non-decreasing measure of accumulated statistical information related to Bhattacharyya's information coefficient. In Section 5.5 we show that each visit to an informative state-action pair adds some strictly positive amount of information to one parameter at least. Hence the number of visits needed for SPRT to trigger is bounded. Using the pigeon-hole principle, we obtain that the number of visits to an informative state action pairs until convergence to an ϵ-optimal policy is also bounded, thus concluding the proof.

Note that from this point on all the probabilities and expectations refer to the stochastic process induced by the PEL algorithm on the actual MDP M_{θ_0}, unless mentioned otherwise.

5.1 An Auxiliary Model

Consider some *fixed* subset of parameters $J \subseteq \Theta$. For every $s \in S$, define the *optimistic parameter* in J as

$$\theta(J, s) = \arg\max_{\theta\in J} V^*_\theta(s)$$

(with ties decided arbitrarily). Define an auxiliary MDP $M_J = \langle S, A, R, p_J, \eta_J\rangle$, where $p_J(s'|s, a) = p_{\theta(J,s)}(s'|s, a)$ and $\eta_J(r'|s, a) = \eta_{\theta(J,s)}(r'|s, a)$. Further define the following stationary policy: $\pi_J(s) = \pi^*_{\theta(J,s)}(s)$. This policy picks at each state the optimal action according to the parameter $\theta(J, s)$ that is optimistic for that state. (In the context of the PEL algorithm, it is evident that as long as the set J_t is equal to J, the algorithm follows this stationary policy.) Denote the value function of the MDP M_J under the policy π_J as $V^{\pi_J}_J$. For notational convenience we the abbreviated notation V_J. Then the auxiliary model is optimistic in the following sense:

Lemma 6 *For any $s \in S$ and $\theta \in J$ the following inequality holds* [1] *:*

$$V_J(s) \geq V^*_\theta(s). \tag{5.1}$$

Proof: Let us consider the difference of the two value functions. Noting the definition of $\theta(J, s)$ and π_J, and substituting the corresponding Bellman backup (2.1) we have

$$\begin{aligned}
&V_J(s) - V^*_\theta(s) \geq V_J(s) - V^*_{\theta(J,s)}(s) \\
&= \gamma \sum_{s'\in S} p_J(s'|s, \pi_J(s)) \left[V_J(s') - V^*_{\theta(J,s)}(s')\right] \\
&\geq \gamma \sum_{s'\in S} p_J(s'|s, \pi_J(s)) \left[V_J(s') - V^*_{\theta(J,s')}(s')\right].
\end{aligned}$$

Repeating the argument n times we obtain that (with $s_0 \equiv s$),

$$\begin{aligned}
&V_J(s) - V^*_\theta(s) \\
&\geq \gamma^n \sum_{s_1,s_2,...,s_n\in S^n} \left(\prod_{i=1}^n p_J(s_i|s_{i-1}, \pi_J(s_{i-1}))\right) \\
&\cdot \left[V_J(s_n) - V^*_{\theta(J,s_n)}(s_n)\right] \\
&\geq -\frac{R_{max}}{1-\gamma}\gamma^n \sum_{s_1,s_2,...,s_n\in S^n} \left(\prod_{i=1}^n p_J(s_i|s_{i-1}, \pi_J(s_{i-1}))\right) \\
&\geq -\frac{R_{max}}{1-\gamma}\gamma^n \xrightarrow{n\to\infty} 0.
\end{aligned}$$

The second inequality follows since the value functions are positive and upper bounded by $\frac{R_{max}}{1-\gamma}$. The third inequality uses the fact that sum of probabilities over all possible histories is equal to 1. ■

[1] Note that the auxiliary model M_J need not be in the family $\{M_\theta\}_{\theta\in\Theta}$. Hence, it may even hold that $V_J(s) > V^*_\theta(s)$ for any $\theta \in \Theta$ and $s \in S$.

5.2 Implicit Explore or Exploit

We next establish that the PEL algorithm implicitly solves a tradeoff between possible exploration and exploitation. In other words, the agent either follows an ϵ-optimal policy or otherwise gains some properly defined information with some positive probability.

The proof is partially based on results from [SL05] and [KS02]. For a stationary policy π denote the discounted H-step value function by

$$V^{\pi}(s,H) \stackrel{\triangle}{=} \mathbb{E}^{\pi,s}\left\{\sum_{t=0}^{H-1}\gamma^{t}r_{t}\right\}.$$

The first lemma addresses the sensitivity of the value function to the time horizon.

Lemma 7 *If $H \geq \frac{1}{1-\gamma}\log\frac{R_{max}}{\epsilon(1-\gamma)}$ then $|V^{\pi}(s,H) - V^{\pi}(s)| \leq \epsilon$ for all policies π and states s.*

Proof: Trivial by bounding the tail of sum of rewards in the definition of the value function (see e.g. Lemma 2 in [KS02]). ■

In the following we use $T_{\text{eff}} = \frac{1}{1-\gamma}\log\frac{4R_{max}}{\epsilon(1-\gamma)}$ as an effective horizon length, beyond which the effect on the discounted return is negligible.

The following lemma bounds the sensitivity of the discounted reward function to perturbations in the transition and reward probabilities. For two probability distributions p and q on the finite set A, we use the l_1 norm to measure their separation:

$$\|p(\cdot) - q(\cdot)\|_1 = \sum_{a\in A}|p(a) - q(a)|\,. \tag{5.2}$$

Lemma 8 *Let $M_1 =< S, A, R, p_1, \eta_1 >$ and $M_2 =< S, A, R, p_2, \eta_2 >$ be two MDPs with non-negative rewards bounded by R_{max}. Let π be some stationary policy and let ϵ_1 be a positive number. If $\|\eta_1(\cdot|s,a) - \eta_2(\cdot|s,a)\|_1 \leq \frac{\epsilon_1(1-\gamma)^2}{R_{max}}$ and $\|p_1(\cdot|s,a) - p_2(\cdot|s,a)\|_1 \leq \frac{\epsilon_1(1-\gamma)^2}{R_{max}}$ for all states s and actions a, then*

$$\max_{s\in S}\left|V^{\pi}_{M_1}(s) - V^{\pi}_{M_2}(s)\right| \leq \epsilon_1.$$

Proof: Follows from Lemma 4 in [SL05], after noting that

$$|\bar{r}_1(s,a) - \bar{r}_2(s,a)| \leq R_{max}\,\|\eta_1(\cdot|s,a) - \eta_2(\cdot|s,a)\|_1\,.$$

■

To state the central result of this subsection, define informative state-action pairs as those pairs for which either the state transition or the reward distribution are distinct under the true and optimistic models. More precisely:

Definition 9 *Recall that θ_0 is the true parameter. Let $\theta(J,s)$ be defined as in Subsection 5.1. For $t \geq 0$, let K_t be the set of state-action pairs (s,a) for which*

$$\left\|\eta_{\theta(J_t,s)}(\cdot|s,a) - \eta_{\theta_0}(\cdot|s,a)\right\|_1 \leq \frac{\epsilon(1-\gamma)^2}{4R_{max}}, \text{ and}$$

$$\left\|p_{\theta(J_t,s)}(\cdot|s,a) - p_{\theta_0}(\cdot|s,a)\right\|_1 \leq \frac{\epsilon(1-\gamma)^2}{4R_{max}}.$$

*We say that the PEL algorithm visited an **informative state-action pair** at time t, if $(s_t,a_t) \notin K_t$.*

The following proposition asserts that occurrence of an ϵ-suboptimal step leads to an explorative interval, where an informative state-action pair is visited with probability of at least $\frac{\epsilon(1-\gamma)}{2R_{max}}$. Recalling the definition of an ϵ-suboptimal time step in Definition 1, let

$$E_1(t) \stackrel{\triangle}{=} \{\theta_0 \in J_t\} \cap \{V^{\mathcal{A}_t}(h_t) < V^{*}_{\theta_0}(s_t) - \epsilon\} \tag{5.3}$$

for $t \geq 0$ denote the "suboptimal" event that time step t is ϵ-suboptimal and the true parameter wasn't eliminated before time t. Let

$$E_2(t) \stackrel{\triangle}{=} \{(s_{t-1},a_{t-1}) \notin K_{t-1}\} \cup \{J_t \neq J_{t-1}\} \tag{5.4}$$

for $t \geq 1$ be the "informative" event that at time step $(t-1)$ either an informative state-action pair was visited or some parameter was eliminated from the set J_{t-1} of plausible parameters at time t. Denote by

$$E_3(t) \stackrel{\triangle}{=} \bigcup_{\tau=t+1}^{t+T_{\text{eff}}} E_2(\tau),\ t \geq 0$$

the event that the informative event $E_2(\tau)$ occurred for τ between $(t+1)$ and $(t+T_{\text{eff}})$. Let $\mathcal{F}_t \stackrel{\triangle}{=} \sigma\{h_t\}$ be the sigma algebra of the history sequence until time step t, then $E_1(t), E_2(t) \in \mathcal{F}_t$, while $E_3(t) \in \mathcal{F}_{t+T_{\text{eff}}}$.

Proposition 10 *For every $t \geq 0$ and t-history h_t that satisfies $E_1(t)$,*

$$\mathbb{P}^{\mathcal{A},s_0}\left\{E_3(t)|\,h_t\right\} > \frac{\epsilon(1-\gamma)}{2R_{max}}.$$

Proof: Let $J = J_t$ denote the set of plausible parameters at time t and let π_J and $\theta(J,s)$ be defined as in Subsection 5.1. Denote by $K = K_t$ the set of non-informative state-action pairs at time t. Then

$$\begin{aligned} V^{\mathcal{A}}_{\theta_0}(h_t) &\equiv \mathbb{E}^{\mathcal{A},s_0}\left\{\sum_{j=t}^{\infty}\gamma^{j-t}r_j\,\middle|\,h_t\right\} \\ &\geq \mathbb{E}^{\mathcal{A},s_0}\left\{\mathbb{I}\{E_3^c\}\sum_{j=t}^{t+T_{\text{eff}}-1}\gamma^{j-t}r_j\,\middle|\,h_t\right\}, \end{aligned}$$

where E_3^c denotes the event complementary to E_3. We wish to replace the policy $\mathcal{A}$ in the last expression with a stationary policy. For that purpose, define an auxiliary MDP M' which coincides with M_{θ_0} on $(s,a) \in K$ and with M_J (see Subsection 5.1) on $(s,a) \notin K$. Denote by $\mathbb{P}^{\pi_J}_{M'}\{\cdot|\,h_t\}$ the probability measure on sequence $(a_i,r_i,s_{i+1})_{i=t}^{\infty}$ induced by the policy π_J on M', with $\mathbb{E}^{\pi_J}_{M'}\{\cdot|\,h_t\}$ the corresponding expectation operator. The t-history h_t determines the sets J, G and K for this auxiliary process.

For any realization in $E_3^c(t)$, the set of plausible parameters J_τ is constant on the interval $\tau \in \{t,...,t+T_{\text{eff}}\}$, hence the PEL algorithm follows the stationary policy π_J on that interval. Moreover, over that interval the PEL algorithm visits only state-action pair in K, hence the measure under MDP

M_{θ_0} coincides with the measure under M' there. Therefore

$$\mathbb{E}^{\mathcal{A},s_0}\left\{\mathbb{I}\left\{E_3^c(t)\right\}\sum_{j=t}^{t+T_{\text{eff}}-1}\gamma^{j-t}r_j\middle|h_t\right\}$$
$$=\mathbb{E}^{\pi_J}_{M'}\left\{\mathbb{I}\left\{E_3^c(t)\right\}\sum_{j=t}^{t+T_{\text{eff}}-1}\gamma^{j-t}r_j\middle|h_t\right\}.$$

Substituting in the previous inequality we obtain:

$$\begin{aligned}V^{\mathcal{A}}_{\theta_0}(h_t) &\geq \mathbb{E}^{\pi_J}_{M'}\left\{\mathbb{I}\left\{E_3^c(t)\right\}\sum_{j=t}^{t+T_{\text{eff}}-1}\gamma^{j-t}r_j\middle|h_t\right\}\\ &= \mathbb{E}^{\pi_J}_{M'}\left\{\sum_{j=t}^{t+T_{\text{eff}}-1}\gamma^{j-t}r_j\middle|h_t\right\}\\ &\quad -\mathbb{E}^{\pi_J}_{M'}\left\{\mathbb{I}\left\{E_3(t)\right\}\sum_{j=t}^{t+T_{\text{eff}}-1}\gamma^{j-t}r_j\middle|h_t\right\}.\end{aligned}$$

The first term is a finite horizon value function $V^{\pi_J}_{M'}(s_t,T_{\text{eff}})$, while the sum in the second expectation can be bounded from above by $\frac{R_{max}}{1-\gamma}$. Hence,

$$V^{\mathcal{A}}_{\theta_0}(h_t)\geq V^{\pi_J}_{M'}(s_t,T_{\text{eff}})-\frac{R_{max}}{1-\gamma}\mathbb{P}^{\pi_J}_{M'}\left\{E_3(t)|\,h_t\right\}. \quad (5.5)$$

Now the first term satisfies, due to Lemma 7, 8 and 6,

$$V^{\pi_J}_{M'}(s_t,T_{\text{eff}})\geq V^{\pi_J}_{M'}(s_t)-\frac{\epsilon}{4}\geq V_J(s_t)-\frac{\epsilon}{2}\geq V^*_{\theta_0}(s_t)-\frac{\epsilon}{2}.$$

For the second term in (5.5), note that $\mathbb{P}^{\pi_J}_{M'}\left\{E_3^c(t)|\,h_t\right\}=\mathbb{P}^{\mathcal{A},s_0}\left\{E_3^c(t)|\,h_t\right\}$, hence $\mathbb{P}^{\pi_J}_{M'}\left\{E_3(t)|\,h_t\right\}=\mathbb{P}^{\mathcal{A},s_0}\left\{E_3(t)|\,h_t\right\}$. Thus,

$$V^{\mathcal{A}}_{\theta_0}(h_t)\geq V^*_{\theta_0}(s_t)-\frac{\epsilon}{2}-\frac{R_{max}}{1-\gamma}\mathbb{P}^{\mathcal{A},s_0}\left\{E_3(t)|\,h_t\right\}.$$

On the other hand, for h_t in $E_1(t)$ the time step t is ϵ-suboptimal, namely $V^{\mathcal{A}}_{\theta_0}(h_t)<V^*_{\theta_0}(s_t)-\epsilon$. Combined with the previous inequality we obtain

$$\mathbb{P}^{\mathcal{A},s_0}\left\{E_3(t)|\,h_t\right\}>\frac{(1-\gamma)\epsilon}{2R_{max}}.$$

■

5.3 Discovery Lemma

Proposition 10 shows that in the T_{eff} steps following an ϵ-suboptimal step there is a probability of at least $\frac{\epsilon(1-\gamma)}{2R_{max}}$ to reach some informative state-action pair or eliminate some parameter from J_t. Based on that, Lemma 12 below essentially bounds the number of ϵ-suboptimal steps in terms of the number of actual visits to informative state-action pairs and parameter eliminations.

The proof of this lemma is somewhat complicated by two facts. First, the events involved are not independent. Second, we need consider only those time instances over which the probability to reach an informative state-action pair exceeds some threshold. Indeed, applying a concentration inequality (such an Hoeffding's or Azuma's) to all time instances, including those where this probability is null or very small, would result in too week a bound. The proposed solution is to apply an appropriate concentration inequality over an appropriate subsequence of (stopping) times.

This argument was introduced by Bernstein in [Ber07] and proceeds through the following proposition.

Proposition 11 (Abstract Discovery Lemma) *Denote by $\{\mathcal{F}_t\}$ a given filtration (i.e., an increasing sequence of σ-algebras) and by $\{D_t\}$ a sequence of events with $D_t\in\mathcal{F}_t$. Let*

$$Z\triangleq\sum_{t=1}^{\infty}\mathbb{I}\left\{\mathbb{P}\left\{D_t|\,\mathcal{F}_{t-1}\right\}>p\right\},$$

where $p>0$ is some given constant. Further, suppose that

$$\mathbb{P}\left\{\sum_{t=1}^{\infty}\mathbb{I}\left\{D_t\right\}\leq M\right\}=1$$

for some integer $M>0$. Then, for $0<\delta<1$,

$$\mathbb{P}\left\{Z\leq\frac{2}{p}\left(M+\frac{4}{p}\log\frac{1}{\delta}\right)\right\}\geq 1-\delta.$$

Proof: See [Ber07]. This proof is repeated in the Appendix A for the benefit of the reader. ■

Let K_t be as in Definition 9 and let N_2 be a positive integer. Recall the definitions of $E_1(t)$, $E_2(t)$, $E_3(t)$ and $\mathcal{F}_t$ from the previous section.

Lemma 12 *For any positive integer N_2, let $T_2(N_2)$ be the time step on which the event $E_2(t)$ occurred for the N_2-th time, namely,*

$$T_2(N_2)=\inf\left\{n\middle|\sum_{k=1}^{n}\mathbb{I}\left\{E_2(k)\right\}=N_2\right\} \quad (5.6)$$

(with $T_2(N_2)=\infty$ is such n does not exist). Then, for all $\epsilon>0$ and $0<\delta<1$,

$$\mathbb{P}^{\mathcal{A},s_0}\left\{\sum_{k=0}^{T_2(N_2)}\mathbb{I}\left\{E_1(k)\right\}\leq N_1\right\}\geq 1-\delta,$$

where

$$N_1\triangleq T_{\mathit{eff}}\frac{4R_{max}}{\epsilon(1-\gamma)}\left[N_2+\frac{8R_{max}}{\epsilon(1-\gamma)}\log\frac{T_{\mathit{eff}}}{\delta_3}\right].$$

Proof: Define the following *discovery event* for $t\geq 0$:

$$D(t)\triangleq\{\theta_0\in J_t\}\bigcap\left\{\sum_{k=1}^{t}\mathbb{I}\left\{E_2(k)\right\}<N_2\right\}\bigcap E_3(t).$$

This event implies the following: by time step t the true parameter wasn't eliminated, and the informative event E_2 was encountered less than N_2 times; furthermore, in the following T_{eff} steps an least one additional event E_2 will occur. Note that $D(t)\in\mathcal{F}_{t+T_{\text{eff}}}$.

In order to employ Proposition 11 let us sample the series of events $D(t)$ and sigma-algebras $\mathcal{F}_t$ with the step of

T_{eff}, i.e., for $i = 0, 1, 2, ...$ and $j \in \{0, .., (T_{\text{eff}} - 1)\}$ denote $D_{i+1}^{(j)} = D(i \cdot T_{\text{eff}} + j)$ and $\mathcal{F}_i^{(j)} = \mathcal{F}_{i \cdot T_{\text{eff}}+j}$. Note that $D_i^{(j)} \in \mathcal{F}_i^{(j)}$. For j as above define

$$Z^{(j)} \triangleq \sum_{i=0}^{\infty} \mathbb{I}\left\{ \mathbb{P}^{\mathcal{A},s_0}\left\{ D_i^{(j)} \middle| \mathcal{F}_i^{(j)} \right\} > \frac{\epsilon(1-\gamma)}{2R_{max}} \right\},$$

and note that

$$\mathbb{P}^{\mathcal{A},s_0}\left\{ \sum_{i=1}^{\infty} \mathbb{I}\left\{ D_i^{(j)} \right\} \leq N_2 \right\} = 1$$

by definition of $D_i^{(j)}$. Noting the definition of N_1, application of Proposition 11 with $p = \frac{\epsilon(1-\gamma)}{2R_{max}}$ yields

$$\mathbb{P}^{\mathcal{A},s_0}\left\{ Z^{(j)} \leq \frac{N_1}{T_{\text{eff}}} \right\} \geq 1 - \frac{\delta_3}{T_{\text{eff}}}.$$

Applying the union bound we obtain

$$\mathbb{P}^{\mathcal{A},s_0}\left\{ \sum_{j=0}^{T_{\text{eff}}-1} Z^{(j)} \leq N_1 \right\} \geq 1 - \delta_3.$$

We conclude the proof by showing that

$$\sum_{t=0}^{T_2(N_2)} \mathbb{I}\{E_1(t)\} \leq \sum_{j=0}^{T_{\text{eff}}-1} Z^{(j)} \tag{5.7}$$

$$\equiv \sum_{t=0}^{T_2(N_2)} \mathbb{I}\left\{ \mathbb{P}^{\mathcal{A},s_0}\{ D_t | \mathcal{F}_t \} > \frac{\epsilon(1-\gamma)}{2R_{max}} \right\}.$$

For some $t < T_2(N_2)$ let h_t be a t-history that satisfies $E_1(t)$ (if such history exists). For this history, $\theta_0 \in J_t$ by definition of $E_1(t)$ and the inequality $\sum_{k=1}^{t} \mathbb{I}\{E_2(k)\} < N_2$ holds by definition of $T_2(N_2)$, hence the discovery event $D(t)$ occurs if and only if the event $E_3(t)$ occurs. Then, by Proposition 10,

$$\mathbb{P}^{\mathcal{A},s_0}\{ D_t | h_t \} > \frac{\epsilon(1-\gamma)}{2R_{max}},$$

therefore $\mathbb{I}\left\{ P^{\mathcal{A},s_0}\{D_t|\mathcal{F}_t\} > \frac{\epsilon(1-\gamma)}{2R_{max}} \right\} \geq \mathbb{I}\{E_1(t)\}$ almost surely. Hence 5.7 is established an the claim follows. ■

5.4 Sequential Hypothesis Testing

The sequential hypothesis test we use in our algorithm was originated by Wald ([Wal52]) and is defined in the following way. Consider a discrete-time stochastic process $\{x_t\}_{t=0}^{\infty}$ taking values in a finite set S. Denote by $x_0^n = \{x_0, ..., x_n\}$ the observations obtained by time n. Let the probability of such observations under hypothesis H_0 be denoted as $p_0(x_0^n)$, and under H_1 as $p_1(x_0^n)$. Note that the discussion here is not limited to Markov processes.

Definition 13 *For any $0 < \delta < 1$ define the stopping time*

$$N^W(\delta) = \inf_n \left\{ n \left| \frac{p_1(x_0^n)}{p_0(x_0^n)} \geq \frac{1}{\delta} \text{ or } \frac{p_1(x_0^n)}{p_0(x_0^n)} \leq \delta \right. \right\},$$

and the decision rule

$$d^W(\delta) = \begin{cases} H_1 & \text{, when } \left.\frac{p_1(x_0^n)}{p_0(x_0^n)}\right|_{n=N^W(\delta)} \geq \frac{1}{\delta} \\ H_0 & \text{, otherwise} \end{cases}.$$

The pair $(N^W(\delta_1), d^W(\delta_1))$ is the Sequential Probability Ratio Test (SPRT).

It was shown by Wald ([Wal52]) that the error probability of the SPRT is bounded by δ:

Theorem 14 (Wald) $\mathbb{P}\left\{ d^W(\delta) = H_0 \middle| H_1 \right\} \leq \delta$ *and* $\mathbb{P}\left\{ d^W(\delta) = H_1 \middle| H_0 \right\} \leq \delta$.

We next establish a useful bound on the stopping time of SPRT, using an auxiliary stopping time for the same process based on the Bhattacharyya coefficient rather than the likelihood ratio. We begin by defining the Bhattacharyya coefficient [Kai67].

Definition 15 (Bhattacharyya coefficient) *Let p and q be probability distributions on the finite set S. Then the Bhattacharyya coefficient is*

$$\rho \triangleq \sum_{s' \in S} p^{1/2}(s') q^{1/2}(s').$$

Note that $\rho \leq 1$ by the Cauchy-Schwarz inequality. The Bhattacharyya distance (or information) is defined as $-\log \rho$. This metric is related to the l_1-norm of $(p - q)$ in the following way:

Lemma 16 $-\log \rho \geq \frac{1}{8} \|p - q\|_1^2$

Proof: Kraft [Kra55] showed the following relation: $\frac{1}{2} \|p - q\|_1 \leq \sqrt{1 - \rho^2}$. Equivalently, $\rho \leq \sqrt{1 - \frac{1}{4} \|p - q\|_1^2}$, hence $\log \rho \leq \frac{1}{2} \log(1 - \frac{1}{4} \|p - q\|_1^2) \leq -\frac{1}{8} \|p - q\|_1^2$, where the last inequality follows since $\log(1 - x) \leq -x$ for all $0 \leq x < 1$. ■

Definition 17 (Bhattacharyya stopping time) *Consider the same processes and hypotheses as in Definition 13. Denote the by*

$$\rho(x_0^n) = \sum_{x_{n+1} \in S} p_0^{1/2}(x_{n+1}|x_0^n) p_1^{1/2}(x_{n+1}|x_0^n)$$

the Bhattacharyya coefficient between $p_0(\cdot|x_0^n)$ and $p_1(\cdot|x_0^n)$. Then the Bhattacharyya stopping time (for $0 < \delta < 1$) is defined as:

$$N^B(\delta) = \inf_n \left\{ n \left| \prod_{t=0}^{n-1} \rho(x_0^t) \leq \delta \text{ or } \right. \right. \tag{5.8}$$

$$\left. \left. p_0(x_n|x_0^{n-1}) = 0 \text{ or } p_1(x_n|x_0^{n-1}) = 0 \right\} \right.$$

We note that the stopping condition $\prod_{t=0}^{n-1} \rho(x_0^t) \leq \delta$ can be written as

$$R_n \triangleq -\sum_{t=0}^{n-1} \log \rho(x_0^t) \geq -\log \delta,$$

where R_n is the cumulative Bhattacharyya distance (or total Bhattacharyya information).

While our algorithm uses the Wald test, the Bhattacharyya stopping time will be more handy for analysis as R_n is a non-decreasing sequence. The following proposition relates these two stopping times.

Proposition 18 *For $0 < \delta < 1$, the inequality*

$$\mathbb{P}\left\{N^W(\delta) > N^B(\delta^{3/2})\right\} \leq \delta$$

holds both under H_0 and H_1.

Proof: Assume that H_0 holds true (the proof is identical under H_1). Since $p_1(x_n|x_0^{n-1}) = 0$ implies $N^W = N^B(\delta^{3/2})$ (if not stopped before), we can focus in the remainder of the proof only on stopping due to the first condition in (5.8). Let $N_1 = N^B(\delta^{3/2})$ and denote the log likelihood ratio of the history up to the stopping time N_1 as:

$$L\left(x_0^{N_1}\right) \triangleq \sum_{t=1}^{N_1} \log \frac{p_1(x_t|x_0^{t-1})}{p_0(x_t|x_0^{t-1})}.$$

Then $N^W(\delta) > N_1$ implies that $L\left(x_0^{N_1}\right) > \log\delta$, hence

$$\mathbb{P}\left\{N^W(\delta) > N_1\right\} \quad \leq \quad \mathbb{P}\left\{L\left(x_0^{N_1}\right) > \log\delta,\ N_1 < \infty\right\}.$$

Chernoff's inequality now implies

$$\begin{aligned}&\mathbb{P}\left\{L\left(x_0^{N_1}\right) > \log\delta,\ N_1 < \infty\right\}\\ &\leq \mathbb{E}\left\{\exp\left\{\frac{1}{2}\left[L\left(x_0^{N_1}\right) - \log\delta\right]\right\}\mathbb{I}_{\{N_1<\infty\}}\right\},\end{aligned}$$

hence

$$\mathbb{P}\left\{N^W(\delta) > N_1\right\} \leq \frac{1}{\sqrt{\delta}} E_C, \tag{5.9}$$

where

$$E_C \triangleq \mathbb{E}\left\{\exp\left\{\frac{1}{2}L\left(x_0^{N_1}\right)\right\}\mathbb{I}_{\{N_1<\infty\}}\right\}.$$

We proceed to bound E_C. Denote

$$d(x_{t+1}|x_0^t) \triangleq p_1^{1/2}(x_{t+1}|x_0^t)p_0^{1/2}(x_{t+1}|x_0^t)$$

so that $\rho(x_0^t) = \sum_{x'\in S} d(x'|x_0^t)$. Further denote

$$D(x_0^k) \triangleq \prod_{t=1}^{k} d(x_t|x_0^{t-1}).$$

Let Q_B be the collection of N_1-histories $x_0^{N_1}$ for which $N_1 < \infty$, namely

$$Q_B = \left\{x_0^k \in S^{k+1} \left| \prod_{i=0}^{k-2}\rho(x_0^i) > \delta^{3/2} \text{ and } \prod_{i=0}^{k-1}\rho(x_0^i) \leq \delta^{3/2}\right.\right\}.$$

Substituting the definition of the expected value we obtain:

$$\begin{aligned}&E_C\\ &= \sum_{x_0^{N_1}\in Q_B} \exp\left\{\frac{1}{2}\sum_{t=1}^{N_1}\log\frac{p_1(x_t|x_0^{t-1})}{p_0(x_t|x_0^{t-1})}\right\}\prod_{t=1}^{N_1}p_0(x_t|x_0^{t-1})\\ &= \sum_{x_0^{N_1}\in Q_B}\prod_{t=1}^{N_1}\left(\frac{p_1(x_t|x_0^{t-1})}{p_0(x_t|x_0^{t-1})}\right)^{1/2}p_0(x_t|x_0^{t-1})\\ &= \sum_{x_0^{N_1}\in Q_B}\prod_{t=1}^{N_1}p_1^{1/2}(x_t|x_0^{t-1})p_0^{1/2}(x_t|x_0^{t-1})\\ &= \sum_{x_0^{N_1}\in Q_B} D(x_0^{N_1}).\end{aligned}$$

Below we show that

$$E_C \equiv \sum_{x_0^{N_1}\in Q_B} D(x_0^{N_1}) \leq \sup_{x_0^{N_1}\in Q_B}\left\{\prod_{t=0}^{N_1-1}\rho(x_0^t)\right\} \leq \delta^{3/2}, \tag{5.10}$$

where the last inequality holds by definition of N_1. Thus, from (5.9) and (5.10),

$$\mathbb{P}\left\{N^W(\delta) > N_1\right\} \leq \frac{1}{\sqrt{\delta}}E_C \leq \delta.$$

Fix an integer $M \geq 1$. With some abuse of notation define $Q_B(k)$ for $k \in \{1, .., M\}$ be the collection of k-histories x_0^k for which $N_1 = k$. We proceed to bound that

$$\sum_{k=1}^{M}\sum_{x_0^k\in Q_B(k)} D(x_0^k) \leq \max_{x_0^{N_1}\in\cup_{k=1}^{M}Q_B(k)}\left\{\prod_{t=0}^{N_1-1}\rho(x_0^t)\right\},$$

hence taking $M \to \infty$ establishes 5.10.

It is handy to artificially extend trajectories in $\bigcup_{k=1}^{M-1} Q_B(k)$ to the length of $(M+1)$. Modify the trajectory $\underline{x} \in Q_B(k)$ for $k < M$ to the trajectory x_0^M by duplicating the state $(\underline{x}_k)$ enough times, i.e., $x_0^M = \{\underline{x}_0, \underline{x}_1, ..., \underline{x}_k, \underline{x}_k, ..., \underline{x}_k\}$. For the modified states $(x_i,\ i \geq k+1)$ redefine $d(x_i|x_0^{i-1}) = \mathbb{I}\{x_i = x_k\}$ so that $D(x_0^k) = D(x_0^M)$. Note that, $\rho(x_0^i) = 1$.

Denote the collection of modified histories by σ_0^M. For $k \leq M$ let

$$\sigma_0^k \triangleq \{x_0^k \in S^{k+1} | \exists y \in \sigma_0^M \text{ s.t. } y_0 = x_0, ..., y_k = x_k\} \tag{5.11}$$

denote the set of the possible first $(k+1)$ states of trajectories from σ_M. For $x_0^{k-1} \in \sigma_0^{k-1}$ denote by

$$\sigma_k(x_0^{k-1}) \triangleq \{x \in S | [x_0^{k-1}; x] \in \sigma_0^k\}$$

the set of states that extend x_0^{k-1} to the sequence in σ_k, where $[\cdot;\cdot]$ denotes concatenation. Denote by

$$\sigma_k^M(x_0^{k-1}) \triangleq \{y_0^M \in \sigma_0^M | y_0 = x_0, ..., y_{k-1} = x_{k-1}\}$$

the collection of all M-histories in σ_0^M equal to x_0^{k-1} at the first $(k-1)$ steps. For $x_0^{k-1} \in \sigma_0^k$ $(k < M)$ denote

$$B(x_0^{k-1}) \triangleq \max_{x_0^M\in\sigma_k^M(x_0^{k-1})}\prod_{t=k}^{M-1}\rho(x_0^t),$$

and set $B(x_0^{M-1})=1$ for $x_0^{M-1}\in\sigma_0^{M-1}$. Note that

$$B(x_0^{k-2})=\max_{x_{k-1}\in\sigma_{k-1}(x_0^{k-2})}\rho([x_0^{k-2};x_{k-1}])B([x_0^{k-2};x_{k-1}]).$$

We next proof by induction that for $0\le k\le M-1$ the following inequality hold:

$$\sum_{x_0^M\in\sigma_0^M}D(x_0^M)\le\sum_{x_0^k\in\sigma_0^k}D(x_0^k)\rho(x_0^k)B(x_0^k).$$

In particular for $k=0$ we obtain

$$\sum_{x_0^M\in\sigma_0^M}D(x_0^M)\le\rho(x_0^0)B(x_0^0)\le\max_{x_0^M\in\sigma_0^M}\left\{\prod_{t=0}^{M-1}\rho(x_0^t)\right\},$$

hence concluding the proof. Let us show the basis for the induction ($k=M-1$)

$$\begin{aligned}&\sum_{x_0^M\in\sigma_0^M}D(x_0^M)\\&=\sum_{x_0^{M-1}\in\sigma_0^{M-1}}D(x_0^{M-1})\sum_{x_M\in\sigma_M(x_0^{M-1})}d(x_M|x_0^{M-1})\\&\le\sum_{x_0^{M-1}\in\sigma_0^{M-1}}D(x_0^{M-1})\rho(x_0^{M-1})B(x_0^{M-1}).\end{aligned}$$

Using the induction hypothesis for $k\ge i+1$ we obtain

$$\begin{aligned}&\sum_{x_0^M\in\sigma_0^M}D(x_0^M)\\&\le\sum_{x_0^{i+1}\in\sigma_0^{i+1}}D(x_0^{i+1})\rho(x_0^{i+1})B(x_0^{i+1})\\&\le\sum_{x_0^i\in\sigma_0^i}D(x_0^i)\sum_{x_{i+1}\in\sigma_{i+1}(x_0^i)}d(x_{i+1}|x_0^i)\\&\quad\cdot\rho([x_0^i;x_{i+1}])B([x_0^i;x_{i+1}])\\&\le\sum_{x_0^i\in\sigma_0^i}D(x_0^i)\sum_{x_{i+1}\in\sigma_{i+1}(x_0^i)}d(x_{i+1}|x_0^i)B(x_0^i)\\&\le\sum_{x_0^i\in\sigma_0^i}D(x_0^i)\rho(x_0^i)B(x_0^i),\end{aligned}$$

hence the induction step holds. ■

5.5 Proof of the Main Result

This subsection builds on our previous results to establish the upper bound on the policy-mistake count (Theorem 5). Consider the PEL algorithm applied to the true MDP M_{θ_0}. The proof proceeds through the following steps. In steps 1-3 we define three "unwanted" events: the event E_4 on which the true parameter θ_0 is eliminated from the plausible parameter set J_t at some point; the event E_5 on which (essentially) there is insufficient number of visits to informative state-action pairs despite a large number of "sub-optimal" steps; and the event E_6 on which a sufficient amount of Bhattacharyya information does not lead to parameter elimination in the SPRT test. We show that the probability of each is bounded by $\frac{\delta}{3}$. In step 4 and step 5 the required upper bound on the PMC is shown to hold on $(E_4\cup E_5\cup E_6)^c$. In step 6 we combine the above to conclude the required result.

Step 1: Let $E_4\triangleq\{\theta_0\notin\cap_{t=1}^{\infty}J_t\}$ be the event that the actual parameter is eliminated from the set J_t of plausible parameters at some point. As explained in Section 4, the elimination step of the algorithm can be interpreted as a SPRT between any pair of parameter in J_t, with the threshold of $\delta'\triangleq\frac{\delta}{3(|\Theta|-1)}$. From Theorem 14 we obtain that the probability of eliminating θ_0 due to any other fixed parameter is less than δ'. Therefore, by union bound the total probability of eliminating θ_0 is less then $(|\Theta|-1)\delta'$, namely,

$$\mathbb{P}^{\mathcal{A},s_0}\{E_4\}\le(|\Theta|-1)\delta'=\frac{\delta}{3}.$$

Step 2: Recall the definition of $E_1(t)$ and T_2 from (5.3) and (5.6). Let $E_5\triangleq\left\{\sum_{t=1}^{T_2(N_2)}\mathbb{I}\{E_1(t)\}>N_1\right\}$ be the event that the sub-optimal event $E_1(t)$ was encountered more than N_1 times before the N_2-th informative event occurred. Here,

$$N_2\triangleq 12(|\Theta|-1)\left(\frac{4R_{max}}{\epsilon(1-\gamma)^2}\right)^2\log(\frac{3(|\Theta|-1)}{\delta})+(|\Theta|-1)$$

(this selection is explained in step 4) and N_1 is selected as in Lemma 12 with $\delta:=\frac{\delta}{3}$, namely,

$$N_1\triangleq\frac{4R_{max}T_{\text{eff}}}{\epsilon(1-\gamma)}\left[N_2+\frac{8R_{max}}{\epsilon(1-\gamma)}\log\frac{3T_{\text{eff}}}{\delta}\right].$$

Then, Lemma 12 implies (for any N_2 and in particular for the one above),

$$\mathbb{P}^{\mathcal{A},s_0}\{E_5\}\le\frac{\delta}{3}.$$

Step 3: Consider hypothesis testing between MDPs M_{θ_0} and M_θ for $\theta\neq\theta_0$. Denote by $N^W(\theta,\delta)$, $R_n(\theta)$ and $N^B(\theta,\delta)$ the corresponding SPRT stopping time, the total Bhattacharyya information and the Bhattacharyya stopping time (see Definitions 13 and 17). Let E_6 be the event on which $N^W(\theta,\delta')>N^B\left(\theta,(\delta')^{3/2}\right)$ holds for some $\theta\neq\theta_0$ (i.e., the relation between Bhattacharyya stopping time and SPRT stopping time defined in Lemma 18 is violated). Using Lemma 18 and the union bound we conclude that

$$\mathbb{P}^{\mathcal{A},s_0}\{E_6\}\le(|\Theta|-1)\delta'=\frac{\delta}{3}.$$

Step 4: Consider a realization $h_\infty=\{s_t,a_t,r_t\}_{t=0}^{\infty}\in E_4^c\cap E_5^c\cap E_6^c$. Recall the definition of the informative event $E_2(t)$ in (5.4). We proceed to show that for this realization,

$$\sum_{t=1}^{\infty}\mathbb{I}\{E_2(t)\}\le N_2. \tag{5.12}$$

Let t be a time step on which an informative state-action pair (s_t,a_t) is visited (see Definition 9). Let us assess the Bhattacharyya distance $-\log\rho_t$ between the joint distribution of (r_t,s_{t+1}) under the true model M_{θ_0} and the auxiliary model M_J. Evidently, it equals to sum of Bhattacharyya distances between $\eta_{\theta_0}(\cdot|s_t,a_t)$ and $\eta_{\theta(t)}(\cdot|s_t,a_t)$, and between $p_{\theta_0}(\cdot|s_t,a_t)$ and $p_{\theta(t)}(\cdot|s_t,a_t)$, where $\theta(t)$ is the optimistic

parameter at time t (see Algorithm 1), namely

$$-\log\rho_t = -\log\left[\sum_{s'\in S} p^{1/2}_{\theta(t}(s'|s_t,a_t)p^{1/2}_{\theta_0}(s'|s_t,a_t)\right] - \log\left[\sum_{r'\in S}\eta^{1/2}_{\theta(t)}(r'|s_t,a_t)\eta^{1/2}_{\theta_0}(r'|s_t,a_t)\right].$$

Since $(s_t,a_t)\notin K_t$, then, by Lemma 16,

$$-\log\rho_t > \frac{1}{8}\left(\frac{\epsilon(1-\gamma)^2}{4R_{max}}\right)^2 \triangleq B_0.$$

Hence each visit to an informative state-action pair $(s_t,a_t)\notin K_t$ increases $R_t(\theta)$ by at least B_0 for at least one $\theta\in J_t$. As the sequence $R_t(\theta)$ is non-decreasing, the total number of such increments until the stopping time $N^B(\theta)$ triggers is upper bounded by $\frac{\log((1/\delta')^{3/2})}{B_0}$. By definition, for $h_\infty\in E_6^c$ the parameter θ is eliminated no later than $t=N^B\left(\theta,(\delta')^{3/2}\right)$, therefore, by the pigeon-hole principle, the total number of visits to informative state-action pairs until all $\theta\neq\theta_0$ are eliminated from J_t is bounded by $(|\Theta|-1)\frac{\log((1/\delta')^{3/2})}{B_0}$. Recall that $E_2(t)$ occurs if an informative state-action pair was visited at time $(t-1)$ or a parameter was eliminated from J_{t-1}. Hence,

$$\begin{aligned}&\sum_{t=1}^{\infty}\mathbb{I}\{E_2(t)\}\\ &\leq\sum_{t=1}^{\infty}\mathbb{I}\{(s_{t-1},a_{t-1})\notin K_{t-1}\}+\sum_{t=1}^{\infty}\mathbb{I}\{J_t\neq J_{t-1}\}\\ &\leq(|\Theta|-1)\frac{\log((1/\delta')^{3/2})}{B_0}+(|\Theta|-1)\equiv N_2.\end{aligned}$$

thus establishing (5.12).

Step 5: Let T_2, N_2 be as in Step 2. For h_∞ as before we argue that $PMC(\epsilon)\leq N_1$. Since $h_\infty\in E_5^c$,

$$\begin{aligned}N_1 &\geq \sum_{t=0}^{T_2(N_2)}\mathbb{I}\{E_1(t)\}\\ &= \left[\sum_{t=0}^{\infty}\mathbb{I}\{E_1(t)\}\right]\mathbb{I}\{T_2(N_2)=\infty\}\\ &\quad+\left[\sum_{t=0}^{\infty}\mathbb{I}\{E_1(t)\}\right]\mathbb{I}\{T_2(N_2)<\infty\}\\ &\quad-\left[\sum_{t=T_2(N_2)+1}^{\infty}\mathbb{I}\{E_1(t)\}\right]\mathbb{I}\{T_2(N_2)<\infty\}.\end{aligned}$$

Note that the argument in Step 4 implies, that for $t>T_2(N_2)$ the set J_t of plausible parameters contains only the true parameter θ_0. For this realization the PEL algorithm follows an optimal policy π_{θ_0} from time $T_2(N_2)$ onward, hence $\sum_{t=T_2(N_2)+1}^{\infty}\mathbb{I}\{E_1(t)\}=0$. Therefore,

$$N_1 \geq \sum_{t=0}^{\infty}\mathbb{I}\{E_1(t)\}=\sum_{t=0}^{\infty}\mathbb{I}\left\{V^{\mathcal{A}_t}(h_t)<V^*_{\theta_0}(s_t)-\epsilon\right\},$$

where equality holds since $\theta_0\in J_t$ for realization in E_4^c (see 5.4). Hence, by definition of PMC,

$$N_1\geq PMC(\epsilon). \tag{5.13}$$

Step 6: The bound (5.13) holds on $h_\infty\in E_4^c\cap E_5^c\cap E_6^c$. But, by the union bound,

$$\mathbb{P}^{\mathcal{A},s_0}\{E_4^c\cap E_5^c\cap E_6^c\}\geq 1-\delta.$$

Substituting N_2 and T_{eff} yields that

$$\begin{aligned}&PMC(\epsilon)\leq \\ &768(|\Theta|-1)\frac{R_{max}^3}{\epsilon^3(1-\gamma)^6}\log\left(\frac{3(|\Theta|-1)}{\delta}\right)\log\frac{4R_{max}}{\epsilon(1-\gamma)}\\ &+(|\Theta|-1)\frac{4R_{max}}{\epsilon(1-\gamma)^2}\log\frac{4R_{max}}{\epsilon(1-\gamma)}\\ &+\frac{32R_{max}^2}{\epsilon^2(1-\gamma)^3}\log\left[\frac{3}{\delta(1-\gamma)}\log\frac{4R_{max}}{\epsilon(1-\gamma)}\right]\log\frac{4R_{max}}{\epsilon(1-\gamma)}\end{aligned} \tag{5.14}$$

with probability of at least $(1-\delta)$. Noting that the first term is the dominant one, the can be simplified to (4.2). ■

6 Conclusion

Parameterized models offer a great potential for reduction of learning time and cost in large RL problems, alongside less structured methods such as function approximation, aggregation and state abstraction. The former can and should be used when the available prior information allows to reduce model uncertainty to a lower dimensional parameter space, thereby allowing explicit modeling of inter-state dependencies and avoiding the pitfalls inherent in the local nature of learning in the general, unstructured model. The development of effective RL methods for parameterized models should therefore be of major interest.

In this paper we have considered the case of parameterized models with with discrete parameter. We proposed the PEL (Parameter Elimination) learning algorithm, which incorporates efficient exploration to achieve polynomial mistake bounds in a PAC sense. As may be expected these bounds are independent of the cardinality of the state and action spaces, and in fact may well apply to continuous spaces under reasonable regularity conditions.

Several nontrivial choices were made in the construction of this algorithm. First, the basic approach taken was that of parameter elimination, rather than on-line parameter estimation. The former has the advantage of reducing the considered parameter set over time, which can quickly converge to a small set if sufficient statistical information is obtained. On the theoretical side, this approach allows the application of sequential hypotheses testing and its related theory for the analysis of the algorithm. On the downside, the possible error of eliminating the true parameter cannot be rectified later, and it is therefore important to keep its probability small. Another choice made in the algorithm is to incorporate an optimistic policy which is defined on a per-state basis, rather than freeze a stationary that is optimal for a certain parameter from a certain state. We believe this approach may add to exploration efficiency, although no direct comparison is available. Further work of immediate interest includes the extension of the PEL algorithm to continuous parameter spaces

though discretization, the consideration of other (estimation-based) algorithms that may be appropriate for such spaces, the incorporation of computational constraints, and consideration of other learning criteria such as the total regret for the average reward problem.

References

[Azu67] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematic Journal*, (19):357–367, 1967.

[Ber07] A. Bernstein. Adaptive state aggregation for reinforcement learning. Master's thesis, Technion - Israel Institute of Technology, 2007.

[BT02] R.I. Brafman and M. Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

[GS92] G. R. Grimmet and D. R. Stirzaker. *Probability and Random Processes*. Oxford University Press, 1992.

[Kai67] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE Transactions of Communication Technology*, com-15(1):52–60, 1967.

[Kak03] S.M. Kakade. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, University College London, 2003.

[KB82] P.R. Kumar and A. Becker. A new family of optimal adaptive controllers for markov chains. *IEEE Trans. Automat. Contr.*, AC-27:137–145, 1982.

[KK99] Michael J. Kearns and Daphne Koller. Efficient reinforcement learning in factored MDPs. In *IJCAI*, pages 740–747, 1999.

[Kra55] C.H. Kraft. Some conditions for consistency and uniform consistency of statistical procedures. *University of California Publications in Statistics*, 1955.

[KS02] M.J. Kearns and S.P. Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49:209–232, 2002.

[KV98] P.R. Kumar and P. Varaiya. *Stochastic Systems: Estimation, Identification and Adaptive Control*. The MIT Press, 1998.

[Man74] P. Mandl. Estimation and control in Markov chains. *Advanced Applied Probability*, 6:40–60, 1974.

[Put94] M.L. Puterman. *Markov Decision Processes. Discrete Stochastic Programming*. Wiley, 1994.

[RW00] L. C. G. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales*. Cambridge University Press, 2000.

[SB98] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 1998.

[SL05] A.L. Strehl and M.L. Littman. A theoretical analysis of model-based interval estimation. *Proceedings of the Twenty-second International Conference on Machine Learning (ICML-05)*, page 857864, 2005.

[SLW+06] A.L. Strehl, L. Li, E. Wiewiora, J. Langford, and M.L. Littman. PAC model-free reinforcement learning. *Proceedings of the* 23^{rd} *International Conference on Machine Learning*, 2006.

[Wal52] A. Wald. *Sequential Analysis*. Wiley, 1952.

A Appendix: An Abstract Discovery Lemma

We are about to prove here the Proposition 11. In words, this theorem tells us that number of time steps with sufficient probability for event D_t to occur (that is, steps having $\mathbb{P}\{D_t|\mathcal{F}_{t-1}\} > p$) is polynomially bounded with high probability, provided that the total number of such occurrences is bounded with probability one.

In order to prove this theorem, we use the following result regarding martingale differences and stopping times.

Lemma 19 *Let $\{M_t\}$ be a martingale difference sequence, adapted to the filtration $\{\mathcal{F}_t\}$, i.e.,*[2]

$$M_t \in \mathcal{F}_t,\ \mathbb{E}\{M_t|\mathcal{F}_{t-1}\} = 0 \text{ (a.s.)},\ \forall t \geq 1.$$

Assume that the underlying martingale process $X_t = \sum_{k=1}^{t} M_k$ is uniformly integrable, that is

$$\forall \epsilon > 0, \exists \lambda > 0 : \mathbb{E}\{|X_t|\mathbb{I}\{|X_t| > \lambda\}\} < \epsilon,\ \forall t \geq 1.$$

Also, let $\{\tau_i\}$ be a sequence of stopping times with respect to $\{\mathcal{F}_t\}$, having $\tau_{i-1} < \tau_i$ a.s. for any i. Assume moreover that $\{\tau_i = t\} \in \mathcal{F}_{t-1}$. Then, $\{M'_i\}_{i=1}^{\infty} \triangleq \{M_{\tau_i}\}_{i=1}^{\infty}$ is a martingale difference sequence, adapted to the filtration $\{\mathcal{F}'_i\}_{i=1}^{\infty} \triangleq \{\mathcal{F}_{\tau_{i+1}-1}\}_{i=1}^{\infty}$.

Proof: First, we see that $M'_i = M_{\tau_i} \in \mathcal{F}_{\tau_i} \subseteq \mathcal{F}_{\tau_{i+1}-1} = \mathcal{F}'_i$ (see [GS92]), for the definition of $\mathcal{F}_T$, where T is a stopping time. Next, we have that

$$\begin{aligned}\mathbb{E}\{M'_i\,\mathcal{F}|'_{i-1}\} &= \mathbb{E}\{M_{\tau_i}\,\mathcal{F}|_{\tau_i-1}\} \\ &= \mathbb{E}\{X_{\tau_i}\,\mathcal{F}|_{\tau_i-1}\} - \mathbb{E}\{X_{\tau_i-1}\,\mathcal{F}|_{\tau_i-1}\} \\ &= X_{\tau_i-1} - X_{\tau_i-1} = 0.\end{aligned}$$

The last equality follows since $X_{\tau_i-1} \in \mathcal{F}_{\tau_i-1}$. Also, it is easy to see that $\tau'_i = \tau_i - 1$ is a stopping time with respect to $\{\mathcal{F}_t\}$, since by hypothesis, $\{\tau'_i = t\} \in \mathcal{F}_t$. Thus, by Optional Sampling Theorem for uniformly integrable martingales (see [RW00], Theorem II.77.5), we have that $\mathbb{E}\left\{X_{\tau_i}\,\mathcal{F}|_{\tau'_i}\right\} = X_{\tau'_i}$, almost surely. ■

In the following lemma we define a martingale that will be used in the proof of Discovery Theorem, and prove that it is uniformly integrable.

Lemma 20 *The following process*

$$X_t \triangleq \sum_{k=1}^{t} \left(\mathbb{I}\{D_k\} - \mathbb{P}\{D_k|\mathcal{F}_{k-1}\}\right),\ t > 0$$

($X_0 \triangleq 0$) is a uniformly integrable martingale adapted to $\{\mathcal{F}_t\}$.

Proof: The fact that $\{X_t\}$ is a martingale adapted to $\{\mathcal{F}_t\}$ follows trivially by definition. To prove uniform integrability, we will use the fact that if there exists a positive random variable Y with $\mathbb{E}\{Y\} < \infty$, such that $|X_t| \leq Y$, $\forall t$, a.s. then the family $\{X_t\}$ if uniformly integrable.

[2] We will use the relation $X \in \mathcal{F}$ to denote the fact that random variable X is measurable with respect to $\mathcal{F}$.

Indeed, for all t, we have that

$$\begin{aligned}|X_t| &\leq \sum_{k=1}^{t} \mathbb{I}\{D_k\} + \sum_{k=1}^{t} \mathbb{P}\{D_k|\mathcal{F}_{k-1}\} \\ &\leq M + \sum_{k=1}^{t} \mathbb{P}\{D_k|\mathcal{F}_{k-1}\}, \qquad \text{(A.1)}\end{aligned}$$

where the second inequality follows by the hypothesis that $\sum_{k=1}^{\infty} \mathbb{I}\{D_k\} \leq M$ with probability one. Now define $Y_t \triangleq \sum_{k=1}^{t} \mathbb{P}\{D_k|\mathcal{F}_{k-1}\}$. First, the expected value of Y_t is bounded by M:

$$\begin{aligned}\mathbb{E}\{Y_t\} &= \mathbb{E}\left\{\sum_{k=1}^{t} \mathbb{P}\{D_k|\mathcal{F}_{k-1}\}\right\} \\ &= \sum_{k=1}^{t} \mathbb{E}\{\mathbb{E}\{\mathbb{I}\{D_k\}|\mathcal{F}_{k-1}\}\} \\ &= \sum_{k=1}^{t} \mathbb{E}\{\mathbb{I}\{D_k\}\} \qquad \text{(A.2)} \\ &= \mathbb{E}\left\{\sum_{k=1}^{t} \mathbb{I}\{D_k\}\right\} \leq M,\ \forall t. \qquad \text{(A.3)}\end{aligned}$$

Also,

$$Y_t(\omega) \leq Y_{t+1}(\omega),\ \forall t, \omega \qquad \text{(A.4)}$$

implying that

$$\mathbb{E}\{Y\}_t \leq \mathbb{E}\{Y\}_{t+1},\ \forall t. \qquad \text{(A.5)}$$

Thus, by (A.2), (A.5), and the monotone convergence of the sequence $a_t \triangleq \mathbb{E}\{Y\}_t$, we know that there exists $a_\infty \triangleq \lim_{t\to\infty} \mathbb{E}\{Y\}_t < \infty$. This, (A.4), and the monotone convergence theorem for Y_t, imply that there exists a random variable $Y_\infty = \lim_{t\to\infty} Y_t$, such that

$$Y_t \leq Y_\infty,\ \text{a.s.}\ , \forall t \qquad \text{(A.6)}$$

and

$$\mathbb{E}\{Y\}_\infty = \mathbb{E}\{\lim\}_{t\to\infty} Y_t = \lim_{t\to\infty} \mathbb{E}\{Y\}_t = a_\infty < \infty. \qquad \text{(A.7)}$$

Substituting (A.6) in (A.1) yields $|X_t| \leq M + Y_\infty \triangleq Y$ with $\mathbb{E}\{Y\} \leq M + \mathbb{E}\{Y\}_\infty < \infty$, where the last inequality holds by (A.7). ■

Proof:[Proof of Abstract Discovery Lemma] Let
$\tau_1 \triangleq \min\{t \geq 0 : \mathbb{P}\{D_t|\mathcal{F}_{t-1}\} > p\}$,
$\tau_i \triangleq \min\left\{t > \tau_{i-1} : \mathbb{P}\left\{D_t\,\mathcal{F}|_{t-1}\right| >\right\} p\right\}$, $\forall t \geq 2$ be an increasing sequence of random times, where $\tau_i = \infty$ if $i > Z$. Now, for any $t = 1, 2, ...$, we have that

$$\{\tau_i \leq t\} \Leftrightarrow \left\{\sum_{k=1}^{t} \mathbb{I}\{\mathbb{P}\{D_k|\mathcal{F}_{k-1}\} > p\} \geq i\right\}$$

and therefore $\{\tau_i \leq t\} \in \mathcal{F}_{t-1}$. Thus, each τ_i, with $i = 1, 2, ...$, is a stopping time (satisfying the conditions of Lemma 19).

Now, by Lemma 20, $\{\mathbb{I}\{D_t\} - \mathbb{P}\{D_t|\mathcal{F}_{t-1}\}\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}$, with uniformly integrable underlying martingale sequence. Therefore, by Lemma 19, the "sampled" process

$\{\mathbb{I}\{D_{\tau_i}\} - \mathbb{P}\{D_{\tau_i}|\mathcal{F}_{\tau_i-1}\}\}$ is a martingale difference sequence, with respect to $\{\mathcal{F}'_i\} \triangleq \{\mathcal{F}_{\tau_{i+1}-1}\}$. Thus, for any $z > 0$, we have that

$$\begin{aligned}
& \mathbb{P}\left\{\sum_{i=1}^{z}\mathbb{I}\{D_{\tau_i}\} \le \frac{p}{2}z,\ Z \ge z\right\} && \text{(A.8)}\\
\le\ & \mathbb{P}\left\{\sum_{i=1}^{z}[\mathbb{I}\{D_{\tau_i}\} - \mathbb{P}\{D_{\tau_i}|\mathcal{F}_{\tau_i-1}\}] \le -\frac{p}{2}z,\ Z \ge z\right\}\\
\le\ & \mathbb{P}\left\{\sum_{i=1}^{z}[\mathbb{I}\{D_{\tau_i}\} - \mathbb{P}\{D_{\tau_i}|\mathcal{F}_{\tau_i-1}\}] \le -\frac{p}{2}z\right\}\\
\le\ & \exp\left(-\frac{p^2}{8}z\right). && \text{(A.9)}
\end{aligned}$$

In this derivation, the first inequality follows since on the event $\{Z \ge z\}, \mathbb{P}\{D_{\tau_i}|\mathcal{F}_{\tau_i-1}\} > p,\ \forall 1 \le i \le z$ by the definition of τ_i. The second inequality follows by omitting the $\{Z \ge z\}$ event from probability. Finally, the third inequality follows by Azuma's Inequality [Azu67], which states that for any martingale difference sequence $\{M_i\}$, with $|M_i| \le c_i$ a.s.,

$$\mathbb{P}\left\{\sum_{i=1}^{z} M_i \le -\alpha\right\} \le \exp\left(-\frac{\alpha^2}{2\sum_{i=1}^{z} c_i^2}\right)$$

holds for any $\alpha > 0$. In our case, $\alpha = \frac{p}{2}z$ and $c_i = 1$.

To complete the proof, for any $z \ge \frac{2M}{p}$, write

$$\begin{aligned}
1 \ =\ & \mathbb{P}\left\{\sum_{k=1}^{\infty}\mathbb{I}\{D_k\} \le M\right\} && \text{(A.10)}\\
\le\ & \mathbb{P}\left\{\sum_{k=1}^{\infty}\mathbb{I}\{D_k\} \le \frac{p}{2}z\right\}\\
\le\ & \mathbb{P}\left\{\sum_{i=1}^{Z}\mathbb{I}\{D_{\tau_i}\} \le \frac{p}{2}z\right\}\\
=\ & \mathbb{P}\left\{\sum_{i=1}^{Z}\mathbb{I}\{D_{\tau_i}\} \le \frac{p}{2}z,\ Z \ge z\right\} && \text{(A.11)}\\
& +\mathbb{P}\left\{\sum_{i=1}^{Z}\mathbb{I}\{D_{\tau_i}\} \le \frac{p}{2}z,\ Z < z\right\}\\
\le\ & \mathbb{P}\left\{\sum_{i=1}^{z}\mathbb{I}\{D_{\tau_i}\} \le \frac{p}{2}z,\ Z \ge z\right\} + \mathbb{P}\{Z < z\}\\
\le\ & \exp\left(-\frac{p^2}{8}z\right) + \mathbb{P}\{Z < z\}.
\end{aligned}$$

Here, the first equality follows by the hypothesis of the theorem. The first inequality follows since $z \ge \frac{2M}{p}$. The second inequality holds since we are counting less time steps where the events D_t occur. Finally, the last inequality follows by (A.9). Thus, we have proved that $1 \le \exp\left(-\frac{p^2}{8}z\right) + \mathbb{P}\{Z < z\}$, which is equivalent to $\mathbb{P}\{Z \ge z\} \le \exp\left(-\frac{p^2}{8}z\right)$. Substitution of $z = \frac{2}{p}\left[M + \frac{p}{4}\log\frac{1}{\delta}\right]$ completes the proof of the Theorem. ■