Model-Based Exploration in Monitored Markov Decision Processes

by

Alireza Kazemipour

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

Trial-and-error is reinforcement learning's core idea. The success of the trial-and-error learning hinges on the assumption that each trial would lead to feedback. As a result, the feedback is used to improve the quality of decisions taken. The assumption that decision makers receive feedback for all their actions at all times does not necessarily translate to real-world scenarios. For example, a human observer may not always be able to provide rewards, a sensor to observe rewards may be limited or broken, or rewards may be unavailable during deployment. Monitored Markov decision processes (Mon-MDPs) have been proposed as a framework for sequential decision making where rewards could be unavailable, or in other words, unobservable to the decision maker. In this thesis, we consider Mon-MDPs. We revisit Mon-MDPs' model of interaction and the objective of decision makers in this model. Then, we introduce our main contribution, the monitored model-based interval estimation with exploration bonus (Monitored MBIE-EB) algorithm. Monitored MBIE-EB is the first algorithm in Mon-MDPs that provably admits a polynomial sample complexity. This polynomial sample complexity is to achieve a minimax-optimal policy in the worst-case. Monitored MBIE-EB pays attention to the structure of the problem at hand and furthermore is able to benefit from prior knowledge about the problem. Prior knowledge about the observability of the rewards is important. We show that Monitored MBIE-EB is able to fully exploit this knowledge, if available. Also, we show Monitored MBIE-EB is also capable of finding the minimax-optimal policy in the absence of privileged prior knowledge. Empirically, we demonstrate the superior performance of Monitored MBIE-EB compared to Directed Exploration-Exploitation, the state-of-art algorithm in Mon-MDPs, on four dozen finite domains.

Preface

This work was carried out in conjunction with Matthew E. Taylor, Michael Bowling, Simone Parisi and Montaser Fathelrhman Hussen Mohammedalamen. Initially, Simone Parisi, Matthew E. Taylor and Michael Bowling posed the question that in realistic scenarios usually agents have to act to observe the reward. The need of acting to observe was overlooked in the reinforcement learning's traditional model. Hence, Simone Parisi, Matthew E. Taylor and Michael Bowling founded the initial idea of monitored Markov decision processes where the agents have to also pay attention to how they can observe the reward. The concept of acting to observe the reward made many traditional algorithms inapplicable. Hence, the next step was to derive a reward-free exploration that independent of the reward could perform exploration. The further step, which significant portions of this thesis build upon, introduced a model-based algorithm where even never-observability of the environment reward could not hinder the agent's efficient exploration. This last work combined two contrasting ideas of optimism and pessimism to provide an efficient algorithm for the worst case. Currently, extending the ideas to settings with function approximation is followed by Simone Parisi and Montaser Fathelrhman Hussen Mohammedalamen.

Parts of Chapter 2 of this thesis has been published as S. Parisi, M. Mohammedalamen, A. Kazemipour, M.E. Taylor, and M. Bowling, "Monitored Markov Decision Processes," In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS '24). My contributions were initiating the code base, discussing the possible improvements at proposed model's each iteration, and the manuscript edits. S. Parisi and M. Mohammedalamen led the experiments' execution, proofs and contributed to writing the manuscript. M.E. Taylor, and M. Bowling were the supervisory authors involved with concept formation and manuscript composition.

Chapter 3 of this thesis has been published as A. Kazemipour, S. Parisi, M.E. Taylor, and M. Bowling, "Model-Based Exploration in Monitored Markov Decision Processes," In Proceedings of the 42nd International Conference on Machine Learning (ICML '25). I was responsible for the execution of experiments and analysis as well as the manuscript composition. S. Parisi, M.E. Taylor, and M. Bowling were the supervisory authors and were involved with concept formation and manuscript composition.

The baseline algorithm in Chapter 4 has been published as S. Parisi, A. Kazemipour, and M. Bowling, “Beyond Optimism: Exploration With Partially Observable Rewards,” In 38th Conference on Neural Information Processing Systems (NeurIPS 2024). I contributed to manuscript edits and the proofs. S. Parisi led the execution of experiments, and the manuscript composition. M. Bowling was the supervisory author and was involved with the proofs and manuscript composition.

This thesis has taken place in the Intelligent Robot Learning (IRL), and Reinforcement Learning Artificial Intelligence (RLAI) labs at the University of Alberta, IRL lab is supported in part by research grants from Alberta Innovates; Alberta Machine Intelligence Institute (Amii); a Canada CIFAR AI Chair, Amii; Digital Research Alliance of Canada; Mitacs; and the National Science and Engineering Research Council (NSERC).

The collaborative nature of this project necessitated the usage of “we” as the first person pronoun in writing of this thesis. However, I remain solely responsible for all the errors present, whether they be due to technical or presentation issues.

Alireza Kazemipour

August, 2025

Acknowledgements

I would like to extend my gratitude toward my supervisors Dr. Matthew E. Taylor and Dr. Michael Bowling for their support and kindness at every step of the journey. Especially, I would like to acknowledge that I will always be indebted to Matt for believing in my email and our short virtual meeting to facilitate my admission to the Department of Computing Science at the University of Alberta. It was and is my ultimate place for my education and I would never want to study anywhere else. Matt made my dream come true. I would also like to admit that working with Mike was beyond my wildest dreams and to this day, I cannot believe I had the privilege to be his student. I am always thankful to Mike for his modesty, support and all he taught me.

I am also grateful to Dr. Csaba Szepesvári for accepting to be my committee examiner, and my collaborators Dr. Simone Parisi and Montaser Fathelrhman Hussen Mohammedalamen who we published multiple papers together and I, as an inexperience MSc student, learned a lot beside them. I would like to thank all my lab mates in Matt's and Mike's student groups as well as Shivam Garg in providing the template format of this thesis, and also Antonie Bodley for helping me in enhancing the clarity of the presented material. I also acknowledge that I used ChatGPT for rephrasing some of my own generated sentences to improve the clarity.

I thank my family: my lovely little sister for listening to me in days that I needed someone to talk to, my dad for always believing in me and inviting to be calm, and my mom who literally sacrificed her best days of life for mine; not any second passes without her in my heart and mind.

Finally, I thank God for making the impossible possible. Only He and I really know the path that led me to studying at the Department of Computing Science at the University of Alberta. A path that by any logical explanation is impossible. I would like to admit that he me left speechless by doing an absolute miracle.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

Imagine purchasing a plant for your apartment that requires regular watering. When you are available, you enjoy taking care of it, ensuring all its needs, including watering, are met. But during the plant's life span, you might sometimes be unavailable, such as when you are away from home for an extended amount of time. In such cases, you might preemptively buy a robot to do this task continuously on your behalf. After a training period, where you reward or punish the robot based on its performance, eventually you train the robot to water the plant reliably. This gives you confidence that the plant's well-being is secure, even in your absence, and you may even consider adding more plants to your care, trusting the robot to maintain them all.

At some point, an occasion might arise in which you are not at home, and the robot must interact with a novel item or plant, in this case, that was not encountered during training. When this happens, what should the robot do? How should the robot adapt if it knows you are away compared to being unaware? Additionally, if the robot needs to navigate between rooms, but not know the exact paths in your absence, how can it reach a plant? How can the robot pay attention to the cost of having your supervision?

In this thesis, we aim to answer the above questions in the context of sequential decision-making and agent behavior. We model the concept of imperfect knowledge about the quality of outcomes and formalize the desired behavior (in scenarios such as the plant-watering robot). We then introduce monitored model-based interval estimation with exploration bonus (Monitored MBIE-EB) algorithm that adopts this behavior with high probability in finite time. Finally, we demonstrate the empirical advantage of Monitored MBIE-EB over Directed Exploration-Exploitation, a previously developed algorithm for this setting.

1.1 Sequential Decision-Making

In this section, we introduce the components of a single decision maker's decision-making process. This introduction provides the concepts needed to study which decisions should be taken when knowledge of potential outcomes is imperfect. Also, we will visualize the example of the plant-watering robot in Figure 1.1.

In this work, decision-makers are called *agents*. We study the decision-making procedure of an agent with respect to the *environment* in which it interacts. In our plant-watering example, the agent is the robot. The robot's environment, consists of the apartment and the plants.

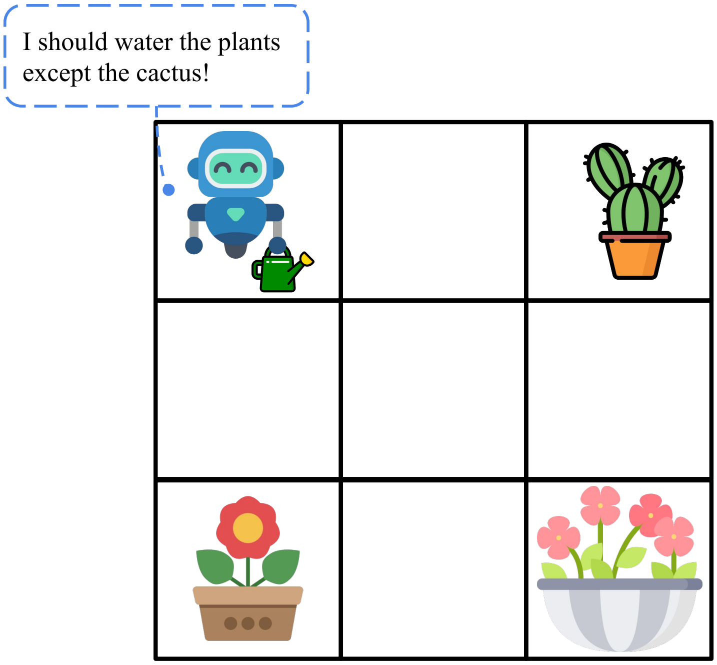


Figure 1.1: This figure depicts the plant-watering robot scenario. The robot is responsible for watering all the plants except the cactus on behalf of the owner, who is absent.

Agents observe information from the environment, such as the time of day or the dryness of the plants in our example. An agent *acts* based on its *observations*—for example, watering a dry plant or stopping when the plant is well-hydrated. To evaluate the quality of these possible actions, agents receive a reward signal that either rewards or penalizes them based on their actions. In our example, watering a dry plant is rewarded, while over-watering is penalized. Hence, the agent's *goal* is to maximize the accumulation of the reward they receive while interacting with the environment [8]. Looking at Figure 1.1, the robot should water the small flower pot with a single flower and the large pot with multiple flowers, but avoid watering the cactus. This behavior maximizes the total reward the agent can obtain.

1.1.1 Learning When The Rewards Are Unobservable

In this section, we highlight one of the unrealistic assumptions that can be made when modeling sequential decision-making: *the reward is always observable by the agent.* We then propose a slightly more realistic alternative. The plant-watering robot scenario indicates the concept of learning through trial and error. The robot is never shown exactly how to reach and water a plant; instead, it follows what it computes to be the best course of action. Then, the agent should determine the correct behavior from the reward signals it received for those actions. Provided a human is available to provide feedback, the robot's learning is possible. However, the robot's reliance on the human's presence is quite limiting. Ultimately, the human bought the robot to handle the tasks for them, giving them time to focus on other activities. Intuitively, one way to address this shortcoming is for the robot to build an internal model of the relationship between actions and rewards, enabling it to evaluate the quality of its own actions. This nearly provides the desired autonomy, but it is not robust against possible unexpected events. For example, in our plan-watering scenario, some unexpected events could include the noisy signal of the robot's sensors leading it to the wrong place, or a neighbor coming in to check on the plants and and puts the robot in the wrong room, etc. Such changes will require self-evaluation of the robot, but how can the robot evaluate itself if it has never experienced the new situation before?

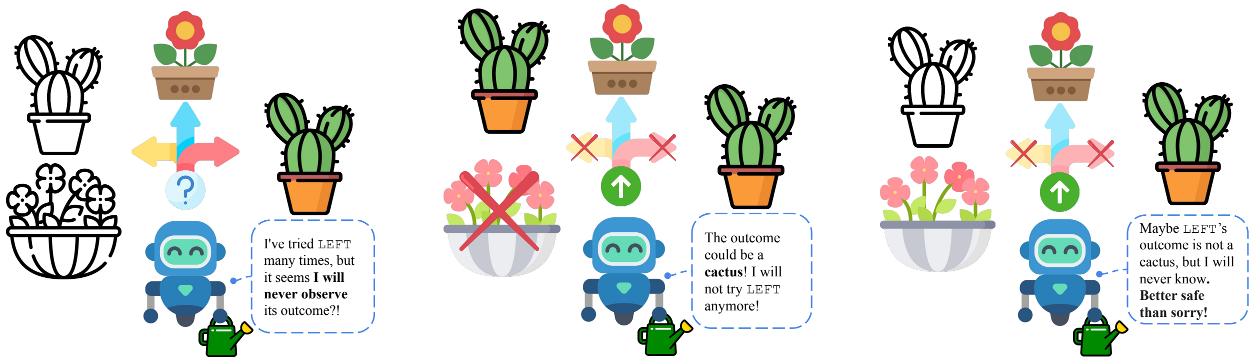
The central difficulty in answering the above question is the degree of expected autonomy. For example, one answer could be that the robot should turn itself off, an automated procedure could inform a neighbor for help, etc. Although the randomness in the plant-watering scenario might seem artificial, it is common because of the world's natural stochasticity. For example, Andrychowicz et al. [2] trained a robot arm only in the simulation to solve a Rubik's Cube, which is a complex task. After the training phase in the simulation, they test the learned behavior on an actual robotic arm with no training during the deployment. However, they mention:

> "A variety of randomizations are applied to the simulator, shrinking the reality gap between the simulated environment and the physical world in order to learn a policy that generalizes to reality." [2]

The added randomization is only applicable if the real-world setup is exactly as designed. Similar to the plant-watering situation, many challenges exist. For example, what if the learned policy is tested on a slightly different robotic arm? Or if the Rubik's Cube varies during testing? One option is to enable the agent to continue learning during deployment. However, should experiment designers be available at all times during this process? After all, the need for constant supervision should be minimized.

On the other hand, even if agents are allowed to continually learn during their interaction with the environment, their hardware may still produce faulty computations, e.g., in large-scale computing clusters, the proximity of circuit boards can introduce noise, leading to erroneous outputs [12].

Hence, it is reasonable to expect agents to learn continually, but they also should account for unexpected events. The solution presented here is to encompass both of these desiderata. Agents should pursue their goals (maximizing the accumulation of rewards) while avoiding the worst-case possible outcome when encountering novel, unexpected situations. This is a design decision appropriate for the worst-case scenario. Returning to our plant-watering example: when the agent encounters an unfamiliar plant, it should water it a few times to gather feedback. If, after several attempts, no reward (not even the constant numerical reward of zero) is received, the agent can reasonably infer that the plant is a cactus and should stop watering it. This behavior is illustrated in Figure 1.2. We argue that this is a reasonable strategy, provided that watering a cactus results in the worst possible outcome. Under this assumption of considering the worst case for unknown outcome, the agent continues watering all known plants and adopts a cautious approach toward unfamiliar ones. If the unknown plant is indeed a cactus, the agent has already avoided harm by stopping early. If it is not a cactus, the agent's behavior is suboptimal—but not catastrophic.


I've tried LEFT many times, but it seems **I will never observe** its outcome?!
The outcome could be a **cactus!** I will not try LEFT anymore!
Maybe LEFT's outcome is not a cactus, but I will never know. **Better safe than sorry!**

(a) `LEFT`'s outcome cannot be observed and will always be unknown.

(b) The agent assumes the worst (cactus) for `LEFT`.

(c) The agent would be pessimistic even if `LEFT`'s true outcome is not the cactus.

Figure 1.2: An example of dealing with unknown outcomes. (a) The agent has to choose between `LEFT`, `UP`, and `RIGHT`. `RIGHT` leads to a cactus, `UP` to a small flower pot, and `LEFT` to either a cactus or a big flower pot (more valuable than a small flower pot), but the agent *can never observe the result* of executing `LEFT`. (b) After sufficient attempts, the agent excludes `LEFT` because its outcome is unknown, and the agent assumes the worst. (c) `LEFT` is ruled out even though it could actually yield the big flower pot. However, since this cannot be known, acting pessimistically complies with our suggestion when dealing with unknown outcomes. Ultimately, the agent balances exploration with the risk that some actions may offer no reward. Thus, after enough exploration, the agent assumes the worst if the action outcome is still unknown.

We have described the plant-watering robot example, which has been assigned to water plants except cacti. Now, there is a need to study how the robot's behavior should be formally studied and defined. Reinforcement learning (RL) is used to study how agents make decisions. In one typical form, known as Markov decision processes (MDPs), RL assumes that rewards are observable for every action, because they specify the goals. However, another formalization, monitored Markov decision processes (Mon-MDPs), accounts for the possibility of unobservable rewards. Despite the Mon-MDPs' existence, no algorithms have been designed yet to guide agents toward the desired behavior, while paying attention to the Mon-MDP's structure at hand. In this thesis, we aspire to fill this gap.

1.2 Contributions

We now summarize the key contributions of this thesis, we:

- define the minimax-optimality in Mon-MDPs replacing the notion of MDPs' optimality.
- present Monitored MBIE-EB, the first model-based minimax-optimal method for Mon-MDPs.
- prove the polynomial time sample complexity of Monitored MBIE-EB.
- show that the dependence of the Monitored MBIE-EB's sample complexity on the stochasticity of observing the reward in Mon-MDPs is essentially unimprovable.
- demonstrate the superior performance of Monitored MBIE-EB compared to Directed-Exploration-Exploitation, the previous state-of-the-art (SOTA) algorithm, on over four dozen finite domains. We show more dramatic results when the dynamics of how the agent can or cannot observe the reward is known apriori.

1.3 Organization

To study sequential decision-making with unobservable rewards, in Chapter 2, we revisit reinforcement learning (RL) that formally models the sequential decision making. In particular, we start with the formulation of Markov decision processes (MDPs) and review the necessary topics studied under this formulation. Then, we introduce monitored Markov decision processes (Mon-MDPs) to extend MDPs to scenarios in which the reward could be unobservable. We formalize the minimax-optimality as the objective on Mon-MDPs. In Chapter 3, we introduce Monitored MBIE-EB that accomplishes the minimax-optimality in Mon-MDPs. We prove the polynomial sample complexity of Monitored MBIE-EB. Finally, in Chapter 4, we show the superior empirical performance of Monitored MBIE-EB compared to the previous SOTA method on some finite domains.

Chapter 2

Background

In this chapter, we introduce the background needed to study sequential decision-making when the agent might not receive the rewards of its actions.

2.1 Markov Decision Processes

In this section, we revisit the MDP definition as the building block of studying sequential decision-making with long-term effects. Throughout this thesis, we only focus on *finite* MDPs. A finite MDP is represented by a tuple $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p \rangle$ [43, 48]. $\mathcal{S}$ is the finite set of the environment's states. Each state contains all the necessary information the agent needs to know about the environment. $\mathcal{A}$ is the finite set of actions the agent can apply in each of the environment's states, $\mathcal{R} \subset \mathbb{R}$ is the finite set of rewards the agent receives from the environment upon taking its actions, and $p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{R} \times \mathcal{S})$[1] is the transition dynamics. In particular, $p\,(s', r'|s, a)$ is the probability of seeing the reward $r' \in \mathcal{R}$ and the next state $s' \in \mathcal{S}$ given the action $a \in \mathcal{A}$ was taken in state $s \in \mathcal{S}$.

In MDPs, to choose its action, it is sufficient to assume the agent follows a memoryless policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$[2]. Following π in M results in a stochastic process $S_0, A_0, R_1, S_1, A_1, R_2, \ldots$ that induces a probability measure $\mathbb{P}$ over some sample space Ω. In the stochastic process $S_0, A_0, R_1, S_1, A_1, R_2, \ldots$, we have that $S_0, S_1, \cdots : \Omega \to \mathcal{S}$, $A_0, A_1, \cdots : \Omega \to \mathcal{A}$, and $R_1, R_2, \cdots : \Omega \to \mathcal{R}$, such that for any $t \geq 0$, $s_0, s_1, \ldots s_{t+1} \in \mathcal{S}$, $a_0, a_1, \ldots a_t \in \mathcal{A}$, and $r_0, r_1, \ldots r_{t+1} \in \mathcal{R}$, it holds

$$\mathbb{P}\left(S_{t+1} = s_{t+1}, R_{t+1} = r_{t+1} | S_0 = s_0, A_0 = a_0, R_1 = r_1, \ldots S_t = s_t, A_t = a_t\right) = p(s_{t+1}, r_{t+1} \mid s_t, a_t), \tag{2.1}$$

[1] $\Delta(\mathcal{X})$ denotes the probability distribution over the set $\mathcal{X}$.

[2] Memoryless policies only use the current environment state to make a decision. Hence, π is a mapping from the state space.

known as the Markov property. Figure 2.1 illustrates the agent-environment interaction in MDPs.

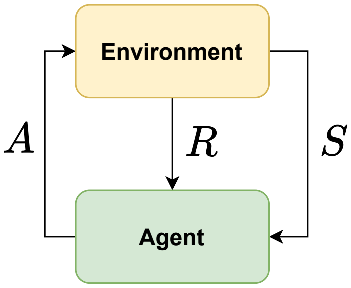


Figure 2.1: MDPs. The agent takes an action A, and in turn, it receives the state of the environment S and the reward R corresponding to the taken action. For the sake of clarity, we have omitted the dependence on time from the notation.

2.1.1 Learning Objective in MDPs

MDPs model the interaction between the agent and the environment. In this section, we revisit the agents' goal during this interaction. One criterion to express the agents' goal, which follows a policy π in M, is maximizing the expected sum of discounted rewards [43], where $0 \leq \gamma < 1$ is the discount factor. This leads to the definition of the state-value and the action-value functions, and the optimal policy. Let $\mathbb{E}$ denote the expectations we get with respect to $\mathbb{P}$. Then, the state-value function of a policy π in MDP M is denoted by V_M^π where:

$$V_M^\pi(s) := \mathbb{E}\left[\sum_{k\geq t}\gamma^{k-t}R_{k+1}\middle| S_t = s\right], \qquad \forall s \in \mathcal{S}. \tag{2.2}$$

Similarly, the action-value function of a policy π in MDP M is denoted by Q_M^π where:

$$Q_M^\pi(s,a) := \mathbb{E}\left[\sum_{k\geq t}\gamma^{k-t}R_{k+1}\middle| S_t = s, A_t = a\right], \qquad \forall s,a \in \mathcal{S}\times\mathcal{A}. \tag{2.3}$$

For a memoryless policy π, the following relation between $V_M^\pi(s)$ and $Q_M^\pi(s,a)$ holds:

$$V_M^\pi(s) = \sum_a \pi(a \mid s) Q_M^\pi(s,a), \qquad \forall s \in \mathcal{S}, \tag{2.4}$$

i.e., the state-value of a policy is the mean of its action-value, when only the randomness in the policy is considered. Having defined the state-value and action-value functions, the agent's goal is defined as finding a policy π^*, called an optimal policy, whose state-value function in M satisfies:

$$\underbrace{V_M^{\pi^*}}_{=:V_M^*}(s) = \max_{\pi \in \Pi} V_M^{\pi}(s), \qquad \forall s \in \mathcal{S}, \tag{2.5}$$

where Π is the set of all policies in M.

2.1.2 Bellman Optimality Equation

The agent's goal was defined as finding π^*. Now, we revisit the Bellman optimality equation as one way of finding π^*. To introduce the Bellman optimality equation, define the immediate expected reward $r(s,a)$ and next-state transition probability $p(\cdot \mid s,a)$ for taking action a at state s as:

$$r(s,a) = \sum_{r' \in \mathcal{R}, s' \in \mathcal{S}} r' \cdot p\left(s', r' \middle| s, a\right), \qquad p\left(s' \middle| s,a\right) = \sum_{r' \in \mathcal{R}} p\left(s', r' \middle| s, a\right), \quad \forall s' \in \mathcal{S}.$$

Then, V_M^* satisfies the following recursive equation known as the Bellman optimality equation [49]:

$$V_M^*(s) = \max_a \left\{ r(s,a) + \gamma \sum_{s'} p\left(s' \middle| s,a\right) V_M^*(s') \right\}, \qquad \forall s \in \mathcal{S}. \tag{2.6}$$

The utility of the Bellman optimality equation is that it is shown [49] by recursively applying this equation (when viewed as a mapping from $\mathbb{R}^{|\mathcal{S}|}$ to $\mathbb{R}^{|\mathcal{S}|}$) on any initial value, V_M^* can be found up to predefined accuracies. Using the Bellman optimality equation, any memoryless policy π that

$$\pi(s) \in \arg\max_a \left\{ r(s,a) + \gamma \sum_{s'} p\left(s' \middle| s,a\right) V_M^*(s') \right\}, \qquad \forall s \in \mathcal{S}, \tag{2.7}$$

is optimal [43]. Hence, Equation (2.7) with Equation (2.4) imply the following results:

$$\begin{aligned} Q_M^*(s,a) &= r(s,a) + \gamma \sum_{s'} p\left(s' \middle| s,a\right) V_M^*(s'), \qquad (2.8)\\ V_M^*(s) &= \max_a Q_M^*(s,a), \\ \pi^*(s) &= \arg\max_a Q_M^*(s,a). \end{aligned}$$

One approach to identify π^* is constructing action-values $Q_M(s,a)_i$ for all state-action pairs (s,a), where $i \in \{0,1,\dots\}$ and $Q_M(s,a)_0$ is chosen arbitrarily. Then perform the update, $Q_M(s,a)_{i+1} = r(s,a) + \gamma \sum_{s'} p\left(s'|s,a\right) \max_{a'} Q_M(s',a')_i$. Then as $i \to \infty$, $Q_M(s,a)_i \to Q_M^*(s,a)$ almost surely. This approach is known as the *value iteration* [48], which makes $Q_M(s,a)_i$ converge to $Q_M^*(s,a)$ at the geometric rate γ^i. In practice, to get an estimate $Q_M(s,a)_i$, such that $Q_M(s,a)_i \geq Q_M^*(s,a) - \epsilon$

for $\epsilon > 0$, it is sufficient to perform value iteration only for $i \in \mathcal{O}\left((1-\gamma)^{-1} \ln\left(\frac{1}{\epsilon(1-\gamma)}\right)\right)$ iterations [30], where it is assumed that $\min\{r' : r' \in \mathcal{R}\} = 0$ and $\max\{r' : r' \in \mathcal{R}\} = 1$.

2.2 Model-Based Interval Estimation with Exploration Bonus

In this section, we revisit Model-Based Interval Estimation with Exploration Bonus (MBIE-EB) [13, 47, 55] that yields a policy satisfying the Bellman optimality equation. The fundamental challenge for the agent is that Q^*_M, as defined in Equation (2.8), assumes the exact model of the immediate expected reward r and next-state transition probabilities p for all state-actions is known. However, neither r nor p is known to the agent in advance, and the agent can only *estimate* them during interaction. Using samples to estimate true values leads to using maximum likelihood estimation (MLE) to construct $\bar{r}$ as an estimate of r and $\bar{p}$ as an estimate of p. Empirical models $\bar{r}$ and $\bar{p}$ for a fixed state-action (s, a), are:

$$\bar{r}(s,a) = \frac{1}{N(s,a)} \sum_{i=1}^{N(s,a)} R_i, \qquad \bar{p}\left(s' \middle| s,a\right) = \frac{1}{N(s,a)} \sum_{i=1}^{N(s,a)} \mathbb{I}\left\{S'_i = s'\right\}, \qquad \forall s' \in \mathcal{S}. \tag{2.9}$$

To compute $\bar{r}$ and $\bar{p}$ above, $N(s,a)$ is the number of times that (s,a) has been visited, R_i and S'_i are the immediate reward and the next-state after the ith visit for $i \in [N(s,a)]$, and $\mathbb{I}$ is the indicator function returning one if the predicate of its argument is true, otherwise zero. Note that $\bar{r}$ and $\bar{p}$ are sample estimates and are not necessarily equal to their true quantities. When they are used in Equation (2.8), they result in the following equation for (s, a):

$$\bar{Q}^*_M(s,a) \coloneqq \bar{r}(s,a) + \gamma \sum_{s'} \bar{p}\left(s' \middle| s,a\right) V^*_M(s'). \tag{2.10}$$

However, the agent cannot act according to $\bar{Q}^*_M$, as defined in Equation (2.10), because V^*_M is also an unknown quantity. The idea of optimism in the face of uncertainty (OFU) turns Equation (2.10) into an equation that does not have any unknown terms, and guides the agent toward state-action where $\bar{r}$ and $\bar{p}$ are the most inaccurate (the fewest number of times visited). MBIE-EB is an algorithmic instantiation of OFU. It states that since the rewards are bounded, without loss of generality between zero and one, V^*_M is deterministic and $\mathbb{E}\left[\bar{Q}^*_M(s,a)\right] = Q^*_M(s,a) \leq \frac{1}{1-\gamma}$, Chernoff-Hoeffding's inequality guarantees with probability at least $1 - \delta_1$:

$$\left|\bar{Q}^*_M(s,a) - Q^*_M(s,a)\right| \leq (1-\gamma)^{-1} \sqrt{\frac{\ln(1/\delta_1)}{2N(s,a)}}. \tag{2.11}$$

Hence, with probability at least $1 - \frac{\delta_1}{2}$ (to split the failure probability equally for the upper and lower bounds of the absolute value), it holds that:

$$\bar{Q}^*_M(s,a) - Q^*_M(s,a) \geq -(1-\gamma)^{-1}\sqrt{\frac{\ln(2/\delta_1)}{2N(s,a)}}$$

$$\underbrace{\bar{Q}^*_M(s,a) + (1-\gamma)^{-1}\sqrt{\frac{\ln(2/\delta_1)}{2N(s,a)}}}_{=:Q^{**}_M(s,a)} \geq Q^*_M(s,a),$$

where Q^{**}_M represents an optimistic (bigger than the optimal) action-value function in M.

By showing that $V^{**}_M(s) = \max_a Q^{**}_M(s,a) \geq V^*_M(s)$ through induction for all states [13, 47], $\bar{Q}^*_M$ is turned into Q^{**}_M that does not have any unknown quantities and the agent can be greedy with respect to. Optimistic action-values, Q^{**}_M, satisfy the following:

$$Q^{**}_M(s,a) = \bar{r}(s,a) + \gamma \sum_{s'} \bar{p}\left(s' \middle| s,a\right) \underbrace{V^{**}_M(s')}_{\max_{a'} Q^{**}_M(s',a')} + (1-\gamma)^{-1}\sqrt{\frac{\ln(2/\delta_1)}{2N(s,a)}}. \tag{2.12}$$

However, care should be taken to use Q^{**}_M, as defined in Equation (2.12), in place of Q^*_M:

1. The condition $V^{**}_M(s) \geq V^*_M(s)$ should not remain in the limit. The agent should eventually follow a policy π such that $V^\pi_M(s) \geq V^*_M(s) - \epsilon, \forall s \in \mathcal{S}, \epsilon > 0$. At the same time, optimism is required for learning accurate models. Hence, MBIE-EB states that the agent should only be optimistic until each state-action is visited at least m times, where m is the least number of visits to learn near-accurate models. Lemmas 2.1, 2.2 and 2.3 quantify the order of values that m should take to have near-accurate models. We refer the reader to Strehl and Littman [47] for their proofs.

   Lemma 2.1 bounds the difference between the action-value of a fixed policy π in two MDPs M_1 and M_2 in terms of how much M_1 and M_2 are different from each other. It is known as the *Simulation Lemma* [23, 47], and it is useful in proving Lemma 2.2.

   Lemma 2.1 (Strehl and Littman [47, Lemma 1]). *Let $M_1 = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}_1, p_1 \rangle$ and $M_2 = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}_2, p_2 \rangle$ be two MDPs with non-negative rewards bounded by $r_{\max}$ and $0 \leq \gamma < 1$. If the following holds for all states and actions:*

   $$|r_1(s,a) - r_2(s,a)| \leq \varphi_1, \qquad \|p_1(\cdot \mid s,a) - p_2(\cdot \mid s,a)\|_1 \leq \varphi_2,$$

   then for all state-action pairs, and deterministic stationary polices we have

   $$|Q^\pi_1 - Q^\pi_2| \leq \frac{\varphi_1 + \gamma r_{\max}\varphi_2}{(1-\gamma)^2}.$$

MBIE-EB performs based on its sample estimates. Lemma 2.2 states that a fixed policy exhibits similar action-values in two MDPs with comparable transition dynamics and expected immediate rewards. Consequently, an agent achieves near-desired behavior by having accurate estimates.

Lemma 2.2 (Strehl and Littman [47, Lemma 2]). *Let $M_1 = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}_1, p_1 \rangle$ and $M_2 = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}_2, p_2 \rangle$ be two MDPs with non-negative rewards bounded by $r_{\max} \geq 1$ and $0 \leq \gamma < 1$. Suppose the following inequalities hold for all state-actions*

$$|r_1(s,a) - r_2(s,a)| \leq \varphi_1, \qquad \|p_1(\cdot \mid s,a) - p_2(\cdot \mid s,a)\|_1 \leq \varphi_2 \ .$$

For any $0 < \epsilon \leq \frac{r_{\max}}{1-\gamma}$ and fixed policy π, there is a constant C, that if $\varphi_1 = \varphi_2 = C\frac{\epsilon(1-\gamma)^2}{r_{\max}}$, then

$$|Q_1^\pi - Q_2^\pi| \leq \epsilon \ .$$

Lemma 2.3 specifies the number of visits to each state-action for MBIE-EB to have accurate models.

Lemma 2.3 (Strehl and Littman [47, Lemma 5 and Theorem 1]). *Let $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p \rangle$ and $\gamma \in [0,1)$. For $\tau = C\frac{\epsilon(1-\gamma)^2}{r_{\max}}$, where C is specified in Lemma 2.2, there exists an $m \in \mathcal{O}\left(\frac{|\mathcal{S}|}{\tau^2} + \frac{1}{\tau^2}\ln\frac{|\mathcal{S}||\mathcal{A}|}{\tau\delta}\right) = \mathcal{O}\left(\frac{|\mathcal{S}|}{\epsilon^2(1-\gamma)^4} + \frac{1}{\epsilon^2(1-\gamma)^4}\ln\frac{|\mathcal{S}||\mathcal{A}|}{\epsilon(1-\gamma)^2\delta}\right)$ such that $|\bar{r}(s,a) - r(s,a)| \leq \tau$ and $\|\bar{p}(\cdot \mid s,a) - p(\cdot \mid s,a)\|_1 \leq \tau$ with probability at least $1-\delta$ hold for all at least m times visited state-action pairs.*

2. The failure probability δ_1 in Equation (2.12) applies to a fixed (s,a), but it should account for all state-actions until each is visited at least m times.

These two requirements are satisfied by specifying $\delta_1 = \frac{\delta}{|\mathcal{S}||\mathcal{A}|m}$ where δ is the global failure probability. With probability at least $1-\delta$, the following action-values are optimistic until all state-actions are visited at least m times:

$$Q_M^{**}(s,a) = \bar{r}(s,a) + \gamma \sum_{s'} \bar{p}\left(s' \middle| s,a\right) \max_{a'} Q_M^{**}(s',a') + (1-\gamma)^{-1}\sqrt{\frac{\ln(2\,|\mathcal{S}|\,|\mathcal{A}|\,m/\delta)}{2N(s,a)}}. \tag{2.13}$$

In summary, Q_M^{**}, in Equation 2.13, is the action-value that MBIE-EB's policy is greedy with respect to. Q_M^{**}, with high probability, converges to the optimal-action values Q^* using only sample estimates of the expected reward $\bar{r}$ and dynamics $\bar{p}$.

2.3 Sample Complexity

In this section, we explain why MBIE-EB is appropriate for finding the optimal policy in finite MDPs. The central goal of developing RL algorithms in MDPs is finding an optimal policy. Ideally, it is desirable to develop algorithms that are not only effective in finding an optimal policy but are also efficient in terms of number of samples they need. Such desideratum rules out asymptotic algorithms such as ε-greedy [48] in favor of other algorithms exhibiting good finite-time performance. One criterion to measure the efficiency of algorithms is measuring the number of time steps required to find a near-optimal policy known as the *sample complexity* [23]. Inspired by the probably approximately correct (PAC) framework [51] used to measure the efficiency of supervised learning algorithms [56], the probably approximately correct MDP (PAC-MDP) framework [23, 28, 47] provides a similar measure for RL algorithms.

Definition 1. *An algorithm is PAC-MDP if, for any MDP $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p \rangle$ with discount factor $0 \leq \gamma < 1$ and $\delta, \epsilon > 0$, it finds an ϵ-optimal policy in M, in time polynomial to*

$$\left(|\mathcal{S}|, |\mathcal{A}|, \ln \frac{1}{\delta}, \frac{1}{\epsilon}, \frac{1}{1-\gamma}, r_{\max} \right).$$

Theorem 2.1 states that MBIE-EB is a PAC-MDP algorithm.

Theorem 2.1 (Strehl and Littman [47, Theorem 2])**.** *Suppose ϵ and δ are two real numbers between 0 and 1 and $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p \rangle$ is any MDP with non-negative rewards bounded by $r_{\max}$. Let discount factor $0 \leq \gamma < 1$. There exist an input $m = m\left(\frac{1}{\epsilon}, \frac{1}{\delta}\right)$ satisfying $m\left(\frac{1}{\epsilon}, \frac{1}{\delta}\right) = \mathcal{O}\left(\frac{|\mathcal{S}|}{\epsilon^2(1-\gamma)^4} + \frac{1}{\epsilon^2(1-\gamma)^4} \ln \frac{|\mathcal{S}||\mathcal{A}|}{\epsilon(1-\gamma)^2\delta}\right)$, such that if MBIE-EB is executed on M and value iteration is done every $H \in \mathcal{O}\left((1-\gamma)^{-1} \ln \frac{1}{\epsilon_1(1-\gamma)}\right)$ steps for H iterations, then the following holds. Let π_t denote MBIE-EB's policy at time t. With probability at least $1-\delta$, $V_M^{\pi_t}(S_t) \geq V_M^*(S_t) - \epsilon$ is true for all but $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|H}{\epsilon^3(1-\gamma)^5}\right) = \widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{\epsilon^3(1-\gamma)^6}\right)$ time steps.*

The main takeaway of Theorem 2.1 is that not only does MBIE-EB find a near-optimal policy using only sample estimates of the reward function and the transition dynamics, but also finds the near-optimal policy in polynomial time. Hence, MBIE-EB is efficient compared to methods that find an optimal policy only asymptotically such as ε-greedy.

2.4 Monitored Markov Decision Processes

In the preceding sections, we explained the MDP formulation, the agents' goal in MDPs, the MBIE-EB algorithm that achieves the goal, and the polynomial sample complexity of MBIE-EB. In this section, we argue against the constant availability of the reward function which can be an unrealistic assumption made in the MDP formulation. This argument motivates the introduction of monitored Markov decision processes that addresses this limitation of MDPs to model scenarios where the reward could be unavailable to the agent.

One central assumption in RL is that upon taking an action, the agent receives a reward. However, receiving the reward at all times may be an unrealistic assumption because an exogenous entity can determine the reward to the agent such as humans [16, 31] or monitoring instrumentation [53]. In such settings, the assumption that the reward is available at all times is not reasonable because of humans' time constraint [41], hardware failure [7, 12, 17], or simply the inaccessibility of rewards during deployment [2]. MDPs do not account for this possibility. They abstract away the process that rewards the agent as being part of the environment. MDPs assume all the required information, including the reward, is provided to the agent in response to its actions.

An alternative to an MDP is a partially observable Markov decision process (POMDP) [22]. The POMDP formulation relaxes the assumption that the agent receives full information from the environment. POMDP expresses that in response to agent's actions, the agent receives a vector of observations O instead of the state and the reward. The observations O may or may not contain enough information about the state of the environment that agent could base its decision on, and the reward is just simply an element of O [48]. Despite the POMDPs' existence, no single study in the POMDP literature exists that investigates the partial observability of the reward [11, 19, 42, 46]. POMDP literature only include work that studied the partial observability of the environment state due to imperfect perception of the agent, not due to unavailability of the process that determines the reward. Figure 2.2 illustrates the agent-environment interaction is POMDPs.

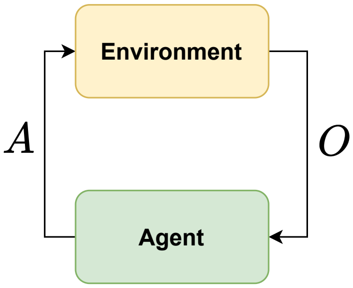


Figure 2.2: POMDPs. The agent takes an action A, and in return it receives the vector of observations O. There is no explicit reward signal because it is one of the O's entries. For the sake of clarity, we have omitted the dependence on time from the notation.

On the other hand, the partial observability of the reward has been studied in the context of RL from human feedback [24], active learning [27], options framework [32] and goal-conditioned policies [14, 54]. Nevertheless, in none of these settings are there rewards that the agent *only sometimes observes* whose observability is predictable and is possibly controlled. Such as the possibility of observing the reward upon turning the light on in a dark room or carrying an object. Consequently, inspired by the problem of partial monitoring [5], Parisi et al. [40] introduced monitored Markov decision processes (Mon-MDP) in an attempt to explicitly bring the process that provides the agent with the reward into the problem definition. The promise of the Mon-MDP formulation is to adequately model partially observable or even never observable rewards. Therefore, Mon-MDPs propose that the agent should also pay attention to how its actions might or might not impact its ability to observe the reward, if possible at all. This idea encompasses not only settings that have a specialized notion of partially observable rewards, but also generalizes all of them. Mon-MDPs state that the reason rewards might be partially observable stems from a process that the agent should account for. As a result, Mon-MDPs preserve the reward-providing process rather than abstracting it away through the traditional concept of the environment. This explicitness with respect to the reward-provider *might* lead to more efficient algorithms, but it surely leads to *more realistic* algorithms than the ones that do not consider the unobservability of the reward at all.

In Mon-MDPs, in addition to the environment that is modeled by an MDP, the agent also interacts with *another MDP* called the monitor. As examples, the monitor may represent a human supervisor or monitoring instrumentation that are some of the common processes specifying the reward to learning agents. As mentioned earlier, the monitor is an MDP thus its definition includes a state space, an action space, a set of rewards and a transition function obeying the Markov property. The agent interacts jointly with the monitor and the environment; therefore to distinguish the quantities, spaces and mappings belonging to the environment or the monitor, we will use superscripts e and m respectively. Let $\mathcal{S}^{\mathrm{m}}$ denote the finite state space of the monitor, $\mathcal{A}^{\mathrm{m}}$ denote the finite action space of the monitor, $\mathcal{R}^{\mathrm{m}}$ denote the finite set of the monitor's reward, and p^{m} denote the transition dynamics of the monitor. In Mon-MDPs $p^{\mathrm{m}} : \mathcal{S}^{\mathrm{m}} \times \mathcal{A}^{\mathrm{m}} \times \mathcal{S}^{\mathrm{e}} \times \mathcal{A}^{\mathrm{e}} \to \Delta(\mathcal{S}^{\mathrm{m}} \times \mathcal{R}^{\mathrm{m}})$. This dependence of p^{m} on $\mathcal{S}^{\mathrm{e}}$ and $\mathcal{A}^{\mathrm{e}}$ captures one aspect of the interplay between the monitor and the environment. In Mon-MDPs, there also exists another mapping f^{m} called the monitor function (not to be confused with the monitor itself). $f^{\mathrm{m}} : \mathcal{R}^{\mathrm{e}} \times \mathcal{S}^{\mathrm{m}} \times \mathcal{A}^{\mathrm{m}} \to \left(\widehat{\mathcal{R}}^{\mathrm{e}} \subset \mathbb{R}\right) \cup \{\bot\}$. The monitor function f^{m} models what reward the agent gets to observe and $\bot$ denotes a symbol when the agent does not receive any numerical feedback in return to the its action. Finally, the joint representation that includes the environment and the monitor comprises the finite joint state space $\mathcal{S} \coloneqq \mathcal{S}^{\mathrm{e}} \times \mathcal{S}^{\mathrm{m}}$, the finite joint action space $\mathcal{A} \coloneqq \mathcal{A}^{\mathrm{e}} \times \mathcal{A}^{\mathrm{m}}$, the finite joint set of rewards $\mathcal{R} \coloneqq \mathcal{R}^{\mathrm{e}} \times \mathcal{R}^{\mathrm{m}}$, the joint transition dynamics $p \coloneqq p^{\mathrm{e}} \otimes p^{\mathrm{m}}$, and the monitor function f^{m} as a tuple $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p, f^{\mathrm{m}} \rangle$.

In Mon-MDPs, it is sufficient to assume the agent follows a memoryless policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ to choose its actions. Following π in M results in a stochastic process $S_0, A_0, R_1, S_1, A_1, R_2, \ldots,$ where $R_{t+1} = (R^{\mathrm{e}}_{t+1}, R^{\mathrm{m}}_{t+1}), t \geq 0$. This policy induces a probability measure $\mathbb{P}$ over some sample space Ω. In this sample space $S_0, S_1, S_2, \cdots : \Omega \to \mathcal{S}$, $A_0, A_1, A_2, \cdots : \Omega \to \mathcal{A}$, and $R_1, R_2, R_3 \cdots : \Omega \to \mathcal{R}$, and Markov property, Equation (2.1), holds. Note that $S_0, A_0, R_1, S_1, A_1, R_2, \ldots$ is the true underlying stochastic process associated with the interaction, but the instead of R^{e}_{t+1} the agent observes the output of f^{m}. Let $\widehat{R}^{\mathrm{e}}_{t+1}$ be the output of f^{m} at time step t, where $\widehat{R}^{\mathrm{e}}_1, \widehat{R}^{\mathrm{e}}_2, \cdots : \Omega \to$

$\widehat{\mathcal{R}}^{\text{e}} \cup \{\bot\}$, then Figure 2.3 illustrates the agent-environment-monitor interaction in Mon-MDPs.

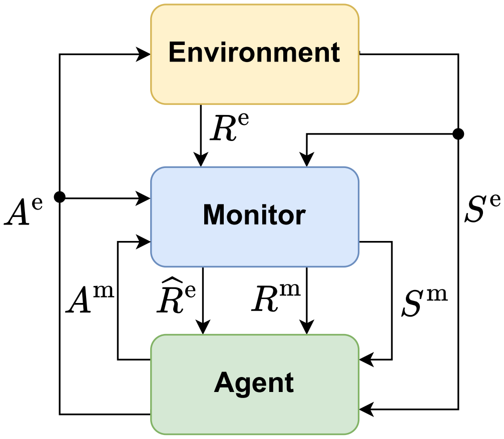


Figure 2.3: Mon-MDPs. The agent interacts not only with the environment, but also with the monitor. Upon taking the joint action $(A^{\text{e}}, A^{\text{m}})$, the agent receives the state of the environment S^{e} and the state of the monitor S^{m}. The monitor receives the environment reward R^{e}, but instead of passing it on, the monitor hands off $\widehat{R}^{\text{e}}$ to the agent. The agent also receives the monitor reward R^{m}. For the sake of clarity, we have omitted the dependence on time from the notation.

2.4.1 Learning Objective in Mon-MDPs

Mon-MDPs model the interaction between the agent, the environment, and the monitor. In this section, we formally revisit the agent's goal during this interaction. The agent is expected to behave optimally with respect to the environment and the monitor *simultaneously*, hence the objective incorporates the maximization of the environment and monitor rewards' sum. Let $\mathbb{E}$ denote the expectations we get with respect to $\mathbb{P}$. Then, using the criterion of maximizing the expected sum of discounted rewards with the discounted factor $0 \leq \gamma < 1$, the state-value and action-value functions of a policy π in Mon-MDP M are

$$V^{\pi}_{M}(s) \coloneqq \mathbb{E}\left[\sum_{k \geq t} \gamma^{k-t}\left(R^{\text{e}}_{k+1} + R^{\text{m}}_{k+1}\right)\middle| S_t = s\right], \quad \forall s \in \mathcal{S},$$

$$Q^{\pi}_{M}(s,a) \coloneqq \mathbb{E}\left[\sum_{k \geq t} \gamma^{k-t}\left(R^{\text{e}}_{k+1} + R^{\text{m}}_{k+1}\right)\middle| S_t = s, A_t = a\right], \quad \forall s, a \in \mathcal{S} \times \mathcal{A}.$$

Note that *even though the agent observes $\widehat{R}^{\text{e}}_{k+1}$ in place of R^{e}_{k+1} for all $k > 0$, the state-value and the action-value functions use R^{e}_{k+1}*. This is the crucial difference between MDPs and Mon-MDPs: the immediate environment reward R^{e}_{k+1} is always generated by the environment, i.e., desired behavior is well-defined. However, the monitor may "hide it" from the agent, possibly even always yielding "unobservable reward" $\widehat{R}^{\mathrm{e}}_{k+1} = \bot$ at all times k for some state-action pairs. For example, consider a task where the reward is given by a human supervisor (the monitor): if the supervisor must leave, the agent will not observe any reward; yet, the task has not changed, i.e., the human — if present — would still give the same rewards. Also similar to MDPs, for memoryless policies we have

$$V^{\pi}_{M}(s) = \sum_{a} \pi(a \mid s) Q^{\pi}_{M}(s, a), \qquad s \in \mathcal{S}, \tag{2.14}$$

i.e., the state-value function is the expected action-value function under the policy's randomness.

The use of the underlying environment reward instead of the reward that the agent actually gets to see could impose serious problems on the agent's learning. Note that the definition of f^{m} does not restrict this function in terms of what real values it should return, if at all ($\widehat{\mathcal{R}}^{\mathrm{e}}$ could be any measurable subset of real numbers). Thus, to rule out pathological cases (such as when $\widehat{R}^{\mathrm{e}}_{k+1} = -R^{\mathrm{e}}_{k+1}$), the monitor function f^{m} is assumed to be *truthful.*

Definition 2 (Parisi et al. [40, Property 3])**.** *f^{m} is truthful, if for $t \geq 0, \widehat{R}^{\mathrm{e}}_{t+1} \in \left\{R^{\mathrm{e}}_{t+1}, \bot\right\}$.*

A truthful monitor guarantees the environment reward would not be observed, unless it is equal to the true environment reward. Therefore, if the agent's goal, similar to Section 2.1.1, is finding a policy π^* such that $V^{\pi^*}_{M}(s) = \max_{\pi'} V^{\pi'}_{M}(s)$ for all joint states, in cases that $\widehat{R}^{\mathrm{e}}_{t+1} = \bot$ at all time steps for a particular environment state-action, finding π^* is impossible. This impossibility is because the agent would never and under no circumstances observe the environment reward. Even if the monitor shows the environment reward a finite number of times, it is impossible to find π^* almost surely. In such cases that the monitor does not reveal the reward for all environment state-action infinitely often, it is said that the monitor is *non-ergodic*, otherwise *ergodic.*

Definition 3 (Parisi et al. [40, Property 2])**.** *f^{m} is ergodic if, under infinite exploration, it returns a real value infinitely often for every environment state-action pair, i.e., $\forall (s^{\mathrm{e}}, a^{\mathrm{e}}) \in \mathcal{S}^{\mathrm{e}} \times \mathcal{A}^{\mathrm{e}}, \exists (s^{\mathrm{m}}, a^{\mathrm{m}}) \in \mathcal{S}^{\mathrm{m}} \times \mathcal{A}^{\mathrm{m}}$ such that for $s \coloneqq (s^{\mathrm{e}}, s^{\mathrm{m}})$ and $a \coloneqq (a^{\mathrm{e}}, a^{\mathrm{m}})$ the following holds:*

$$\mathbb{P}\left(\limsup_{t\to\infty} \widehat{R}^{\mathrm{e}}_{t+1} \neq \bot \middle| S_t = s, A_t = a\right) > 0.$$

Therefore, the traditional notion of optimality cannot be extended to Mon-MDPs when the monitor is non-ergodic, as optimality becomes unattainable in such cases. Parisi et al. [40] called such Mon-MDPs *unsolvable.* Intuitively, Parisi et al. [40] argued if the agent can never know that a certain state-action yields the highest (or lowest) environment reward, then the agent can never learn to visit (or avoid) that state-action. Nonetheless, assuming every environment reward is observable (sooner or later) is a stringent condition, not suitable for real-world tasks — reward instrumentation may have limited coverage, human supervisors may never be available in the evening, or training before deployment may not guarantee full state coverage. Hence, Parisi et al. [40] introduced a minimax-optimal objective that coincides with finding an optimal policy in solvable Mon-MDPs. Additionally, minimax-optimality empowers the agent to find an alternative policy to the optimal policy in unsolvable MDPs.

When the agent is interacting with a Mon-MDP with a non-ergodic function, there could be at least one environment state-action such that $\widehat{R}^{\text{e}}_{t+1}$ associated with that is always $\perp$. Thus, from the agent's perspective, there may be finitely many possible Mon-MDPs it could be interacting with. Each of these possible Mon-MDPs is associated with one distinct element of $\mathcal{R}^{\text{e}}$. Because any rewards could be the true underlying environment reward that is obscured by $\perp$. Therefore, the agent is facing a set of possible Mon-MDPs that it cannot distinguish them from each other. Let $[M]_{\mathbb{I}}$ be such a set (formal definition in Appendix A). If M is solvable, all environment rewards can be observed infinitely-often, thus $[M]_{\mathbb{I}} = \{M\}$. Otherwise, from the agent's perspective, there are possibly finitely-many Mon-MDPs in $[M]_{\mathbb{I}}$. Let $M_{\downarrow}$ be the worst-case Mon-MDP, i.e., the one where all never-observable rewards are $r^{\text{e}}_{\text{min}} = -r^{\text{e}}_{\text{max}} = -\max\{|r'| : r' \in \mathcal{R}^{\text{e}}\}$ (without loss of generality and to reduce the clutter in the theoretical part of this work, assume $r^{\text{e}}_{\text{max}} = 1$):

$$M_{\downarrow} \in \underset{M' \in [M]_{\mathbb{I}}}{\arg\min}\, r^{\text{e}}_{M'}(s^{\text{e}}, a^{\text{e}}), \qquad \forall (s^{\text{e}}, a^{\text{e}}) \in \mathcal{S}^{\text{e}} \times \mathcal{A}^{\text{e}}, \tag{2.15}$$

where $r^{\text{e}}_{M'}$ is the expected environment reward in Mon-MDP M'. In words, Equation (2.15) states that $M_{\downarrow}$ is a Mon-MDP whose expected environment reward is minimized over all Mon-MDPs indistinguishable from M. Then, the objective in Mon-MDPs is finding a *minimax-optimal policy* π in M as the optimal policy of the worst-case Mon-MDP the agent could possibly be facing, i.e.,

$$V^{\pi}_{M}(s) = \max_{\pi' \in \Pi} V^{\pi'}_{M_{\downarrow}}(s), \qquad \forall s \in \mathcal{S}. \tag{2.16}$$

where Π is the set of all policies in M (and $M_{\downarrow}$). We denote a minimax-optimal policy by $\pi^*_{\downarrow}$ and its corresponding state-value function with $V^*_{\downarrow}$. If M is solvable then $[M]_{\mathbb{I}} = \{M\}$, and Equation (2.16) and Equation (2.5) become equivalent, i.e., the minimax-optimal policy is simply the optimal policy.

Minimax-optimality advocates taking a *pessimistic* approach. Pessimism has already been suggested in the context of batch RL — where the agent cannot interact with the environment and has only access to static dataset— when the coverage of the offline dataset is insufficient [10, 21, 26, 44], online RL when models are imperfect [20] or the reward function is imprecise [45], safe RL when safety constraints impacts the exploration [1], and robotics to exercise caution [52].

2.4.2 Solutions for Mon-MDPs

We have defined the agent's goal as finding a policy that holds in Equation (2.16). In this section, we concisely revisit the Bellman optimality equation for Mon-MDPs that lays the ground for finding the minimax-optimal policies. Let $r^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}})$ be the expected environment reward, $r^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}})$ be expected monitor reward, and $p\,(s'|s, a)$ be joint next-state transition probability for the joint state-action $(s, a){\equiv}(s^{\mathrm{e}}, s^{\mathrm{m}}, a^{\mathrm{e}}, a^{\mathrm{m}})$. Define $\mathcal{E}_t$ to be the event where there exits a monitor state-action $(s^{\mathrm{m}}_0, a^{\mathrm{m}}_0)$ such that $\mathcal{E}_t = \left\{\widehat{R}^{\mathrm{e}}_{t+1} \neq \bot \middle| S^{\mathrm{e}}_t = s^{\mathrm{e}}, S^{\mathrm{m}}_t = s^{\mathrm{m}}_0, A^{\mathrm{e}}_t = a^{\mathrm{e}}, A^{\mathrm{m}}_t = a^{\mathrm{m}}_0\right\}$, i.e., the immediate environment reward is observed upon taking the joint action $(a^{\mathrm{e}}, a^{\mathrm{m}}_0)$ at the joint state $(s^{\mathrm{e}}, s^{\mathrm{m}}_0)$ at time step t. then

$$
\begin{aligned}
r^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}}) &= \begin{cases} \sum_{(r^{\mathrm{e}}_{\circ}, r^{\mathrm{m}}_{\circ})\in\mathcal{R}, s'\in\mathcal{S}} r^{\mathrm{e}}_{\circ} \cdot p\,(s', (r^{\mathrm{e}}_{\circ}, r^{\mathrm{m}}_{\circ})|s, a)\,, & \text{if } \mathbb{P}\left(\limsup_{t\to\infty} \mathcal{E}_t\right) > 0; \\ r_{\min} = -r_{\max}, & \text{otherwise}, \end{cases} \\
r^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}}) &= \sum_{(r^{\mathrm{e}}_{\circ}, r^{\mathrm{m}}_{\circ})\in\mathcal{R}, s'\in\mathcal{S}} r^{\mathrm{m}}_{\circ} \cdot p\left(s', (r^{\mathrm{e}}_{\circ}, r^{\mathrm{m}}_{\circ})\middle|s, a\right)\,, \\
p\left(s'\middle|s, a\right) &= \sum_{r'\in\mathcal{R}} p\left(s', r'\middle|s, a\right)\,.
\end{aligned}
$$

Then, similar to MDPs, it can be shown that $V^*_{\downarrow}$ satisfies the following Bellman optimality equation,

$$
V^*_{\downarrow}(s) = \max_a \left\{ r^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}}) + r^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}}) + \gamma \sum_{s'} p\left(s'\middle|s, a\right) V^*_{\downarrow}(s') \right\}. \tag{2.17}
$$

Therefore, similar to MDPs, for any initial value, iterative application of the value iteration converges to $V^*_{\downarrow}$. The minimax-optimal state-value $V^*_{\downarrow}$ defined in Equation (2.17) with Equation (2.14) that shows the relationship between state-value and action-value functions, immediately results in

$$
\begin{aligned}
Q^*_{\downarrow}(s, a) &= r^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}}) + r^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}}) + \gamma \sum_{s'} p\left(s'\middle|s, a\right) V^*_{\downarrow}(s'), \qquad (2.18) \\
V^*_{\downarrow}(s) &= \max_a Q^*_{\downarrow}(s, a), \\
\pi^*(s) &= \arg\max_a Q^*_{\downarrow}(s, a).
\end{aligned}
$$

However, no prior work similar to MBIE-EB exists that learns models of $r^{\mathrm{e}}, r^{\mathrm{m}}$, and p to perform value iteration and find $\pi^*_{\downarrow}$ associated with $Q^*_{\downarrow}$. All of the previous work used the ε-greedy exploration strategy to learn r^{e} and then directly estimated $Q^*_{\downarrow}$ without learning r^{m} or p [39, 40]. They all considered only Mon-MDPs with an ergodic monitor function. Hence, in those work's settings $\pi^*_{\downarrow}$ is equal to π^*. This means previous work were only concerned with solvable Mon-MDPs that, as was mentioned before, do not highlight the power of Mon-MDPs. The power of Mon-MDPs is to model scenarios that the agent might never get feedback for particular state-actions. Hence, previous work did not need their agents pursue the notion of minimax-optimality instead the traditional optimality. This simplification undermined the need for Mon-MDPs' introduction.

2.5 The Baseline: Directed Exploration-Exploitation

In this section, we describe Directed-Exploration-Exploitation (Directed-E^2) [39] as the baseline algorithm we use in our experiments. Parisi et al. [39] showed that Directed-E^2 outperforms conventional exploration strategies including ε-greedy, pure optimistic initialization, ε-greedy with count-based bonus [6], ε-greedy with upper confidence bound (UCB) bonus [3, 29], and ε-greedy with long-term UCB [38] on 48 finite Mon-MDPs. Directed-E^2 is the most performant algorithm on Mon-MDPs developed so far. It uses two estimates. The first is the ordinary action-values that uses a reward model, similar to Equation (3.1) in place of the immediate environment reward. However, Directed-E^2 does not use the pessimistic part of Equation (3.1). This reward model sets the agent free from the partial observability of the environment reward once the reward model is learned. The second action-value tries to maximize the successor representations. This successor representations maximizer denoted by Ψ, is dubbed as visitation-values. Directed-E^2 uses visitation-values for exploration. It uses visitation-values to keep the visitation of every joint state-action in a comparable range guaranteeing that every state-action is visited sufficiently. Consequently, since Directed-E^2 visits every state-action infinitely-often, every knowable quantity can be learned. In the limit of infinite exploration, Directed-E^2 becomes greedy with respect to the task respecting action-values for maximizing the expected of sum of discounted rewards. The intensity of trying to keep the visitation counts in the same order is determined by the hyperparameter $\bar{\beta} > 0$; the lower $\bar{\beta}$ is, the more frequently the agent tries to visit the least visited state-action. Algorithm 1 shows the Directed-E^2's pseudocode.

Algorithm 1 Directed-E^2

```
$t = 0$ // Total time steps
for episodes $k \coloneqq 1, 2, \ldots$ do
    for steps $h \coloneqq 1, 2, \ldots$ do
        $(s^{\mathrm{g}}, a^{\mathrm{g}}) = \arg\min_{s,a} N(s, a)$
        $\beta_t = \frac{\log t}{N(s^{\mathrm{g}}, a^{\mathrm{g}})}$
        if $\beta_t > \bar{\beta}$ then
            $A_h = \arg\max_a \Psi(a \mid S_h, s^{\mathrm{g}}, a^{\mathrm{g}})$ // Explore
        else
            $A_h = \arg\max_a Q(S_h, a)$ // Exploit
        $t \coloneqq t + 1$
        Perform action $A_h$
```

In summary, Directed-E^2 at each time step first ensures that all state-actions' visitation are comparable. If this condition holds, Directed-E^2 acts greedy with respect to action-values that maximize the accumulation of discounted rewards. In summary, in this chapter we started by introducing MDPs as the traditional model to formalize the agent-environment interaction. We revisited the agents' goal in MDPs as finding a memoryless deterministic policy maximizing the state-value function. We briefly discussed the Bellman optimality equation which gives rise to the value iteration as an iterative procedure to find an optimal policy. However, the Bellman optimality equation assumes having access to the true model of the environment. To address this unrealistic assumption, we revisited MBIE-EB. We explained that sample complexity is one of the properties to measure the efficiency of algorithms. MBIE-EB guaranteed with high-probability (using empirical models) near-optimal policies can be found with polynomial sample complexity. Then, we introduced Mon-MDPs that formalize the interaction of the agent not only with the environment but also with the monitor. The monitor represents the process that provides the agent with the reward. We discussed that due to potential unavailability of the monitor, the agent might be unable to receive reward for some state-actions. This inaccessibility to the reward made the notion of optimality in MDPs inapplicable in Mon-MDPs. As a results, we introduced the notion of minimax-optimality for Mon-MDPs. We defined a minimax-optimal policy for any unobserved reward, assumes the minimum possible value. Finally, while revisiting Directed-E^2 as the SOTA algorithm in Mon-MDPs, we pointed out that, no algorithm similar to MBIE-EB exits for Mon-MDPs.

Chapter 3

Methodology

In this chapter we introduce Monitored MBIE-EB as an extension of the MBIE-EB's idea to Mon-MDPs. Further, we extend the notion of PAC-MDP to PAC-Mon-MDP to specify the sample complexity of Monitored MBIE-EB.

3.1 MBIE-EB in Mon-MDPs

Since MBIE-EB uses MLEs to build its models, in order to extend the idea of MBIE-EB to Mon-MDPs, in this section, we first define models in Mon-MDPs. Then, we introduce appropriate bonuses that would extend MBIE-EB to Mon-MDPs. Remember from Equation (2.18), in Mon-MDPs the minimax-optimal action-value for a fixed joint state-action $(s, a) \equiv (s^{\mathrm{e}}, s^{\mathrm{m}}, a^{\mathrm{e}}, a^{\mathrm{m}})$ is

$$Q^*_{\downarrow}(s, a) = r^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}}) + r^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}}) + \gamma \sum_{s'} p\left(s' \middle| s, a\right) V^*_{\downarrow}(s').$$

Define $N(s^{\mathrm{e}}, a^{\mathrm{e}}), N(s^{\mathrm{m}}, a^{\mathrm{m}})$, and $N(s, a)$ as the number of times the environment reward at $(s^{\mathrm{e}}, a^{\mathrm{e}})$ has been *observed*, the number of times the monitor reward at $(s^{\mathrm{m}}, a^{\mathrm{m}})$ has been observed, and the number of times the joint state-action (s, a) had been *visited*. The MLEs of $r^{\mathrm{e}}, r^{\mathrm{m}}$ and p are

$$\begin{aligned}
\bar{r}^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}}) &= \begin{cases} \frac{1}{N(s^{\mathrm{e}},a^{\mathrm{e}})} \sum_{i=1}^{N(s^{\mathrm{e}},a^{\mathrm{e}})} R^{\mathrm{e}}_i, & N(s^{\mathrm{e}}, a^{\mathrm{e}}) \neq 0, \\ r^{\mathrm{e}}_{\text{min}} = -r^{\mathrm{e}}_{\text{max}}, & \text{otherwise}, \end{cases} \qquad (3.1) \\
\bar{r}^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}}) &= \frac{1}{N(s^{\mathrm{m}}, a^{\mathrm{m}})} \sum_{j=1}^{N(s^{\mathrm{m}},a^{\mathrm{m}})} R^{\mathrm{m}}_j, \\
\bar{p}\left(s' \middle| s, a\right) &= \frac{1}{N(s, a)} \sum_{k=1}^{N(s,a)} \mathbb{I}\left\{S'_k = s'\right\}, \qquad \forall s' \in \mathcal{S},
\end{aligned}$$

where $R^{\mathrm{e}}_i, R^{\mathrm{m}}_j$, and S'_k are the ith observed immediate environment reward, the jth observed immediate monitor reward and the next joint state after the kth visit. $\mathbb{I}$ is the indicator function returning one if the predicate of its argument is true and zero otherwise. These MLEs results in following sample estimate of $Q^*_\downarrow(s,a)$ denoted by $\bar{Q}^*_\downarrow(s,a)$,

$$\bar{Q}^*_\downarrow(s,a) \coloneqq \bar{r}^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}}) + \bar{r}^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}}) + \gamma \sum_{s'} \bar{p}(s' \mid s,a) V^*_\downarrow(s').$$

In MDPs, for a fixed (s,a) the expected immediate reward $r(s,a)$ and the discounted expected next-state's value $\gamma \sum_{s'} p(s' \mid s,a)V^*(s')$ are both mappings from the same space, $\mathcal{S} \times \mathcal{A}$. However, in Mon-MDPs the expected immediate environment reward $r^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}})$, the expected immediate monitor reward $r^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}})$, and $\gamma \sum_{s'} p(s' \mid s,a)V^*_\downarrow(s')$ are all mappings from different input spaces. $r^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}})$ is a mapping from the environment state-action space, $r^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}})$ is a mapping from the monitor state-action space, and $\gamma \sum_{s'} p(s' \mid s,a)V^*(s')$ is a mapping from the joint state-action space. As a result, empirical counts for each of these mappings possibly increase at a different rate. We cannot bound the deviation of $\bar{Q}^*_\downarrow(s,a)$ from its expected value $Q^*_\downarrow(s,a)$ by a single application of the Chernoff-Hoeffding's bound as we did in Equation (2.11). Nevertheless, using a union bound argument, we will show later the required optimism in $Q^{**}_\downarrow(s,a)$, which is the optimistic action-values analogous to Equation (2.13), defined in Equation (3.2), with probability at least $1-\delta$ holds. In Equation (3.2), m is the least number of samples required per state-action until their MLEs are sufficiently close to their true values

$$\begin{aligned} Q^{**}_\downarrow(s,a) = \bar{r}^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}}) + \bar{r}^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}}) + \gamma \sum_{s'} \bar{p}(s' \mid s,a) V^{**}_\downarrow(s') + \\ \frac{\beta^{\mathrm{e}}}{\sqrt{N(s^{\mathrm{e}}, a^{\mathrm{e}})}} + \frac{\beta^{\mathrm{m}}}{\sqrt{N(s^{\mathrm{m}}, a^{\mathrm{m}})}} + \frac{\beta}{\sqrt{N(s,a)}}, \end{aligned} \tag{3.2}$$

where

$$V^{**}_\downarrow(s') = \max_{a'} Q^{**}(s',a'), \quad \beta^{\mathrm{e}} = \sqrt{2\ln\left(\frac{12\,|\mathcal{S}|\,|\mathcal{A}|\,m}{\delta}\right)}, \quad \beta^{\mathrm{m}} = \sqrt{2\ln\left(\frac{12\,|\mathcal{S}|\,|\mathcal{A}|\,m}{\delta}\right)}, \text{ and}$$
$$\beta = \frac{2\gamma}{1-\gamma}\sqrt{2\ln\left(\frac{12\,|\mathcal{S}|\,|\mathcal{A}|\,m}{\delta}\right)}.$$

In summary, we introduced models $\bar{r}^{\mathrm{e}}, \bar{r}^{\mathrm{m}}, \bar{p}$, and bonuses which extends MBIE-EB to Mon-MDPs.

3.2 Observation Stage

MBIE-EB directly uses optimistic initial values to cover cases that the visitation count $N(s, a)$ is zero for state-action (s, a). However, incorporating $\bar{r}^{\mathrm{e}}$ defined in Equation (3.1) into $Q_{\downarrow}^{**}$, defined in Equation (3.2), hinders the optimism required for $Q_{\downarrow}^{**}$. Because when the count $N^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}})$ is zero, the pessimistic value of $-r_{\max}^{\mathrm{e}}$ is used, not the optimistic initial values. For example, back to our plant-watering robot example in Figure 1.1, if the robot waters a large flower pot (which is the correct action) during the training but due to some circumstances the owner cannot give the reward to the robot, the robot becomes prematurely pessimistic about watering the large flower pot. Therefore, we want the robot to explore more to ensure that it will be given feedback infinitely-often, not based on a random one-off event.

In this section, we describe our solution to this challenge. Algorithms based on the principle of optimism in the face of uncertainty (OFU) [4, 9, 18, 47] prove the maintenance of the optimism through induction. The base case of the induction, which corresponds to having zero counts, is proved using optimistic initialization. However, when $N(s^{\mathrm{e}}, a^{\mathrm{e}})$ is zero, Equation (3.1) prescribes assigning a pessimistic value to $\bar{r}^{\mathrm{e}}$ instead of using optimistic initial values to define $Q_{\downarrow}^{**}$. This is problematic in cases that the observability of the environment reward is stochastic. If the agent does not get to observe the environment in its first visit, then it becomes pessimistic about the unobserved reward (possibly for ever).

Our response to the challenge of becoming prematurely pessimistic is intuitively described as having a "cautiously optimistic" agent. Before asking the agent to follow $Q_{\downarrow}^{**}$, the agent should first undergo an additional stage to assess the observability of the environment rewards. This stage tries to determine, with high probability, whether the environment reward for each environment state-action is observable or not. Then based on the findings of this stage, the agent is able to use $\bar{r}^{\mathrm{e}}$ such that with high probability the assigned pessimism is actually for environment rewards that are effectively never-observable. Otherwise, the principle of optimism can be safely used. We call this stage as the *observation* stage.

In the observation stage, the agent transforms the underlying Mon-MDP $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p, f^{\mathrm{m}} \rangle$ it is facing, into an MDP $\widetilde{M} = \langle \mathcal{S}, \mathcal{A}, \{0, 1\}, \widetilde{p} \rangle$. In MDP $\widetilde{M}$, $\mathcal{S}$ is the state space, which is the joint state space of M, $\mathcal{A}$ is the action space, which is the joint action space of M, the set of rewards is $\{0, 1\}$, and $\widetilde{p}$ is the transition dynamics. At time step t the immediate reward $\widetilde{R}_{t+1}$ for the joint state-action $(S_t, A_t) \equiv (S_t^{\mathrm{e}}, S_t^{\mathrm{m}}, A_t^{\mathrm{e}}, A_t^{\mathrm{m}})$ is defined as

$$\widetilde{R}_{t+1} = \mathbb{I}\left\{ \widehat{R}_{t+1}^{\mathrm{e}} \neq \perp \middle| S_t, A_t \right\} \cdot \mathbb{I}\left\{ N\left(S_t^{\mathrm{e}}, A_t^{\mathrm{e}}\right) = 0 \right\}. \tag{3.3}$$

Maximizing this reward corresponds to selecting actions $a \in \mathcal{A}$ that results in observing the environment reward in states $s \in \mathcal{S}$ where their environment reward has not been observed *so far*, i.e., $N\left(S_t^{\mathrm{e}}=s^{\mathrm{e}}, A_t^{\mathrm{e}}=a^{\mathrm{e}}\right)=0$. This stage is only concerned with discovering which environment rewards are observable, it only focuses on state-actions that their environment reward has not been observed so far. Otherwise, the agent has figured out whether environment reward for which particular state-actions is observable and, upon enough visits, a good sample estimate of the environment reward's mean can be computed. The term $\mathbb{I}\left\{N\left(S_t^{\mathrm{e}}, A_t^{\mathrm{e}}\right)=0\right\}$ in Equation (3.3) introduces non-stationarity to the problem as once the environment reward for $(S_t^{\mathrm{e}}=s^{\mathrm{e}}, A_t^{\mathrm{e}}=a^{\mathrm{e}})$ is observed, the value of the indicator function flips. We argue that this non-stationarity eventually washes out. The agent would observe everything that is observable and anything that is never-observable will remain never-observable. Also, when $N(S_t^{\mathrm{e}}=s^{\mathrm{e}}, A_t^{\mathrm{e}}=a^{\mathrm{e}})$ is positive, it will remain positive, thus the value of $\mathbb{I}\left\{N(S_t^{\mathrm{e}}=s^{\mathrm{e}}, A_t^{\mathrm{e}}=a^{\mathrm{e}})=0\right\}$ would result in a deterministic reward of zero because $N(S_t^{\mathrm{e}}=s^{\mathrm{e}}, A_t^{\mathrm{e}}=a^{\mathrm{e}})$ would always be bigger than zero and the argument inside the indicator function is false. This deterministic reward makes the optimization in $\widetilde{M}$ easier. In the worst-case scenario, $\mathbb{I}\left\{N(S_t^{\mathrm{e}}=s^{\mathrm{e}}, A_t^{\mathrm{e}}=a^{\mathrm{e}})=0\right\}$ would be one for all state-actions.

In order to make the transformation from M to $\widetilde{M}$ in the observation stage more tangible, consider the example of Figure 3.1. In this example, the environment has only a single state corresponding to the cell that the agent is located at. The agent has four cardinal actions {`LEFT, DOWN, RIGHT, UP`}, which all of them result in staying in the current state and an environment reward of zero. Moreover, the agent has two monitor actions `ASK, NOT ASK`. Each cardinal action with `ASK`ing is denoted by $\Longrightarrow$ and each cardinal action with `NOT ASK`ing is denoted by $\longrightarrow$. If

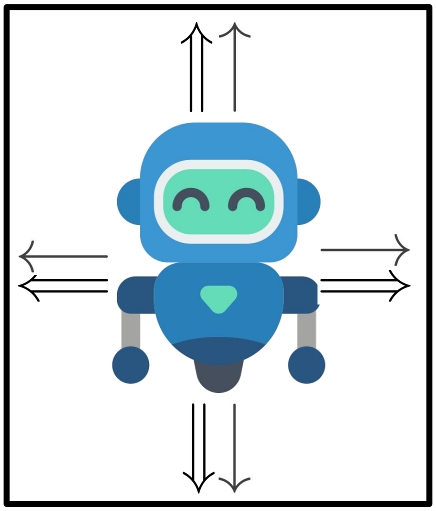


Figure 3.1: An example of a Mon-MDP. The agent has four cardinal actions and two separate monitor actions to `ASK` or `NOT ASK` for reward. Each cardinal action with `ASK`ing is denoted by $\Longrightarrow$ and each cardinal action with `NOT ASK`ing is denoted by $\longrightarrow$. If the agent `ASK`s, it observes the reward for the taken cardinal action. But, by `ASK`ing the agent also pays a cost. If the agent does `NOT ASK`, it does not observe the reward for the taken cardinal action and does not pay any costs.

the agent moves and also `ASK`s, then it will observe the environment reward of zero. If it does `NOT ASK`, then it will observe $\perp$. Suppose at the current time step t the agent selects the `RIGHT` action and `ASK`s. Also, suppose at a time step $t' < t$, the agent had already `ASK`ed and consequently had observed the reward of zero *only* for going `RIGHT`. Then, $\widetilde{R}_{t+1}$ in the associated $\widetilde{M}$ is

$$\widetilde{R}_{t+1} = \begin{array}{|c|c|c|c|c|c|c|c|} \hline \Longleftarrow & \Downarrow & \Longrightarrow & \Uparrow & \longleftarrow & \downarrow & \longrightarrow & \uparrow \\ \hline 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline \end{array}.$$

We can execute MBIE-EB on $\widetilde{M}$ to find and optimal policy seeking observability. For a state-action (s, a) define $p\left(s'|s, a\right) = \sum_{r' \in \{0,1\}} \widetilde{p}\left(s', r'|s, a\right), \forall s' \in \mathcal{S}$. Then, identical to Equation (2.9), we have

$$\bar{r}(s, a) = \frac{1}{N(s, a)} \sum_{i=1}^{N(s,a)} \widetilde{R}_i, \qquad \bar{p}\left(s'|s, a\right) = \frac{1}{N(s, a)} \sum_{i=1}^{N(s,a)} \mathbb{I}\left\{S'_i = s'\right\}, \qquad \forall s' \in \mathcal{S}.$$

However, $\bar{r}(s, a)$ is zero for all state-actions. Because, for all state-actions that their environment reward has been observed $\mathbb{I}\left\{N(S^{\mathrm{e}}_t = s^{\mathrm{e}}, A^{\mathrm{e}}_t = a^{\mathrm{e}}) = 0\right\}$ is zero, and for all state-actions that their reward is yet to be observed, $\mathbb{I}\left\{\widehat{R}^{\mathrm{e}}_{t+1} \neq \perp \middle| S_t = s, A_t = a\right\}$ is zero. Hence, the optimistic action-values $\widetilde{Q}^{**}$ in $\widetilde{M}$, analogous to Equation (2.13), with probability at least $1 - \frac{\delta}{2}$ is

$$\widetilde{Q}^{**}(s, a) = \gamma \sum_{s'} \bar{p}(s' \mid s, a) \max_{a'} \widetilde{Q}^{**}(s', a') + (1 - \gamma)^{-1} \sqrt{\frac{\ln\left(\frac{2|\mathcal{S}||\mathcal{A}|m}{\delta}\right)}{2N(s, a)}}, \tag{3.4}$$

where m is the number of samples required until $\bar{p}$ is close to its true value, indicated by Lemma 2.3.

However, inspired by the derivation of $Q^{**}_{\downarrow}$ in Equation (3.2) to have different bonuses on the mean reward and the next state's discounted value, we use the upper confidence bound of $\bar{r}$. This use of the upper confidence bound makes the agent more optimistic with respect to the yet-to-be-observed rewards. More optimism means the agent is more eager to explore the observability of the yet-to-be-observed rewards. Since $\bar{r}$ is Bernoulli, KL-UCB [15, 34] is the suitable option to compute the upper confidence bound of $\bar{r}$. Hence, $\widetilde{Q}^{**}$ is turned into

$$\widetilde{Q}^{**}(s, a) = \text{KL-UCB}(0, N(s, a)) \cdot \mathbb{I}\left\{N(s^{\mathrm{e}}, a^{\mathrm{e}}) = 0\right\} + \gamma \sum_{s'} \bar{p}(s' \mid s, a) \max_{a'} \widetilde{Q}^{**}(s', a') + \frac{\beta^{\text{obs}}}{\sqrt{N(s, a)}}, \tag{3.5}$$

where KL-UCB$(\bar{\mu}, n) =$

$$\max_{\mu} \left\{\mu \in [0, 1] : d(\bar{\mu}, \mu) \leq \frac{\beta^{\text{KL-UCB}}}{n}\right\}\Bigg|_{\bar{\mu}=0} = \max_{\mu} \left\{\mu \in [0, 1] : \ln\left(\frac{1}{1 - \mu}\right) \leq \frac{\beta^{\text{KL-UCB}}}{n}\right\},$$

$\beta^{\text{KL-UCB}} = \ln\left(\frac{8|\mathcal{S}||\mathcal{A}|m}{\delta}\right)$, and $\beta^{\text{obs}} = (1-\gamma)^{-1}\sqrt{0.5\ln\left(\frac{8|\mathcal{S}||\mathcal{A}|m}{\delta}\right)}$.

In the KL-UCB's definition, d is the relative entropy between two Bernoulli distributions. Note that KL-UCB$(0, N(s,a))$ explicitly shows that the empirical probability of observing the environment reward *so far* has been zero. This use of the KL-UCB instead of the empirical mean $\bar{r}$, as used in the original MBIE-EB, mandates proving that $\widetilde{Q}^{**}$ remains optimistic, otherwise, other components of MBIE-EB is untouched. Lemma 3.1 gives us a high probability guarantee that $\widetilde{Q}^{**}$ is an optimistic action-value. The proof is provided in Appendix C.1.

Lemma 3.1. *Let $\widetilde{M} = \langle \mathcal{S}, \mathcal{A}, \{0,1\}, \widetilde{p}\rangle$ be an MDP obtained by transforming a truthful Mon-MDP using observability reward defined in Equation (3.3) and let $0 \leq \gamma < 1$. If*

$$\beta^{\textit{KL-UCB}} = \ln\left(\frac{8\,|\mathcal{S}|\,|\mathcal{A}|\,m}{\delta}\right) \quad \textit{and} \quad \beta^{obs} = (1-\gamma)^{-1}\sqrt{0.5\ln\left(\frac{8\,|\mathcal{S}|\,|\mathcal{A}|\,m}{\delta}\right)},$$

then $\widetilde{Q}^{**}(s,a) \geq \widetilde{Q}^{*}(s,a)$ with probability at least $1-\frac{\delta}{4}$ for all state-actions that have not been visited at least m times defined in Lemma 2.3.

In summary, in this section we introduced the observation stage, where the agent tries to determine the observability of the environment rewards with high probability. We showed that the agent in this stage follows $\widetilde{Q}^{**}$, as specified in Equation (3.5).

3.3 Monitored MBIE-EB

In this section, we explain our full proposed algorithm, *Monitored MBIE-EB*. Monitored MBIE-EB puts together the rationale of MBIE-EB and the observation stage into a single procedure. The agent operates in slices of episodes each with a maximum length of H. Monitored MBIE-EB uses $\kappa^*(k)$, where k is the episode counter, and $\kappa^* : \mathbb{N} \cup \{0\} \to \mathbb{R}$ is a sublinear function, to determine if the agent should go through the observation stage. Otherwise, if the agent does not use the observation stage, it attempts to obtain minimax-optimality by following the policy corresponding to $Q^{**}_{\downarrow}$. Algorithm 2 shows the pseudocode of Monitored MBIE-EB.

In summary, the insight behind Monitored MBIE-EB is seeking the observability of rewards while ensuring convergence to a near-minimax-optimal policy in polynomial time, or asymptotically, depending on how the agent schedules to enter the observation stage using $\kappa^*(k)$.

Algorithm 2 Monitored MBIE-EB

```
$\kappa \leftarrow 0$
for episodes $k \coloneqq 1, 2, 3, \dots$ do
    if $\kappa \leq \kappa^*(k)$ then
        // Observation Stage
        $Q \leftarrow \widetilde{Q}^{**}$ // Equation (3.5)
        $\kappa \leftarrow \kappa + 1$
    else
        $Q \leftarrow Q_{\downarrow}^{**}$ // Equation (2.13)
    for steps $h \coloneqq 1, 2, \dots, H$ do
        Follow the greedy policy with respect to $Q$.
```

3.3.1 Sample Complexity

Similar to Section 2.3 where the sample complexity of MBIE-EB was given, in this section, we determine the Monitored MBIE-EB's sample complexity as our measure of its efficiency. In Mon-MDPs partial observability of the environment reward naturally makes any algorithm take more time. The lower the probability is, the more samples are required to confidently approximate the statistics of the environment reward. As a result we extend the definition of PAC-MDP explained in Section 2.3 to PAC-Mon-MDP:

Definition 4. *An algorithm is PAC-Mon-MDP minimax-optimal if for any Mon-MDP $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p, f^{\mathrm{m}} \rangle$ with a truthful monitor function, discount factor $0 \leq \gamma < 1$ and any $\delta, \epsilon > 0$, the algorithm finds an ϵ-optimal policy in $M_{\downarrow}$, the worst-case Mon-MDP in the equivalence class of M, in time polynomial to $\left(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{\epsilon}, \ln\frac{1}{\delta}, \frac{1}{1-\gamma}, r^{\mathrm{e}}_{\max} + r^{\mathrm{m}}_{\max}, \frac{1}{\rho}\right)$, where $0 < \rho \leq 1$ is the minimum non-zero probability of observing the environment reward, embedded in f^{m}.*

Since the observation stage uses an instantiation of MBIE-EB, by virtue of Theorem 2.1 there exists a $k^* \in \widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|H}{\epsilon^3(1-\gamma)^5}\right)$ such that if $\kappa^*(k) = k^*$ (constant function), then the agent with high probability learns the observability status of each environment reward in polynomial time. This choice of $\kappa^*(k)$ gives Monitored MBIE-EB the possibility of becoming PAC-Mon-MDP conditioned on maintaining polynomial sample complexity after the agent has left the observation stage. A matter we will show is taken care of in Theorem 3.1, but before proving the theorem we state a series of useful lemmas in proving it.

Lemma 3.2 bounds the difference between a fixed policy's action-value in two Mon-MDPs M_1 and M_2 in terms of how much M_1's and M_2's immediate expected reward and transition dynamics are different from each other. It is useful in proving Lemma 3.3. Lemma 3.2 is the adaptation of the *Simulation Lemma* [23, 47] from MDPs to Mon-MDPs. The proof is provided in Appendix C.2.

Lemma 3.2. *Let $M_1 = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}_1, p_1, f^{\mathrm{m}} \rangle$ and $M_2 = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}_2, p_2, f^{\mathrm{m}} \rangle$ be two Mon-MDPs and $0 \leq \gamma < 1$. Assume the following bounds hold*

$$-r^{\mathrm{e}}_{\max} \leq r^{\mathrm{e}}_1, r^{\mathrm{e}}_2 \leq r^{\mathrm{e}}_{\max} \quad \textit{and} \quad -r^{\mathrm{m}}_{\max} \leq r^{\mathrm{m}}_1, r^{\mathrm{m}}_2 \leq r^{\mathrm{m}}_{\max} \ .$$

Also, assume the following conditions hold for all joint state-actions $(s, a) \equiv (s^{\mathrm{e}}, s^{\mathrm{m}}, a^{\mathrm{e}}, a^{\mathrm{m}})$,

$$\begin{aligned} |r^{\mathrm{e}}_1(s^{\mathrm{e}}, a^{\mathrm{e}}) - r^{\mathrm{e}}_2(s^{\mathrm{e}}, a^{\mathrm{e}})| &\leq \varphi^{\mathrm{e}}, \\ |r^{\mathrm{m}}_1(s^{\mathrm{m}}, a^{\mathrm{m}}) - r^{\mathrm{m}}_2(s^{\mathrm{m}}, a^{\mathrm{m}})| &\leq \varphi^{\mathrm{m}}, \ \textit{and} \\ \|p_1(\cdot \mid s, a) - p_2(\cdot \mid s, a)\|_1 &\leq \varphi, \end{aligned}$$

then for any stationary deterministic policies π it holds that

$$|Q^{\pi}_1(s, a) - Q^{\pi}_2(s, a)| \leq \frac{\varphi^{\mathrm{e}} + \varphi^{\mathrm{m}} + 2\varphi\gamma(r^{\mathrm{e}}_{\max} + r^{\mathrm{m}}_{\max})}{(1-\gamma)^2}.$$

Monitored MBIE-EB uses empirical models. Lemma 3.3 demonstrates that a fixed policy π exhibits similar action-values in two Mon-MDPs with comparable transition dynamics and immediate expected rewards. Consequently, an agent achieves near-desired behavior by ensuring the accuracy of its built models. Later we will use this lemma to measure how many samples are required to obtain accurate models. The proof is provided in Appendix C.3.

Lemma 3.3. *Let $M_1 = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}_1, p_1, f^{\mathrm{m}} \rangle$ and $M_2 = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}_2, p_2, f^{\mathrm{m}} \rangle$ be two Mon-MDPs and $0 \leq \gamma < 1$. Assume the following bounds hold*

$$-r^{\mathrm{e}}_{\max} \leq r^{\mathrm{e}}_1, r^{\mathrm{e}}_2 \leq r^{\mathrm{e}}_{\max} \quad \textit{and} \quad -r^{\mathrm{m}}_{\max} \leq r^{\mathrm{m}}_1, r^{\mathrm{m}}_2 \leq r^{\mathrm{m}}_{\max} \ .$$

Suppose further for all joint state-actions $(s, a) \equiv (s^{\mathrm{e}}, s^{\mathrm{m}}, a^{\mathrm{e}}, a^{\mathrm{m}})$ the following are satisfied,

$$|r^{\mathrm{e}}_1(s^{\mathrm{e}}, a^{\mathrm{e}}) - r^{\mathrm{e}}_2(s^{\mathrm{e}}, a^{\mathrm{e}})| \leq \varphi^{\mathrm{e}}, \quad |r^{\mathrm{m}}_1(s^{\mathrm{m}}, a^{\mathrm{m}}) - r^{\mathrm{m}}_2(s^{\mathrm{m}}, a^{\mathrm{m}})| \leq \varphi^{\mathrm{m}}, \quad \|p_1(\cdot \mid s, a) - p_2(\cdot \mid s, a)\|_1 \leq \varphi.$$

There exists a constant C such that for any $0 < \epsilon \leq \frac{(r^{\mathrm{e}}_{\max} + r^{\mathrm{m}}_{\max})}{1-\gamma}$, and any stationary policy π, if $\varphi^{\mathrm{e}} = \varphi^{\mathrm{m}} = \varphi = C\frac{\epsilon(1-\gamma)^2}{r^{\mathrm{e}}_{\max} + r^{\mathrm{m}}_{\max}}$, then

$$|Q^{\pi}_1(s, a) - Q^{\pi}_2(s, a)| \leq \epsilon \ .$$

Lemma 3.4 determines the number of visits to each state-action to have accurate estimates of the transition dynamics and the immediate mean reward. The proof is provided in Appendix C.4.

Lemma 3.4. *Let $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p, f^{\mathrm{m}} \rangle$, $0 \leq \gamma < 1$ and ρ be the minimum non-zero probability of observing the environment reward in M. For $\tau = C\frac{\epsilon(1-\gamma)^2}{r^{\mathrm{e}}_{\max} + r^{\mathrm{m}}_{\max}}$, where C is a constant specified in*

Lemma 3.3, if $\rho^{-1} > \mathcal{O}(|\mathcal{S}|)$, then there exists an m,

$$m \in \mathcal{O}\left(\frac{1}{\rho\tau^2}\ln\frac{|\mathcal{S}|\,|\mathcal{A}|}{\tau\delta}\right) = \mathcal{O}\left(\frac{1}{\rho\epsilon^2(1-\gamma)^4}\ln\frac{|\mathcal{S}|\,|\mathcal{A}|}{\epsilon(1-\gamma)^2\delta}\right),$$

and if $\rho^{-1} < \mathcal{O}(|\mathcal{S}|)$, then there exists an m,

$$m \in \mathcal{O}\left(\frac{|\mathcal{S}|}{\tau^2} + \frac{1}{\tau^2}\ln\frac{|\mathcal{S}|\,|\mathcal{A}|}{\tau\delta}\right) = \mathcal{O}\left(\frac{|\mathcal{S}|}{\epsilon^2(1-\gamma)^4} + \frac{1}{\epsilon^2(1-\gamma)^4}\ln\frac{|\mathcal{S}|\,|\mathcal{A}|}{\epsilon(1-\gamma)^2\delta}\right),$$

such that with probability at least $1-\delta$, $|\bar{r}(s,a) - r(s,a)| \leq \tau$ and $\|\bar{p}(\cdot \mid s,a) - p(\cdot \mid s,a)\|_1 \leq \tau$ hold for all state-actions that have been visited at least m times.

Lemma 3.5 describes the difference in a policy's state-value between two distinct Mon-MDPs, given that their transition dynamics and rewards are identical on certain state-actions (those in set $\mathcal{K}$), and arbitrarily different on the other state-actions. If the difference in the value of the same policy between these two Mon-MDPs is large, the probability of reaching a state that differentiates the two Mon-MDPs is also high. Lemma 3.5 is the adaptation of the *Induced Inequality* [23, 47] from MDPs to Mon-MDPs. The proof is provided in Appendix C.5.

Lemma 3.5. *Let M be a Mon-MDP, $\mathcal{K}$ a set of state-actions, M' a Mon-MDP equal to M on $\mathcal{K}$ (identical transition and reward function), π a policy, and H be some positive integer. Let $\mathcal{E}_M$ be the event that a state-action not in $\mathcal{K}$ is encountered in a trial generated by starting from S_0 and following π for H steps in M, then*

$$V^{\pi}_{M}(S_0)_H \geq V^{\pi}_{M'}(S_0)_H - \frac{\mathbb{P}(\mathcal{E}_M)}{2(1-\gamma)}.$$

Lemma 3.6 shows Monitored MBIE-EB exhibits the principle of OFU in the worst-case Mon-MDP the agent could be facing during the interaction. Optimism incentives the agent to visit state-actions, where the estimates are not accurate. The proof is provided in Appendix C.6

Lemma 3.6. *Let M be any truthful Mon-MDP. With probability at least $1-\frac{5\delta}{6}$, $Q^{**}_{\downarrow}(s,a) \geq Q^{*}_{\downarrow}(s,a)$ for all joint state-actions that have not been visited at least m times.*

Theorem 3.1 is our main result for Monitored MBIE-EB describing its sample complexity.

Theorem 3.1. *Suppose that ϵ, and δ are two real numbers between zero and one. Suppose $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p, f^{\mathrm{m}} \rangle$ is any truthful Mon-MDP and ρ is the minimum non-zero probability of observing the environment reward in M. Let $0 \leq \gamma < 1$ be the discount factor. There exist an input m as in Lemma 3.4 such that if Monitored MBIE-EB is executed on M and value iteration is done every $H \in \mathcal{O}\left((1-\gamma)^{-1}\ln\frac{1}{\epsilon(1-\gamma)}\right)$ time steps for H iterations, then the following holds. Let π_t denote* *the policy of Monitored MBIE-EB at time t and S_t denote the state at time t. With probability at least $1-\delta$, $V_{\downarrow}^{\pi_t}(S_t) \geq V_{\downarrow}^{*}(S_t) - \epsilon$ is true for all but $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\epsilon^3(1-\gamma)^6\rho}\right)$ time steps.*

Proof. Let $\epsilon_1 = \epsilon_2 = \frac{\epsilon}{2}$, and $\delta_1 = \delta_2 = \frac{\delta}{2}$. According to Theorem 2.1 there exists $k^* \in \widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{\epsilon_1^3(1-\gamma)^6}\right)$ such that if $\kappa^*(k) = k^*$, then the observability of environment rewards are determined with probability at least $1-\delta_1$. Let ϵ_3 be an arbitrary positive real number, whose precise value we will specify later. Let $H = \mathcal{O}\left((1-\gamma)^{-1} \ln \frac{1}{\epsilon_3(1-\gamma)}\right)$ be a positive integer large enough so that for all Mon-MDPs M' with discount factor γ, policies π and states s, the output of the value iteration for H steps, $V_{M'}^{\pi}(s)_H$, is ϵ_3 close to the true value $V_{M'}^{\pi}(s)$.

First, we argue that after each (s,a) has been experienced a polynomial number of times m, the empirical model learned from those experiences, $\bar{r}^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}}), \bar{r}^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}})$ and $\bar{p}(\cdot \mid s,a)$ will be sufficiently close to their true values $r^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}}), r^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}})$, and $p(\cdot \mid s,a)$, hence using models will result in a near-minimax-optimal policy. We want that the state-value function of any policy according to $\bar{r}^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}}), \bar{r}^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}})$, and $\bar{p}(\cdot \mid s,a)$ is no more than ϵ_3 away from its true value according to $r^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}}), r^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}})$, and $p(\cdot \mid s,a)$ (but otherwise the same), with high probability. It follows from Lemma 3.3 that it is sufficient to require $|\bar{r}^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}}) - r^{\mathrm{e}}(s^{\mathrm{e}}, a^{\mathrm{e}})| \leq \tau, |\bar{r}^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}}) - r^{\mathrm{m}}(s^{\mathrm{m}}, a^{\mathrm{m}})| \leq \tau$, and $\|\bar{p}(\cdot \mid s,a) - p(\cdot \mid s,a)\|_1 \leq \tau$ for $\tau = C\frac{\epsilon_3(1-\gamma)^2}{r^{\mathrm{e}}_{\max}+r^{\mathrm{m}}_{\max}}$, where C is the constant specified in Lemma 3.3. Using Lemma 3.4 it follows

$$m \geq \begin{cases} C_1\left(\frac{1}{\rho\epsilon_3^2(1-\gamma)^4}\ln\frac{|\mathcal{S}||\mathcal{A}|}{\epsilon_3(1-\gamma)^2\delta_2}\right), & \text{if } \rho^{-1} > \mathcal{O}\left(|\mathcal{S}|\right); \\ C_2\left(\frac{|\mathcal{S}|}{\epsilon_3^2(1-\gamma)^4} + \frac{1}{\epsilon_3^2(1-\gamma)^4}\ln\frac{|\mathcal{S}||\mathcal{A}|}{\epsilon_3(1-\gamma)^2\delta_2}\right), & \text{if } \rho^{-1} \leq \mathcal{O}\left(|\mathcal{S}|\right), \end{cases} \tag{3.6}$$

for some positive constants C_1 and C_2

Consider some fixed time step t. Let S_t be the current state. Define $\mathcal{K}$ to be the set of all state-actions that have been experienced at least m times. Let us call $\mathcal{K}$ the set of known state-actions. Recall that the Monitored MBIE-EB agent (denoted by π_t) chooses its next action by following $\pi^{**} = \arg\max_a Q_{\downarrow}^{**}(S_t, a)$ at time step t that corresponds to Mon-MDP $M_{\downarrow}^{**}$. Let $M_{\downarrow}'$ be the Mon-MDP equal to $M_{\downarrow}$ on $\mathcal{K}$ (equal reward and transition dynamics) and equal to $M_{\downarrow}^{**}$ on $\mathcal{S} \times \mathcal{A} - \mathcal{K}$. Define $\bar{M}_{\downarrow}$ be the Mon-MDP, where the intimidate expected rewards are $\bar{r}^{\mathrm{e}}$ and $\bar{r}^{\mathrm{m}}$ and $\bar{p}$ is the transition dynamics. Let $\bar{M}_{\downarrow}'$ be the the Mon-MDP that is equal to $\bar{M}_{\downarrow}$ on $\mathcal{K}$ and equal to $M_{\downarrow}^{**}$ on $\mathcal{S} \times \mathcal{A} - \mathcal{K}$. From our choice of m,

$$\left|V_{\bar{M}_{\downarrow}'}^{**}(s) - V_{M_{\downarrow}'}^{**}(s)\right| \leq \epsilon_3 \tag{3.7}$$

holds for all states, with probability at least $1-\delta_2$. Also note that $M_{\downarrow}^{**}$ is identical to $\bar{M}_{\downarrow}'$ except that some state-actions (precisely those in $\mathcal{K}$) have additional bonuses. The state-actions in $\mathcal{K}$ that their environment reward is observable will have these extra bonuses $\frac{\beta^{\mathrm{e}}}{\sqrt{\rho m}}, \frac{\beta^{\mathrm{m}}}{\sqrt{m}}$ and $\frac{\beta}{\sqrt{m}}$. That is upon visiting joint state-action $(s, a) \equiv (s^{\mathrm{e}}, s^{\mathrm{m}}, a^{\mathrm{e}}, a^{\mathrm{m}})$ for m times

$$N(s, a) = m, \quad N(s^{\mathrm{m}}, a^{\mathrm{m}}) \geq m, \quad \text{and} \quad N(s^{\mathrm{e}}, a^{\mathrm{e}}) \geq \rho m.$$

In the worst-case $\frac{\beta^{\mathrm{m}}}{\sqrt{N(s,a)}} = \frac{\beta^{\mathrm{m}}}{\sqrt{m}}$, $\frac{\beta^{\mathrm{e}}}{\sqrt{N(s^{\mathrm{e}},a^{\mathrm{e}})}} = \frac{\beta^{\mathrm{e}}}{\sqrt{\rho m}}$, and $\frac{\beta}{\sqrt{N(s,a)}} = \frac{\beta}{\sqrt{m}}$. For state-actions in $\mathcal{K}$, where rewards are unobservable, we add bonuses of $\frac{\beta^{\mathrm{m}}}{\sqrt{m}}$ and $\frac{\beta}{\sqrt{m}}$. Hence,

$$V_{\downarrow}^{**} = V_{M_{\downarrow}^{**}}^{**} \leq V_{\bar{M}_{\downarrow}'}^{**} + (1-\gamma)^{-1}\left(\frac{\beta^{\mathrm{e}}}{\sqrt{\rho m}} + \frac{\beta^{\mathrm{m}}}{\sqrt{m}} + \frac{\beta}{\sqrt{m}}\right).$$

For our analysis, we require that

$$(1-\gamma)^{-1}\left(\frac{\beta^{\mathrm{e}}}{\sqrt{\rho m}} + \frac{\beta^{\mathrm{m}}}{\sqrt{m}} + \frac{\beta}{\sqrt{m}}\right) \leq \epsilon_3. \tag{3.8}$$

In Equation (3.2) we defined

$$\beta^{\mathrm{e}} = \sqrt{2\ln\left(\frac{6\,|\mathcal{S}|\,|\mathcal{A}|\,m}{\delta_2}\right)}, \quad \beta^{\mathrm{m}} = \sqrt{2\ln\left(\frac{6\,|\mathcal{S}|\,|\mathcal{A}|\,m}{\delta_2}\right)}, \text{ and } \quad \beta = \frac{2\gamma}{(1-\gamma)}\sqrt{2\ln\left(\frac{6\,|\mathcal{S}|\,|\mathcal{A}|\,m}{\delta_2}\right)}.$$

It is not hard to show we can make m large enough so that Equations (3.6) and (3.8) hold, yet small enough that Lemma 3.4 holds. This comes by bounding each term in Equation (3.8) by $\frac{\epsilon_3}{3}$:

$$\begin{cases} 3\frac{\beta^{\mathrm{e}}}{\sqrt{\rho m}} & \leq \epsilon_3(1-\gamma) \\ 3\frac{\beta^{\mathrm{m}}}{\sqrt{m}} & \leq \epsilon_3(1-\gamma) \\ 3\frac{\beta}{\sqrt{m}} & \leq \epsilon_3(1-\gamma) \end{cases}$$

Which results in the following inequalities

$$\begin{cases} 18\frac{\ln\left(\frac{6|\mathcal{S}||\mathcal{A}|m}{\delta_2}\right)}{\rho\epsilon_3^2(1-\gamma)^2} & \leq m, \\ 18\frac{\ln\left(\frac{6|\mathcal{S}||\mathcal{A}|m}{\delta_2}\right)}{\epsilon_3^2(1-\gamma)^2} & \leq m, \\ 72\frac{\gamma^2\ln\left(\frac{6|\mathcal{S}||\mathcal{A}|m}{\delta_2}\right)}{\epsilon_3^2(1-\gamma)^4} & \leq m. \end{cases}$$

Lemma B.1 shows these inequalities satisfy Lemma 3.4. Given Equation (3.8) holds, we have

$$V_{\downarrow}^{**} \leq V_{\bar{M}_{\downarrow}'}^{**} + \epsilon_3. \tag{3.9}$$

Let $\mathcal{E}_{M_{\downarrow}}$ be an event that some some state-action not in $\mathcal{K}$ is experienced after following π^{**} from

S_t for H steps. According to Lemma 3.5

$$V_{\downarrow}^{\pi_t}(s)_H \geq V_{M'_{\downarrow}}^{**}(s)_H - \frac{\mathbb{P}\left(\mathcal{E}_{M_{\downarrow}}\right)}{2(1-\gamma)}. \tag{3.10}$$

Consider two mutually exclusive cases. First, suppose $\mathbb{P}\left(\mathcal{E}_{M_{\downarrow}}\right) \geq 2\epsilon_3(1-\gamma)$, i.e., an agent following π_t will encounter an unknown state-action in H steps with probability at least $2\epsilon_3(1-\gamma)$. Using Chernoff-Hoeffding's inequality, after $\mathcal{O}\left(\frac{m|\mathcal{S}||\mathcal{A}|H}{\epsilon_3(1-\gamma)} \ln \frac{1}{\delta_2}\right)$ time steps, where $\mathbb{P}\left(\mathcal{E}_{M_{\downarrow}}\right) \geq 2\epsilon_3(1-\gamma)$ is satisfied, all state-actions will become known with probability at least $1 - \frac{\delta_2}{2}$.

Now suppose that $\mathbb{P}\left(\mathcal{E}_{M_{\downarrow}}\right) < 2\epsilon_3(1-\gamma)$, then we have

$$\begin{aligned}
V_{\downarrow}^{\pi_t}(S_t) &\geq V_{\downarrow}^{\pi_t}(S_t)_H - \epsilon_3 \\
&\geq V_{M'_{\downarrow}}^{**}(S_t)_H - \frac{\mathbb{P}\left(\mathcal{E}_{M_{\downarrow}}\right)}{2(1-\gamma)} - \epsilon_3 \\
&\geq V_{M'_{\downarrow}}^{**}(S_t)_H - 2\epsilon_3 \\
&\geq V_{M'_{\downarrow}}^{**}(S_t) - 3\epsilon_3 \\
&\geq V_{\bar{M}'_{\downarrow}}^{**}(S_t) - 4\epsilon_3 \\
&\geq V_{\downarrow}^{**}(S_t) - 5\epsilon_3 \\
&\geq V_{\downarrow}^{*}(S_t) - 5\epsilon_3.
\end{aligned}$$

The first step uses the bound on $\left|V_{\downarrow}^{\pi_t}(S_t) - V_{\downarrow}^{\pi_t}(S_t)_H\right|$ and the choice of H. The second step is the application of Equation (3.10). The third step is from our assumption. The forth step follows from the choice of H. The fifth step follows from Equation (3.7). The sixth step follows from Equation (3.9). And the last step uses Lemma 3.6.

Therefore, by $\epsilon_3 = \frac{\epsilon_2}{5} = \frac{\epsilon}{10}$, the policy of Monitored MBIE-EB is ϵ-minimax-optimal with probability at least $1-\delta_2 = 1-\frac{\delta}{2}$ for all, but $\mathcal{O}\left(\frac{m|\mathcal{S}||\mathcal{A}|H}{\epsilon(1-\gamma)} \ln \frac{1}{\delta}\right)$ many time steps. In the worst-case,

$$m \in \mathcal{O}\left(\frac{1}{\rho\epsilon^2(1-\gamma)^4} \ln \frac{|\mathcal{S}|\,|\mathcal{A}|}{\epsilon(1-\gamma)^2\delta}\right). \tag{3.11}$$

Comparing Equation (3.11) with the bound of $\kappa^*(k)$, the overall bound is equal to

$$\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|\,|\mathcal{A}|\,H}{\epsilon^3(1-\gamma)^5\rho}\right) = \widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|\,|\mathcal{A}|}{\epsilon^3(1-\gamma)^6\rho}\right). \qquad \square$$

In words, Theorem 3.1 implies that we can have two positive numbers δ and ϵ, which we divide each of them into two equal numbers δ_1 and δ_2, and ϵ_1 and ϵ_2, where $\delta_1 = \delta_2 = \frac{\delta}{2}$, and $\epsilon_1 = \epsilon_2 = \frac{\epsilon}{2}$. Then, with probability at least $1-\delta_1$, the observation stage determines the observability of all environment rewards in $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{\epsilon_1^3(1-\gamma)^6}\right)$ time steps. After the observation stage, with probability at least $1-\delta_2$, the minimax-optimal policy is found in $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\epsilon_2^3(1-\gamma)^6\rho}\right)$ time steps. The bounds add, yielding a polynomial sample complexity dominated by the second term.

3.3.2 Dependence On ρ^{-1} Is Unimprovable

The main distinction between our derived bound in Theorem 3.1 and sample complexities given for MDPs such as the bound of Theorem 2.1 is the existence of ρ^{-1}. In this section, by showing the existence of ρ^{-1} in a lower bound, we conclude that the dependence of our bound in Theorem 3.1 on ρ^{-1} is essentially unimprovable. Note that lower bounds quantify the difficulty of learning for a given problem for *any algorithm*. Given a lower bound, no algorithm's performance can be better than what the lower bound indicates. Hence, providing a lower bound with its dependence on ρ^{-1} proves the tightness of the Monitor MBIE-EB's sample complexity on term ρ^{-1}.

To provide the lower bound, we consider the problem of stochastic bandits with finitely-many arms (multi-armed bandit for brevity) [29] as a simpler form of sequential decision-making. A multi-armed bandit is a special case of MDPs where the state-space $\mathcal{S}$ is a singleton and the discount factor γ equals zero. Mannor and Tsitsiklis [35] has proved a tight lower bound on the sample complexity of learning in multi-armed bandits. We follow the setup of Mannor and Tsitsiklis [35] as follows: The agent has $k+1$ arms (actions). Each arm $a \in [k]$ is associated with a sequence of identically distributed Bernoulli random variables X_{at} with unknown mean μ_a. Here, X_{at} corresponds to the reward obtained the tth time that arm a is tried. We assume that random variables X_{at} for $a = 1, \ldots, k+1$, and $t = 1, \ldots$ are independent. The last arm $a = k+1$ has a known mean of zero and pulling this arm terminates the interaction. A policy is a mapping that given a history, chooses a particular arm. We only consider policies that are guaranteed to eventually pull arm $k+1$ with probability one, for every possible vector of $[\mu_1, \ldots, \mu_k, 0]$ (otherwise the expected number of steps of interaction would be infinite). Given a particular policy and multi-armed Bernoulli bandit, we let $\mathbb{P}$ and $\mathbb{E}$ denote the induced probability measure and the expectation with respect to this measure. This probability measure captures both the randomness in the arms and the policy. Let T be total number of steps at which the policy chooses arm $k+1$ and terminates the interaction. Also, let A_t denote the arm chosen at time step t. We say that a policy is (ϵ, δ)-correct if

$$\mathbb{P}\left(\mu_{A_{T-1}} > \max_a \mu_a - \epsilon\right) \geq 1-\delta,$$

for every $[\mu_1, \ldots, \mu_{k+1}] \in [0,1]^{k+1}$, where $\mu_{k+1} = 0$. Theorem 3.2 shows the lower bound on $\mathbb{E}\left[T-1\right]$ for every (ϵ, δ)-correct policy. We refer the reader to the original work for the proof.

Theorem 3.2 (Mannor and Tsitsiklis [35, Theorem 1]). *There exists positive constants c_1, c_2, ϵ_0, and δ_0, such that for every $k \geq 2, \epsilon \in (0, \epsilon_0)$, and $\delta \in (0, \delta_0)$, and for every (ϵ, δ)-correct policy,* *there exists some* $[\mu_1, \ldots, \mu_k, 0]$ *such that*

$$\mathbb{E}\left[T-1\right] \geq c_1 \frac{k}{\epsilon^2} \ln \frac{c_2}{\delta}.$$

In particular, ϵ_0 *and* δ_0 *can be taken equal to* $\frac{1}{8}$ *and* $\frac{e^{-4}}{4}$*, respectively.*

In the following corollary we use the result of Theorem 3.2 to provide the lower bound in a multi-armed bandit, where the reward of each arm is revealed with only the probability of ρ.

Corollary 3.2.1. *Under the Theorem 3.2's conditions with the addition that the each arm's reward is only revealed with probability* $0 < \rho < 1$ *and with probability* $1 - \rho$ *the symbol* $\perp$ *is revealed, then*

$$\mathbb{E}\left[T-1\right] \geq c_1 \frac{k}{\rho\epsilon^2} \ln \frac{c_2}{\delta}.$$

Proof. Since we only consider policies that terminates with probability one, then there exists an $n \in \mathbb{N}$ such that $T \leq n$. Let X_i denote the reward obtained at round $i = 1, 2, \ldots$. We have,

$$\begin{aligned}
\mathbb{E}\left[T\right] &= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{a=1}^{k} \mathbb{I}\left\{A_t = a\right\}\right] + 1 \\
&= \sum_{t=1}^{n}\sum_{a=1}^{k} \left(\mathbb{E}\left[\mathbb{I}\left\{A_t = a\right\}\right]\right) + 1 \\
&= \sum_{t=1}^{n}\sum_{a=1}^{k} \left(\mathbb{E}\left[\mathbb{I}\left\{A_t = a\right\} | X_t \neq \perp\right] + \mathbb{E}\left[\mathbb{I}\left\{A_t = a\right\} | X_t = \perp\right]\right) + 1 \\
&\geq \sum_{t=1}^{n}\sum_{a=1}^{k} \left(\mathbb{E}\left[\mathbb{I}\left\{A_t = a\right\} | X_t \neq \perp\right]\right) + 1 \\
&= \sum_{t=1}^{n}\sum_{a=1}^{k} \left(\frac{\mathbb{E}\left[\mathbb{I}\left\{A_t = a, X_t \neq \perp\right\}\right]}{\mathbb{P}\left(X_t \neq \perp\right)}\right) + 1 && \text{(Conditional expectation's definition)} \\
&= \frac{1}{\rho}\sum_{t=1}^{n}\sum_{a=1}^{k} \left(\mathbb{E}\left[\mathbb{I}\left\{A_t = a, X_t \neq \perp\right\}\right]\right) + 1 \\
&= \frac{1}{\rho}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a=1}^{k} \mathbb{I}\left\{A_t = a, X_t \neq \perp\right\}\right] + 1 \\
&= \frac{1}{\rho} c_1 \frac{k}{\epsilon^2} \ln \frac{c_2}{\delta} + 1 && \text{(Theorem 3.2).}
\end{aligned}$$

Thus,

$$\mathbb{E}\left[T-1\right] \geq c_1 \frac{k}{\rho\epsilon^2} \ln \frac{c_2}{\delta}. \qquad \square$$

The existence of ρ^{-1} in the lower bound of Corollary 3.2.1 asserts that the dependence of the Monitor MBIE-EB's sample complexity on ρ^{-1} is tight and unimprovable.

In summary, in this chapter we extended the MBIE-EB's idea to Mon-MDPs. However, due to the contrasting demands of pessimism in Mon-MDPs and optimism by MBIE-EB, we introduced the observation stage. In the observation stage, the agent determines in which state-actions it should be pessimistic. We introduced Monitored MBIE-EB that combines the MBIE-EB's extension with the observation stage. We showed Monitored MBIE-EB enjoys a polynomial sample complexity. Finally, we provided a lower bound in a special case of the stochastic multi-armed bandit problem with partially observable rewards. The lower bound proved that the dependence of Monitored MBIE-EB's sample complexity on the probability of observing the environment rewards is tight.

Chapter 4

Empirical Evaluation

In this chapter, we present the details for Monitored MBIE-EB's practical implementation[7]. We demonstrate Monitored MBIE-EB's superiority against the current SOTA method, Directed Exploration-Exploitation [39], over four dozen domains.

4.1 Practical Implementation

The Monitored MBIE-EB's advantages extend beyond theory. In this section, we explain the Monitored MBIE-EB's practical implementation. This explanation is necessary because theoretically justified parameters for Monitored MBIE-EB present challenges in practice. These parameters appear in the Theorem 3.1's bound and the value of $\kappa^*(k)$. First, we rarely have a particular ϵ and δ in mind, preferring algorithms that produce ever-improving approximations with ever-improving probability. The second is the constant $\kappa^*(k) = k^*$, which places the observation stage only at the start of training. Third, running value iteration from scratch (with an initial value of 0) prior to each episode to compute $Q_{\downarrow}^{**}$ or $\widetilde{Q}^{**}$ is computationally wasteful.

In practice, we follow the pattern of Lattimore and Szepesvári [29], with the confidence bounds growing slightly faster than logarithmically. By defining $g(x) = 1 + x\ln^2(x), x \in \mathbb{R}_{\geq 0}$, Table 4.1 summarizes how we replaced theoretically justified Monitored MBIE-EB's parameters with ones that can be used in practice. After using the practical expressions, the scale parameters β, β^{m}, β^{e}, β^{obs}, and $\beta^{\text{KL-UCB}}$ are tuned manually for each domain. The log base of $\kappa^*(k) = \log k$, is also manually tuned for each domain. Finally, $Q_{\downarrow}^{**}$ or $\widetilde{Q}^{**}$ are both initialized optimistically. We do fifty steps of value iteration [43] before every episode to improve $Q_{\downarrow}^{**}$ and fifty steps of value iteration before episodes at the observation stage to improve $\widetilde{Q}^{**}$. At each invocation, value iteration is initialized with the most recent output, maintaining separate updates for $Q_{\downarrow}^{**}$ and $\widetilde{Q}^{**}$.

[7] Code: https://github.com/IRLL/Exploration-in-Mon-MDPs.

Theoretical	Practical	Unexplained variables
$\beta = \frac{2\gamma}{(1-\gamma)}\sqrt{2\ln\,(\frac{12\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert m}{\delta}}\,)$	$\beta\sqrt{g(\ln N(s))}$	$N(s)$ is the number of visits to state s
$\beta^{\mathrm{m}} = \sqrt{2\ln\,(\frac{12\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert m}{\delta}}\,)$	$\beta^{\mathrm{m}}\sqrt{g(\ln N(s^{\mathrm{m}}))}$	$N(s^{\mathrm{m}})$ is the number of visits to state s^{m}
$\beta^{\mathrm{e}} = \sqrt{2\ln\,(\frac{12\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert m}{\delta}}\,)$	$\beta^{\mathrm{e}}\sqrt{g(\ln N(s^{\mathrm{e}}))}$	$N(s^{\mathrm{e}})$ is the number of observed rewards at state s^{e}
$\beta^{\mathrm{obs}} = (1-\gamma)^{-1}\sqrt{0.5\ln\left(\frac{8\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert m}{\delta}\right)}$	$\beta\sqrt{g(\ln N(s))}$	$N(s)$ is the number of visits to state s
$\beta^{\text{KL-UCB}} = \ln\left(\frac{8\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert m}{\delta}\right)$	$\beta^{\text{KL-UCB}}g\left(\ln N(s)\right)$	$N(s)$ is the number of visits to state s
$\kappa^*(k) = k^* = \widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert H}{\epsilon_1^3(1-\gamma)^5}\right)$	$\kappa^*(k) = \log k$	-

Table 4.1: Replacing Monitored MBIE-EB's theoretical parameters with practical alternatives.

4.2 Evaluation

This thesis claims Monitored MBIE-EB takes advantage of: the Mon-MDP structure, the possibility of a known monitor, and dealing with unsolvable Mon-MDPs. This section divides these claims into four research questions (RQs) to investigate if Monitored MBIE-EB:

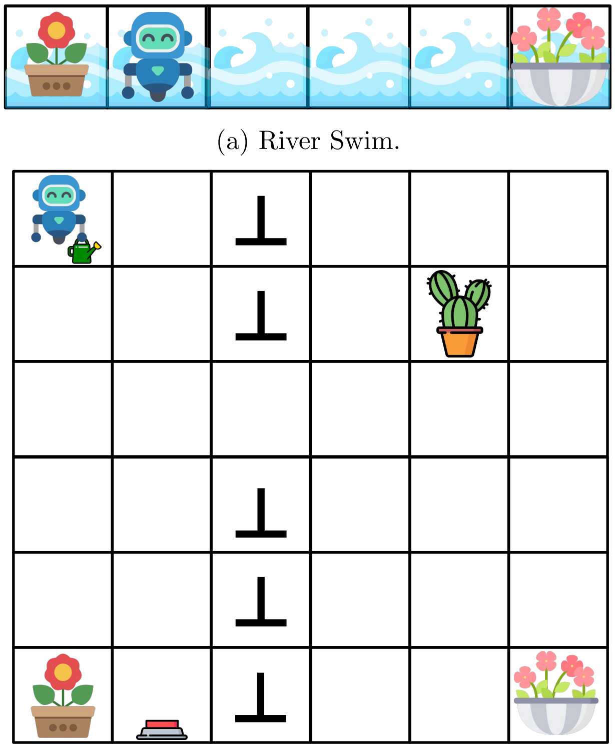
- RQ1) Explores efficiently in hard-exploration tasks?
- RQ2) Acts pessimistically when rewards are unobservable?
- RQ3) Identifies and learns about difficult to observe rewards?
- RQ4) Takes advantage of a known model of the monitor?

To address these questions, we first present results on two tasks with two monitors, followed by results on 48 benchmarks from Parisi et al. [39], designed to showcase the Directed-E^2's effectiveness.

4.2.1 Environment and Monitor Description

In this section, we explain the dynamics of *River Swim*, a classic environment to test how Monitored MBIE-EB performs the exploration. Also, we explain the dynamics of *Bottleneck*, an environment we designed to create variations of reward observability in Mon-MDPs. We also describe the dynamics of the monitors used in each experiment.

River Swim (Figure 4.1a) is a difficult exploration task with two actions. It was initially designed to highlight the MBIE-EB's strength in performing deep, efficient exploration [37, 47]. Moving `LEFT` always succeeds, but moving `RIGHT` may not — the river current may cause the agent to stay at the same location or even be pushed to the left. The goal state is on the far right with a reward of 1. However, the `LEFT` action at the leftmost tile yields a reward of 0.1, and it is much easier to reach. Other states have zero rewards. Agents often struggle to find the optimal policy (always



(a) River Swim.

(b) Bottleneck.

Figure 4.1: River Swim and Bottleneck. River Swim is appropriate for testing exploration capabilities. Bottleneck is useful to create different non-ergodic Mon-MDPs.

move `RIGHT`), and, converge to always move `LEFT`. We pair River Swim with the *MDP Monitor*, which ensures rewards are always visible, letting us focus on the algorithm's exploration ability.

As one of our contributions, Bottleneck environment, (Figure 4.1b) accepts five deterministic actions: `LEFT`, `UP`, `RIGHT`, `DOWN`, `WATER`, which move the agent around the grid and pour water. Episodes end when the agent executes `WATER` in either small flower pot (with a reward of 0.1) or in the big flower pot state (with a reward of 1). Reaching the cactus yields -10, and other states yield 0. However, states denoted by $\perp$ have *never-observable* rewards of -10, i.e., $R^{\mathrm{e}}_{t+1} = -10$ at time step t for actions leading to those states but $\widehat{R}^{\mathrm{e}}_{t+1}$ is equal to $\perp$. In these experiments, we pair Bottleneck with the *Button Monitor*, where the monitor state S^{m}_{t} at time step t is either `ON` or `OFF` (initialized at random) and is switched if the agent executes `DOWN` in the button state. When the monitor is `ON`, the agent receives $R^{\mathrm{m}}_{t+1} = -0.2$ at every time step and observes the current environment reward (except for$\perp$ cells). The minimax-optimal policy follows the shortest path to the big flower pot, while avoiding the cactus and $\perp$ states. The minimax-optimal policy should turn the monitor `OFF`, if it was initialized to `ON` at the beginning of the episode. To evaluate how Monitored MBIE-EB performs when observability is stochastic, we consider two versions of the

Button Monitor: one where the monitor works as intended and rewards are observable with 100% probability if ON, and the other where the rewards are observable only with 5% probability if ON (and 0% of the time when OFF). Also, we show the result when the monitoring process is known to the agent in advance by the Known monitor baseline.

4.2.2 Results

In this section, we present the results of executing Monitored MBIE-EB on River Swim and Bottleneck. In all benchmarks, the discount factor is $\gamma = 0.99$. The full set of hyperparameters appear in Appendix D.7, and full evaluation details (e.g., episode lengths, evaluation frequencies, etc.) are in Appendix D.1. Results shown in Figures 4.2 and 4.3 are at test time, i.e., when during training, the agent's learning is paused and the agent follows the current greedy policy without exploring.

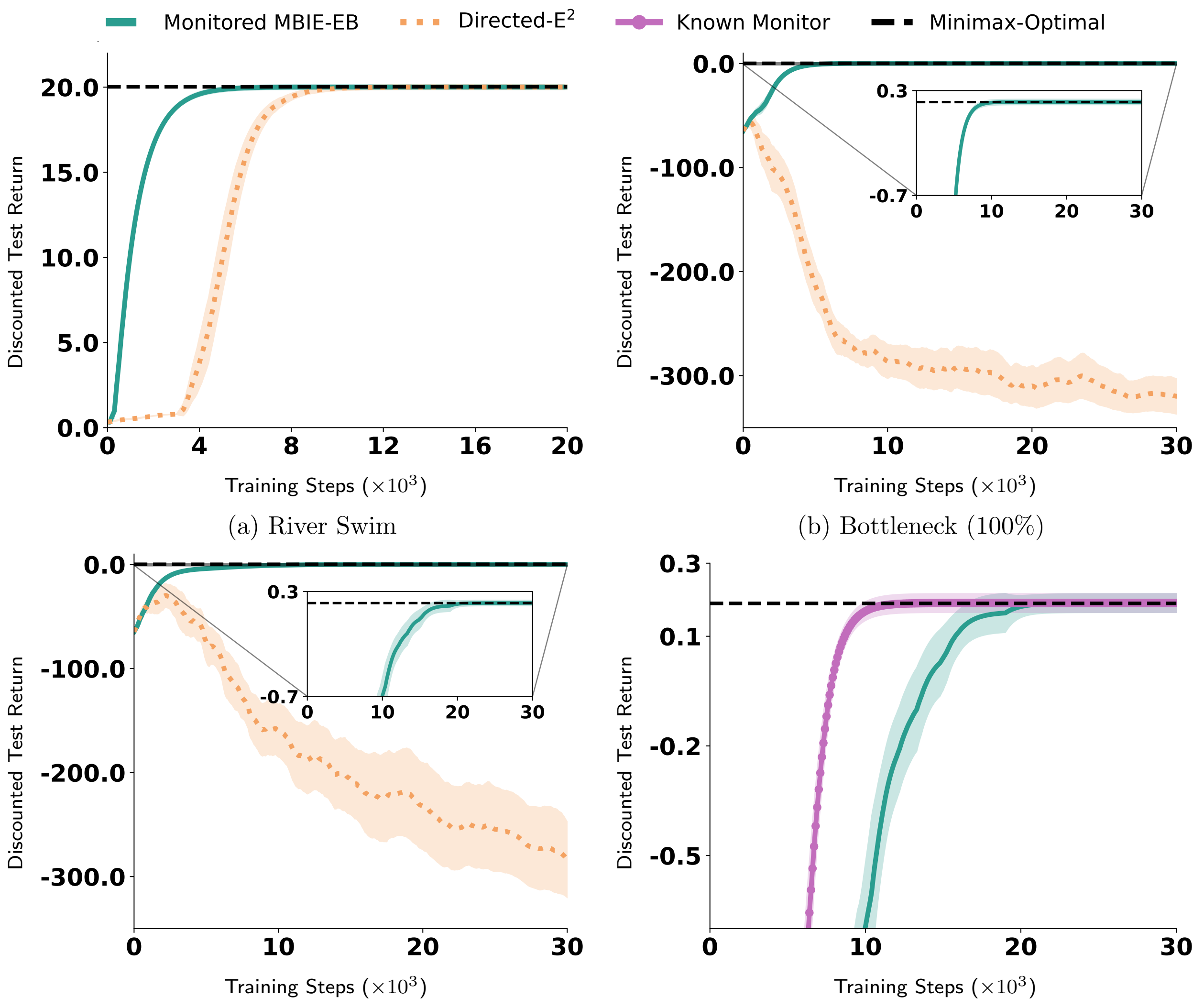
To answer RQ1, consider the results in Figure 4.2a. In this case, the performance of Monitored MBIE-EB significantly outperforms that of Directed-E^2. This task is difficult for any ϵ-greedy exploration strategy (such as the one of Directed-E^2).

To answer RQ2, consider Figure 4.2b. In this case, states marked with $\perp$ are never observable by the agent, regardless of the monitor state. Because the minimum environment reward in this task is $r^{\mathrm{e}}_{\min} = -10$, the minimax-optimal policy is to avoid states marked by $\perp$ while reaching the goal state. Monitored MBIE-EB is able to find this minimax-optimal policy, whereas Directed-E^2 does not because it does not learn to avoid unobservable rewards[1]. This result highlights the impact of pessimism: unsolvable Mon-MDPS require pessimism when the reward cannot be observed.

To answer RQ3, consider Figure 4.2c. Despite the difficulty to observe rewards, Monitored MBIE-EB is able to learn the minimax-optimal policy. This shows that Monitored MBIE-EB is still appropriately pessimistic, successfully avoiding $\perp$ states and the cactus, and reaches the goal state. Because rewards are only visible one out of twenty times (when the monitor is ON), learning is much slower than in Figure 4.2b, matching ρ^{-1} in Theorem 3.1's bound.

To answer RQ4, consider the Known monitor's results in Figure 4.2d, demonstrating the Monitored MBIE-EB's performance when provided the model of the Button Monitor 5%. Results indicate Monitored MBIE-EB's convergence speed increases significantly, as Monitored MBIE-EB takes (on average) 30% fewer steps to find the optimal policy. This feature of Monitored MBIE-EB is particularly important in settings where the agent has already learned about the monitor previously, or the practitioner provides the agent with an accurate model of the monitor. The agent needs only to learn about the environment, and does not need to explore the monitoring process.

[1] Directed-E^2 describes initializing its reward model randomly, relying on the Mon-MDP being solvable, independent of the initialization. For unsolvable Mon-MDPs, this is not true, and Directed-E^2 depends significantly on initialization. In fact, while not noted by Parisi et al. [39], pessimistic initialization with Directed-E^2 results in asymptotic convergence for unsolvable Mon-MDPs.



(a) River Swim

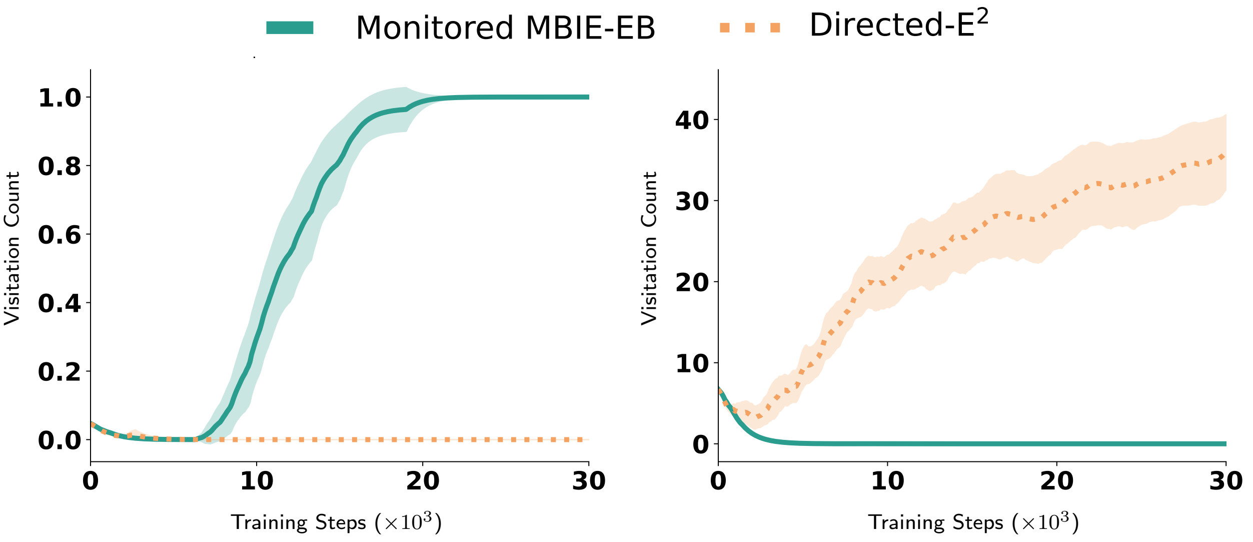
(b) Bottleneck (100%)

(c) Bottleneck (5%)

(d) Bottleneck (5%)

Figure 4.2: Discounted return at test time, averaged over 30 seeds (shaded areas denote 95% confidence intervals). Monitored MBIE-EB (in green) outperforms Directed-E^2 (in orange) and always converges to the minimax-optimal policy (the dashed black line). (a) shows the superior exploration of Monitored MBIE-EB compared to Directed-E^2. (b) shows that Monitored MBIE-EB finds the minimax-optimal policy while Directed-E^2 does not. (c) and (d) both show results in the Bottleneck with the 5% Button Monitor, but with different axis ranges to highlight the improvement if Monitored MBIE-EB already knows details of the monitor (in purple).

To better understand the above results, Figure 4.3 shows how many times the agent visits the goal state and $\perp$ states per testing episode. Both algorithms initially visit the goal state (Figure 4.3a) during random exploration (i.e., when executing the policy after zero time steps of training). Monitored MBIE-EB appropriately explores for some training episodes (recall that rewards are only observed in `ON` and even then only 5% of the time), and then learns to always go



(a) Visits to the goal

(b) Visits to $\perp$

Figure 4.3: Visits to important states at test time in the Bottleneck with 5% Button Monitor. Results are averaged over 30 trials, and shaded areas denote 95% confidence intervals. Directed-E^2 fails to focus on the goal and instead keeps visiting $\perp$ states, whereas Monitored MBIE-EB reduces its visitation frequency instead, ultimately visiting only the goal.

to the goal. Both initially visit $\perp$ states (Figure 4.3b). However, while Monitored MBIE-EB learns to be appropriately pessimistic over time and avoids them, Directed-E^2 never updates its (random) initial estimate of the value of $\perp$ states and incorrectly believes they should continue to be visited. This lack of update explains why Directed-E^2 performs even worse in Figure 4.2c.

Finally, Figure 4.4 presents results comparing Monitored MBIE-EB across all domains and monitor benchmarks from Parisi et al. [39]. In these 48 benchmarks, Monitored MBIE-EB significantly outperforms Directed-E^2 in all but five of them, where they perform similarly. Since the confidence intervals don not overlap, the performance of Monitored MBIE-EB over 43 of the benchmarks are statistically significant. Appendix D.1 contains the details of all 48 benchmarks.

In summary, in this chapter we discussed possible changes to make Monitored MBIE-EB practical on finite domains. We revisited Directed-E^2 as the SOTA algorithm in Mon-MDPs and used it as the benchmark of empirical performance of Monitored MBIE-EB. We showed efficient exploration of Monitored MBIE-EB on River Swim. We also demonstrated the minimax-optimality of Monitored MBIE-EB on Bottleneck, where some rewards are never-observable. We provided evidence that knowing the monitor speeds up the learning. Finally, we showed Monitored MBIE-EB comprehensively outperforms Directed-E^2 on domains that where designed for Directed-E^2.

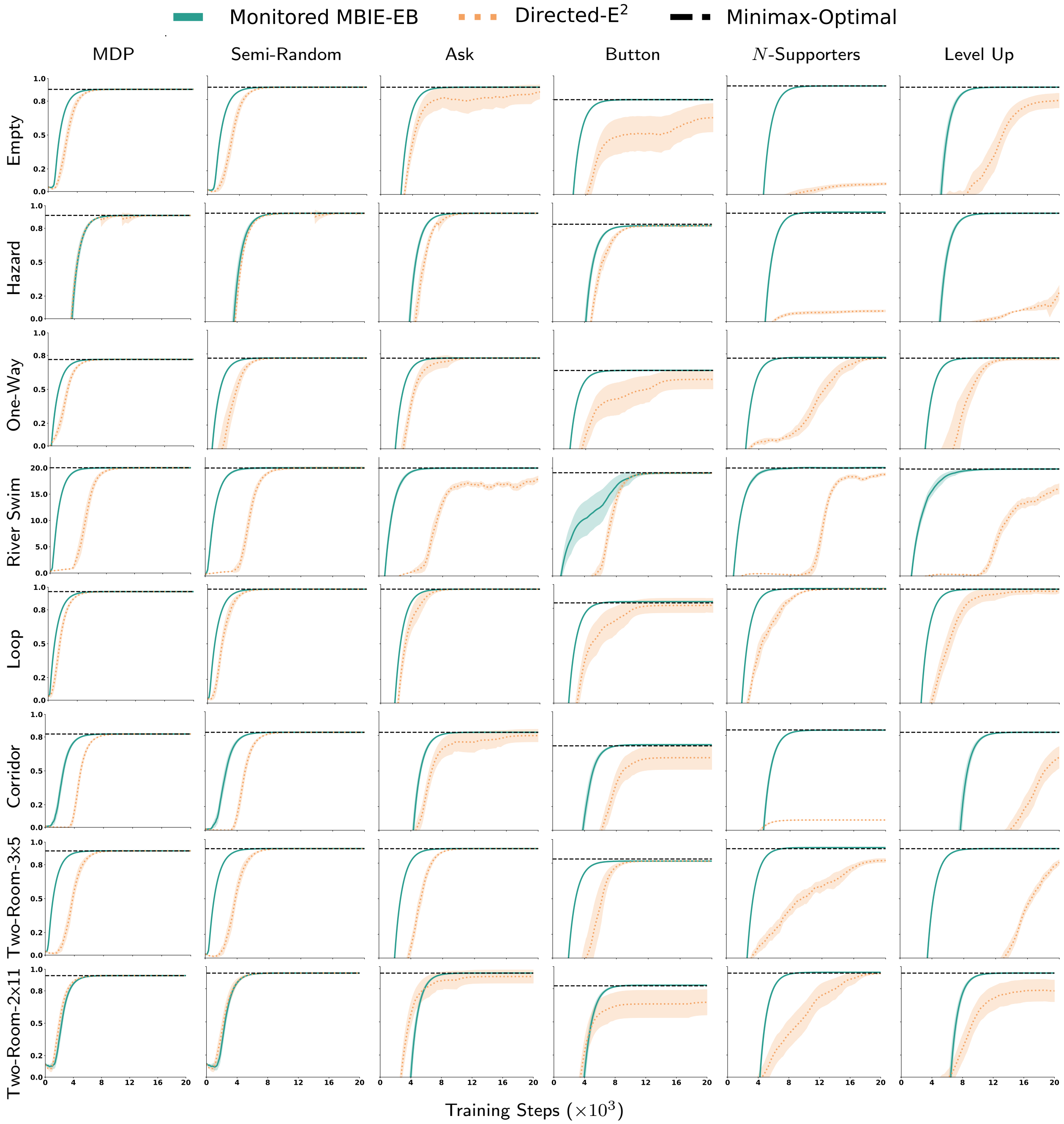


Figure 4.4: Performance on 48 benchmarks from Parisi et al. [39]. Monitored MBIE-EB outperforms Directed-E^2 in 43 of them and performs on par in the remaining five.

Chapter 5

Conclusion and Future Work

The core premise of this work is that incorporating reward observability into algorithm design enables the effective handling of sequential decision-making problems in which the reward signal is entirely unobservable to the agent. We demonstrated the effectiveness of incorporating reward observability by proposing Monitored MBIE-EB and the rationale behind every crucial step in Monitored MBIE-EB's development from theory to practice. We showed Monitored MBIE-EB is effective and efficient in finding the optimal policy in MDPs. Hence it is applicable in the traditional MDP formulation. Additionally, we demonstrated that Monitored MBIE-EB is effective and efficient in Mon-MDPs, where traditional algorithms developed for MDPs fail. Monitored MBIE-EB's Effectiveness in Mon-MDPs was further supported by illustrating the superior empirical performance of Monitored MBIE-EB compared to the SOTA algorithm Directed-E^2 in Mon-MDPs. Furthermore, since Monitored MBIE-EB is a model-based algorithm, this property allowed to incorporate prior knowledge about the monitor, which resulted in faster learning. In this chapter we state the limitations of Monitored MBIE-EB which leads to laying out the potential future avenues to extend the current work. Finally, we provide a conclusion.

5.1 Limitations and Future work

There are a number of limitations to our approach that suggest directions for future improvements. First, Mon-MDPs contain an exploration-exploitation dilemma, but with an added twist — the agent needs to treat never observed rewards pessimistically in order to achieve a minimax-optimal solution; however, the agent should continue exploring those states to get more confident about their unobservability. Much like early algorithms for the exploration-exploitation dilemma in MDPs [25], our approach separately optimizes a model for observing and one for seeking a minimax-optimal solution. A more elegant approach would be to simultaneously optimize for both. Second, the notion of the minimax-optimality presented in this work does not necessarily lead to a no-regret algorithm. In other words, there is a possibility that assuming the minimum reward for unobservable ones is unnecessarily pessimistic and since the agent is excessively pessimistic about the unobservable rewards, it suffers from regret when the underlying rewards are not the minimum possible. One solution is to develop a randomized algorithm that stochastically deals with unobservable rewards. Third, our approach uses explicit counts to drive its exploration, which limits it to enumerable Mon-MDPs. Adapting psuedocount-based methods [6, 33, 36, 50] can help making Monitored MBIE-EB more applicable to large or continuous state spaces. Finally, the decision of when to stop trying to observe rewards and instead optimize is essentially an optimal stopping time problem [29], and there may be considerable innovations that could improve the bounds along with empirical performance.

5.2 Conclusion

We introduced Monitored MBIE-EB for Mon-MDPs that addressed many of the shortcomings of previous algorithms. Monitored MBIE-EB admits the first sample complexity bounds for Mon-MDPs, while being applicable to both solvable and unsolvable Mon-MDPs, for which it is also the first. We showed the dependence of Monitored MBIE-EB's sample complexity on the inverse of the probability of observing the reward is tight. Furthermore, Monitored MBIE-EB exploits the structure of the Mon-MDP and can take advantage of knowledge of the monitor process, if available. These features were shown to not just be theoretical. We showed these innovations result in empirical improvements in Mon-MDP benchmarks, comprehensively outperforming the previous best learning algorithm.

References

[1] S. Ainsworth, M. Barnes, and S. Srinivasa. Mo'states mo'problems: Emergency stop mechanisms from observation. *Advances in Neural Information Processing Systems*, 32, 2019. (cited on page 17.)

[2] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 2020. (cited on pages 3 and 13.)

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47:235–256, 2002. (cited on page 19.)

[4] M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, 2017. (cited on page 23.)

[5] G. Bartók, D. P. Foster, D. Pál, A. Rakhlin, and C. Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, pages 967–997, 2014. (cited on page 14.)

[6] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying Count-Based Exploration and Intrinsic Motivation. In *Advances in Neural Information Processing Systems*, volume 29, 2016. (cited on pages 19 and 44.)

[7] D. P. Bossev, A. R. Duncan, M. J. Gadlage, A. H. Roach, M. J. Kay, C. Szabo, T. J. Berger, D. A. York, A. Williams, K. LaBel, et al. Radiation Failures in Intel 14nm Microprocessors. In *Military and Aerospace Programmable Logic Devices (MAPLD) Workshop*, 2016. (cited on page 13.)

[8] M. Bowling, J. D. Martin, D. Abel, and W. Dabney. Settling the reward hypothesis. In *International Conference on Machine Learning*, pages 3003–3020. PMLR, 2023. (cited on page 2.)

[9] R. I. Brafman and M. Tennenholtz. R-max: A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 2002. (cited on page 23.)

[10] J. Buckman, C. Gelada, and M. G. Bellemare. The Importance of Pessimism in Fixed-Dataset Policy Optimization. In *International Conference on Learning Representations*, 2021. (cited on page 17.)

[11] I. Chadès, L. V. Pascal, S. Nicol, C. S. Fletcher, and J. Ferrer-Mestres. A primer on partially observable Markov decision processes POMDPs. *Methods in Ecology and Evolution*, 2021. (cited on page 13.)

[12] H. D. Dixit, S. Pendharkar, M. Beadon, C. Mason, T. Chakravarthy, B. Muthiah, and S. Sankar. Silent data corruptions at scale. *arXiv preprint arXiv:2102.11245*, 2021. (cited on pages 4 and 13.)

[13] E. Even-Dar, S. Mannor, and Y. Mansour. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of Machine Learning Research*, 2006. (cited on pages 9 and 10.)

[14] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is All You Need: Learning Skills without a Reward Function. In *International Conference on Learning Representations*, 2019. (cited on page 13.)

[15] A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, 2011. (cited on page 25.)

[16] J. Hejna and D. Sadigh. Inverse preference learning: Preference-based RL without a reward function. *Advances in Neural Information Processing Systems*, 2024. (cited on page 13.)

[17] P. H. Hochschild, P. Turner, J. C. Mogul, R. Govindaraju, P. Ranganathan, D. E. Culler, and A. Vahdat. Cores that don't count. In *Proceedings of the Workshop on Hot Topics in Operating Systems*, page 9–16, 2021. (cited on page 13.)

[18] T. Jaksch, R. Ortner, and P. Auer. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 2010. (cited on page 23.)

[19] R. Jaulmes, J. Pineau, and D. Precup. Active Learning in Partially Observable Markov Decision Processes. In *Machine Learning: ECML 2005*, 2005. (cited on page 13.)

[20] N. Jiang. PAC reinforcement learning with an imperfect model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. (cited on page 17.)

[21] Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021. (cited on page 17.)

[22] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 1998. (cited on page 13.)

[23] S. M. Kakade. *On the sample complexity of reinforcement learning.* University of London, University College London (United Kingdom), 2003. (cited on pages 10, 12, 27, and 29.)

[24] C. Kausik, M. Mutti, A. Pacchiano, and A. Tewari. A Framework for Partially Observed Reward-States in RLHF. *arXiv preprint arXiv:2402.03282*, 2024. (cited on page 13.)

[25] M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 2002. (cited on page 43.)

[26] R. Kidambi, A. Rajeswaran, P. Netrapalli, and T. Joachims. MOReL: Model-Based Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21810–21823, 2020. (cited on page 17.)

[27] D. Krueger, J. Leike, O. Evans, and J. Salvatier. Active Reinforcement Learning: Observing Rewards at a Cost. *arXiv:2011.06709*, 2020. (cited on page 13.)

[28] T. Lattimore and M. Hutter. PAC bounds for discounted MDPs. In *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23*, pages 320–334, 2012. (cited on page 12.)

[29] T. Lattimore and C. Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020. (cited on pages 19, 33, 36, and 44.)

[30] L. Li. *A unifying framework for computational reinforcement learning theory.* Rutgers The State University of New Jersey, School of Graduate Studies, 2009. (cited on page 9.)

[31] J. MacGlashan, M. K. Ho, R. Loftin, B. Peng, G. Wang, D. L. Roberts, M. E. Taylor, and M. L. Littman. Interactive learning from policy-dependent human feedback. In *International conference on machine learning*, 2017. (cited on page 13.)

[32] M. C. Machado and M. Bowling. Learning purposeful behaviour in the absence of rewards. *arXiv preprint arXiv:1605.07700*, 2016. (cited on page 13.)

[33] M. C. Machado, M. G. Bellemare, and M. Bowling. Count-based exploration with the successor representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. (cited on page 44.)

[34] O.-A. Maillard, R. Munos, and G. Stoltz. A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of the 24th annual Conference On Learning Theory*, 2011. (cited on page 25.)

[35] S. Mannor and J. N. Tsitsiklis. The Sample Complexity of Exploration in the Multi-Armed Bandit Problem. *Journal of Machine Learning Research*, 5:623–648, 2004. (cited on page 33.)

[36] J. Martin, S. Narayanan Sasikumar, T. Everitt, and M. Hutter. Count-Based Exploration in Feature Space for Reinforcement Learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI*, pages 2471–2478, 2017. (cited on page 44.)

[37] I. Osband, B. V. Roy, D. J. Russo, and Z. Wen. Deep Exploration via Randomized Value Functions. *Journal of Machine Learning Research*, 2019. (cited on pages 37 and 67.)

[38] S. Parisi, D. Tateo, M. Hensel, C. D'eramo, J. Peters, and J. Pajarinen. Long-term visitation value for deep exploration in sparse-reward reinforcement learning. *Algorithms*, 15(3):81, 2022. (cited on page 19.)

[39] S. Parisi, A. Kazemipour, and M. Bowling. Beyond Optimism: Exploration With Partially Observable Rewards. In *Advances in Neural Information Processing Systems*, 2024. (cited on pages 18, 19, 36, 37, 39, 41, 42, 64, 66, 67, and 73.)

[40] S. Parisi, M. Mohammedalamen, A. Kazemipour, M. E. Taylor, and M. Bowling. Monitored Markov Decision Processes. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2024. (cited on pages 14, 16, 17, 18, and 50.)

[41] P. M. Pilarski, M. R. Dawson, T. Degris, F. Fahimi, J. P. Carey, and R. S. Sutton. Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. In *2011 IEEE International Conference on Rehabilitation Robotics*, 2011. (cited on page 13.)

[42] J. Pineau, G. Gordon, and S. Thrun. Anytime point-based approximations for large POMDPs. *Journal of Artificial Intelligence Research*, 2006. (cited on page 13.)

[43] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. (cited on pages 6, 7, 8, and 36.)

[44] P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021. (cited on page 17.)

[45] K. Regan and C. Boutilier. Robust Policy Computation in Reward-Uncertain MDPs Using Nondominated Policies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24: 1127–1133, 2010. (cited on page 17.)

[46] D. Silver and J. Veness. Monte-Carlo planning in large POMDPs. *Advances in neural information processing systems*, 2010. (cited on page 13.)

[47] A. L. Strehl and M. L. Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences (JCSS)*, 2008. (cited on pages 9, 10, 11, 12, 23, 27, 29, 37, 53, 56, and 61.)

[48] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, 2018. (cited on pages 6, 8, 12, and 13.)

[49] C. Szepesvári. *Algorithms for reinforcement learning.* Springer nature, 2022. (cited on page 8.)

[50] H. Tang, R. Houthooft, D. Foote, A. Stooke, O. Xi Chen, Y. Duan, J. Schulman, F. DeTurck, and P. Abbeel. # Exploration: A study of count-based exploration for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 2017. (cited on page 44.)

[51] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27:1134–1142, 1984. (cited on page 12.)

[52] A. Vemula, Y. Oza, J. A. Bagnell, and M. Likhachev. Planning and execution using inaccurate models with provable guarantees. *Robotics: Science and Systems*, 2020. (cited on page 17.)

[53] T. L. Vu, S. Mukherjee, R. Huang, and Q. Huang. Barrier Function-based Safe Reinforcement Learning for Emergency Control of Power Systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, 2021. (cited on page 13.)

[54] D. Warde-Farley, T. V. de Wiele, T. Kulkarni, C. Ionescu, S. Hansen, and V. Mnih. Unsupervised Control Through Non-Parametric Discriminative Rewards. In *International Conference on Learning Representations*, 2019. (cited on page 13.)

[55] M. Wiering and J. Schmidhuber. Efficient model-based exploration. In *5th International Conference on Simulation of Adaptive Behavior*, 1998. (cited on page 9.)

[56] T. Zhang. *Mathematical Analysis of Machine Learning Algorithms.* Cambridge University Press, 2023. (cited on page 12.)

Appendix A

Solvability of Mon-MDPs

In this chapter we formally define when a Mon-MDP is solvable and when unsolvable. This chapter is useful to clarify why a Mon-MDP is called unsolvable.

Definition 5 (Parisi et al. [40, Definition 1])**.** *Let $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p, f^{\mathrm{m}} \rangle$ be a truthful Mon-MDP. Let $\mathcal{M}$ be the set of all Mon-MDPs that differ from M only in expected environment reward r^{e}. Define Π_M to be the set of all policies in M and $\Pi = \bigcup_{M' \in \mathcal{M}} \Pi_{M'}$ to be the set of all Mon-MDPs' policies in $\mathcal{M}$. Further, let $\tau_\ell = \left\{ \left(S_i, A_i, \widehat{R}^{\mathrm{e}}_{i+1}, R^{\mathrm{m}}_{i+1}, S_{i+1}\right)_{i=0}^{\ell-1} \middle| \pi, M \right\}$ be a trajectory of length ℓ in M when following a policy π, where $\mathbb{E}\left[\widehat{R}^{\mathrm{e}}_{i+1} \middle| \widehat{R}^{\mathrm{e}}_{i+1} \neq \bot, S_i, A_i\right] = r^{\mathrm{e}}(S^{\mathrm{e}}_i, A^{\mathrm{e}}_i)$ almost surely. Let $\mathcal{T}_L = \bigcup_{\mathcal{M} \times \Pi} \left(\cup_{l=0}^{L-1} \tau_\ell\right)$ be the set of all L length trajectories in $\mathcal{M}$. The indistinguishability relation $\mathbb{I}$ between Mon-MDPs $M_1, M_2 \in \mathcal{M}$ is defined:*

$$M_1 \mathbb{I} M_2 : \forall L \in \mathbb{N}, \forall \tau \in \mathcal{T}_L, \mathbb{P}\left(\tau | M_1\right) = \mathbb{P}\left(\tau | M_2\right) \quad \textit{almost surely.}$$

It follows directly from the definition that the indistinguishability is an equivalence relation:

1. **Reflexive.** Every Mon-MDP M is indistinguishable from itself: $M \mathbb{I} M$.

2. **Symmetric.** If M_1 is indistinguishable from M_2, so is M_2 from M_1: $M_1 \mathbb{I} M_2 \Leftrightarrow M_1 \mathbb{I} M_2$

3. **Transitive.** If M_1 and M_2 are indistinguishable, and M_2 and M_3, so are M_1 and M_3: $M_1 \mathbb{I} M_2 \wedge M_2 \mathbb{I} M_3 \Rightarrow M_1 \mathbb{I} M_3$.

As an equivalence relation, $\mathbb{I}$ partitions Mon-MDPs into disjoint classes. If $|[M]_{\mathbb{I}}| = 1$, the agent can eventually identify M and its optimal policy, making M solvable. If $|[M]_{\mathbb{I}}| > 1$, M is unsolvable, as it is indistinguishable from at least another Mon-MDP with possibly different optimal policies.

Appendix B

Propositions

This chapter revisits the auxiliary propositions useful in proving this thesis's lemmas and theorems.

Lemma B.1. *For any $a, x > 0$, $x \geq 2a \ln a$ implies $x \geq a \ln x$.*

Proof. First we prove that for $\forall x > 0$, $x > 2 \ln x$. Consider the function $y = x - 2 \ln x$. It is enough to prove that $y > 0$ is always true on its domain. we have that

$$\frac{dy}{dx} = 1 - \frac{2}{x}, \quad \frac{d^2 y}{d^2 x} = \frac{2}{x^2} > 0.$$

Therefore, $y = x - \ln x$ is convex. Also,

$$\frac{dy}{dx} = 1 - \frac{2}{x} = 0 \Rightarrow x = 2$$

Hence, the minimum of $y = x - \ln x = 2 - \ln 2 > 0$ and $x > 2 \ln x, \forall x > 0$.

Back to the inequalities in the lemma, let $z = \frac{x}{a}$, then we need to prove that $z \geq 2 \ln a$ implies $z \geq \ln(za)$. There are only two possible cases:

- If $a \geq z$, then:

  $$\ln za = \ln z + \ln a \leq 2 \ln a \leq z$$

  Which is by the assumption of $z \geq 2 \ln a$.

- If $a < z$, then:

  $$\ln a < \ln z \Rightarrow \ln za \leq 2 \ln z < z$$

  which is by our proof in the beginning that $z \geq 2 \ln z, \forall z > 0$. □

Lemma B.2. *Let Ω be an outcome space, and each of $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^n$ each be n random variables on Ω. It holds that,*

$$\left\{\sum_{i=1}^n X_i \geq \sum_{i=1}^n Y_i\right\} \subseteq \left\{\bigcup_{i=1}^n (X_i \geq Y_i)\right\}.$$

Proof. Proof by contradiction.

Suppose $\{\sum_{i=1}^n X_i \geq \sum_{i=1}^n Y_i\} \supset \{\bigcup_{i=1}^n (X_i \geq Y_i)\}$, then it means there exists an $\omega \in \Omega$ such that $\sum_{i=1}^n X_i(\omega) \geq \sum_{i=1}^n Y_i(\omega)$ but $X_1(\omega) < Y_1(\omega), X_2(\omega) < Y_2(\omega), \cdots X_n(\omega) < Y_n(\omega)$ that results in $\sum_{i=1}^n X_i(\omega) < \sum_{i=1}^n Y_i(\omega)$ which is a contradiction. □

Corollary B.0.1. *Let $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^n$ be n random variables on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. It holds that,*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq \sum_{i=1}^n Y_i\right) \leq \sum_{i=1}^n \mathbb{P}\left(X_i \geq Y_i\right).$$

Proof. Using Lemma B.2 and due to monotonicity of measures we have

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq \sum_{i=1}^n Y_i\right) \leq \mathbb{P}\left(\bigcup_{i=1}^n (X_i \geq Y_i)\right).$$

By applying the union bound the inequality is obtained. □

Lemma B.3 (Chernoff-Hoeffding's inequality)**.** *For $(X_i)_{i=1}^n$ independent samples on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $X_i \in [a_i, b_i]$ for all i and $\epsilon > 0$, we have:*

$$\mathbb{P}\left(\mathbb{E}\left[X_1\right] - \frac{1}{n}\sum_{i=1}^n X_i \geq \epsilon\right) \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Appendix C

Proofs

In this chapter, we prove the lemmas that are stated in the main body of this work.

C.1 Proof of Lemma 3.1

Proof. Let $N(s,a)$ denote the number of times a fixed state-action (s,a) is visited. According to Strehl and Littman [47, Lemma 7], by choosing $\beta^{\text{obs}} = (1-\gamma)^{-1}\sqrt{0.5 \ln\left(\frac{4|\mathcal{S}||\mathcal{A}|m}{\delta}\right)}$,

$$\widetilde{Q}^{**}(s,a) = \gamma \sum_{s'} \bar{p}\left(s' \middle| s,a\right) \max_{a'} \widetilde{Q}^{**}\left(s',a'\right) + \frac{\beta^{\text{obs}}}{\sqrt{N(s,a)}} \geq \widetilde{Q}^{*}(s,a),$$

with probability at least $1 - \frac{\delta}{4|\mathcal{S}||\mathcal{A}|m}$. On the other hand, by choosing $\beta^{\text{KL-UCB}} = \ln\left(\frac{4|\mathcal{S}||\mathcal{A}|m}{\delta}\right)$,

$$\text{KL-UCB}\left(0, N(s,a)\right) = \max\left\{\mu \in [0,1] : d(0,\mu) \leq \frac{\beta^{\text{KL-UCB}}}{N(s,a)}\right\} \geq \bar{r}(s,a),$$

holds with probability at least $1 - \frac{\delta}{4|\mathcal{S}||\mathcal{A}|m}$. Now consider the following random variables:

$$X_1 = \widetilde{Q}^{*}(s,a) - \gamma \sum_{s'} \bar{p}\left(s' \middle| s,a\right) \max_{a'} \widetilde{Q}^{**}\left(s',a'\right), \qquad X_2 = \bar{r}(s,a) = 0.$$

We have

$$\mathbb{P}\left(X_1 \geq \underbrace{\frac{\beta^{\text{obs}}}{\sqrt{N(s,a)}}}_{Y_1}\right) \leq \frac{\delta}{4\left|\mathcal{S}\right|\left|\mathcal{A}\right|m},$$

$$\mathbb{P}\left(X_2 \geq \underbrace{\max\left\{\mu \in [0,1] : d(0,\mu) \leq \frac{\beta^{\text{KL-UCB}}}{N(s,a)}\right\}}_{Y_2}\right) \leq \frac{\delta}{4\left|\mathcal{S}\right|\left|\mathcal{A}\right|m}.$$

Using Corollary B.0.1 we have

$$\mathbb{P}\left(X_1 + X_2 \geq Y_1 + Y_2\right) \leq \frac{\delta}{2\left|\mathcal{S}\right|\left|\mathcal{A}\right|m}.$$

Thus, with probability at least $1 - \frac{\delta}{2|\mathcal{S}||\mathcal{A}|m}$ we must have that $X_1 + X_2 \leq Y_1 + Y_2$. By replacing the explicit values of X_1 and X_2, we have

$$\begin{aligned}
\widetilde{Q}^*(s,a) - \gamma \sum_{s'} \bar{p}\left(s' \middle| s,a\right) \max_{a'} \widetilde{Q}^{**}\left(s',a'\right) &\leq Y_1 + Y_2 \\
\widetilde{Q}^*(s,a) &\leq \gamma \sum_{s'} \bar{p}\left(s' \middle| s,a\right) \max_{a'} \widetilde{Q}^{**}\left(s',a'\right) + Y_1 + Y_2 \\
\widetilde{Q}^*(s,a) &\leq \widetilde{Q}^{**}(s,a) + \max_{\mu}\left\{\mu \in [0,1] : d(0,\mu) \leq \frac{\beta^{\text{KL-UCB}}}{N(s,a)}\right\}.
\end{aligned}$$

By abusing the notation for $\widetilde{Q}^{**}$ to incorporate the maximization term, $\widetilde{Q}^*(s,a) \leq \widetilde{Q}^{**}(s,a)$. By using the union bound over $\mathcal{S}, \mathcal{A}$ and m the above inequality holds for all state-actions until they are visited m times with probability at least $1 - \frac{\delta}{2}$. □

C.2 Proof of Lemma 3.2

Proof. Let

$$\Delta \coloneqq \max_{(s,a)} \left|Q_1^\pi(s,a) - Q_2^\pi(s,a)\right|, \quad \Delta^{\text{e}} \coloneqq r_1^{\text{e}}\left(s^{\text{e}},a^{\text{e}}\right) - r_2^{\text{e}}\left(s^{\text{e}},a^{\text{e}}\right), \quad \Delta^{\text{m}} \coloneqq r_1^{\text{m}}\left(s^{\text{m}},a^{\text{m}}\right) - r_2^{\text{m}}\left(s^{\text{m}},a^{\text{m}}\right),$$

$$\Delta^{\text{p}} \coloneqq \gamma \sum_{s'} p_1\left(s' \middle| s,a\right) V_1^\pi\left(s'\right) - \gamma \sum_{s'} p_2\left(s' \middle| s,a\right) V_2^\pi\left(s'\right).$$

Then,

$$
\begin{aligned}
&|Q_1^\pi(s,a) - Q_2^\pi(s,a)| = |\Delta^{\mathrm{e}} + \Delta^{\mathrm{m}} + \Delta^{\mathrm{p}}| \le |\Delta^{\mathrm{e}}| + |\Delta^{\mathrm{m}}| + |\Delta^{\mathrm{p}}| \\
&\le \varphi^{\mathrm{e}} + \varphi^{\mathrm{m}} + \gamma \left| \sum_{s'} \left(p_1\left(s'\middle|s,a\right) V_1^\pi\left(s'\right) - p_2\left(s'\middle|s,a\right) V_2^\pi\left(s'\right) \right) \right| \\
&= \varphi^{\mathrm{e}} + \varphi^{\mathrm{m}} + \\
&\gamma \left| \sum_{s'} \left(p_1\left(s'\middle|s,a\right) V_1^\pi\left(s'\right) - p_1\left(s'\middle|s,a\right) V_2^\pi\left(s'\right) + p_1\left(s'\middle|s,a\right) V_2^\pi\left(s'\right) - p_2\left(s'\middle|s,a\right) V_2^\pi\left(s'\right) \right) \right| \\
&= \varphi^{\mathrm{e}} + \varphi^{\mathrm{m}} + \gamma \left| \sum_{s'} \left(p_1\left(s'\middle|s,a\right) \left(V_1^\pi\left(s'\right) - V_2^\pi\left(s'\right)\right) + \left(p_1\left(s'\middle|s,a\right) - p_2\left(s'\middle|s,a\right)\right) V_2^\pi\left(s'\right) \right) \right| \\
&\le \varphi^{\mathrm{e}} + \varphi^{\mathrm{m}} + \gamma \left| \sum_{s'} p_1\left(s'\middle|s,a\right) \left(V_1^\pi\left(s'\right) - V_2^\pi\left(s'\right)\right) \right| + \gamma \left| \sum_{s'} \left(p_1\left(s'\middle|s,a\right) - p_2\left(s'\middle|s,a\right)\right) V_2^\pi\left(s'\right) \right| \\
&\le \varphi^{\mathrm{e}} + \varphi^{\mathrm{m}} + \gamma \left| \sum_{s'} p_1\left(s'\middle|s,a\right) \left(V_1^\pi\left(s'\right) - V_2^\pi\left(s'\right)\right) \right| + \frac{2\gamma\varphi\left(r_{\max}^{\mathrm{e}} + r_{\max}^{\mathrm{m}}\right)}{1-\gamma}.
\end{aligned}
$$

By taking the $\max_{(s,a)}$ from the both sides we have

$$
\begin{aligned}
\Delta &\le \varphi^{\mathrm{e}} + \varphi^{\mathrm{m}} + \gamma\Delta + \frac{2\gamma\varphi\left(r_{\max}^{\mathrm{e}} + r_{\max}^{\mathrm{m}}\right)}{1-\gamma} \\
(1-\gamma)\Delta &\le \varphi^{\mathrm{e}} + \varphi^{\mathrm{m}} + \frac{2\gamma\varphi\left(r_{\max}^{\mathrm{e}} + r_{\max}^{\mathrm{m}}\right)}{1-\gamma} \\
\Delta &\le \frac{\varphi^{\mathrm{e}} + \varphi^{\mathrm{m}}}{1-\gamma} + \frac{2\gamma\varphi\left(r_{\max}^{\mathrm{e}} + r_{\max}^{\mathrm{m}}\right)}{(1-\gamma)^2} \\
\Delta &\le \frac{\varphi^{\mathrm{e}} + \varphi^{\mathrm{m}} + 2\gamma\varphi(r_{\max}^{\mathrm{e}} + r_{\max}^{\mathrm{m}})}{(1-\gamma)^2}.
\end{aligned}
$$

Since $|Q_1^\pi(s,a) - Q_2^\pi(s,a)| \le \Delta$, the proof is completed. □

C.3 Proof of Lemma 3.3

Proof. Using Lemma 3.2, we should show that

$$
\frac{\varphi^{\mathrm{e}} + \varphi^{\mathrm{m}} + 2\varphi\gamma(r_{\max}^{\mathrm{e}} + r_{\max}^{\mathrm{m}})}{(1-\gamma)^2} = 2\varphi\frac{1 + \gamma(r_{\max}^{\mathrm{e}} + r_{\max}^{\mathrm{m}})}{(1-\gamma)^2} \le \epsilon,
$$

which yields:

$$
\varphi \le \frac{\epsilon(1-\gamma)^2}{2\left(1 + \gamma(r_{\max}^{\mathrm{e}} + r_{\max}^{\mathrm{m}})\right)}.
$$

By our assumption that $r_{\max}^{\mathrm{m}} = r_{\max}^{\mathrm{e}} = 1$, choosing $C = \frac{1}{2+2(r_{\max}^{\mathrm{e}}+r_{\max}^{\mathrm{m}})} = \frac{1}{6}$ completes the proof. □

C.4 Proof of Lemma 3.4

In truthful Mon-MDPs the agent can face two kinds of joint state-actions: 1) state-actions that lead to observing the environment reward e.g., moving and asking for reward 2) state -action pairs that do not lead to observing the environment reward e.g., moving and not asking for reward. Let us denote these sets as the **observable** and the **unobservable** respectively.

Proof.

Number of samples for the observable set. Since we have three unknown quantities $r^{\mathrm{e}}, r^{\mathrm{m}}$, and p —that are mappings from different input spaces— we need Lemmas 3.2 and 3.3 that are straight adaptations of Strehl and Littman [47, Lemmas 1 and 2].

Using Lemma 3.3 if we want to find an $\epsilon-$minimax-optimal policy for the state-actions that are in the observable set, by choosing $\tau = \frac{1}{6}\epsilon(1-\gamma)^2$ we must have:

$$\left|\bar{r}^{\mathrm{e}}\left(s^{\mathrm{e}}, a^{\mathrm{e}}\right)-r^{\mathrm{e}}\left(s^{\mathrm{e}}, a^{\mathrm{e}}\right)\right| \leq \tau, \quad\left|\bar{r}^{\mathrm{m}}\left(s^{\mathrm{m}}, a^{\mathrm{m}}\right)-r^{\mathrm{m}}\left(s^{\mathrm{m}}, a^{\mathrm{m}}\right)\right| \leq \tau, \quad \|\bar{p}(\cdot \mid s, a)-p(\cdot \mid s, a)\|_1 \leq \tau .$$

On the other hand, we know that if $(s, a) \equiv (s^{\mathrm{e}}, s^{\mathrm{m}}, a^{\mathrm{e}}, a^{\mathrm{m}})$ has been visited $N(s, a)$ times, its monitor reward has been observed $N\left(s^{\mathrm{m}}, a^{\mathrm{m}}\right)$ times, and its environment reward has been observed $N\left(s^{\mathrm{e}}, a^{\mathrm{e}}\right)$ times, with probabilities at least $1-\delta^{\mathrm{e}}, 1-\delta^{\mathrm{m}}$, and $1-\delta$:

$$\|\bar{p}(\cdot \mid s, a)-p(\cdot \mid s, a)\|_1 \leq \sqrt{\frac{2\left[\ln \left(2^{|\mathcal{S}|}-2\right)-\ln \delta\right]}{N(s, a)}} \tag{C.1}$$

$$\left|\bar{r}^{\mathrm{m}}\left(s^{\mathrm{m}}, a^{\mathrm{m}}\right)-r^{\mathrm{m}}\left(s^{\mathrm{m}}, a^{\mathrm{m}}\right)\right| \leq \sqrt{\frac{2 \ln (2 / \delta^{\mathrm{m}})}{N\left(s^{\mathrm{m}}, a^{\mathrm{m}}\right)}} \tag{C.2}$$

$$\left|\bar{r}^{\mathrm{e}}\left(s^{\mathrm{e}}, a^{\mathrm{e}}\right)-r^{\mathrm{e}}\left(s^{\mathrm{e}}, a^{\mathrm{e}}\right)\right| \leq \sqrt{\frac{2 \ln (2 / \delta^{\mathrm{e}})}{N\left(s^{\mathrm{e}}, a^{\mathrm{e}}\right)}} \tag{C.3}$$

Thus, in order to find m, the least number of visits to (s, a), we make connections between $m, N(s, a), N\left(s^{\mathrm{m}}, a^{\mathrm{m}}\right)$, and $N\left(s^{\mathrm{e}}, a^{\mathrm{e}}\right)$. If a joint state-action is visited m times, then:

$$m=N(s, a), \quad m \leq N\left(s^{\mathrm{m}}, a^{\mathrm{m}}\right) \leq \sum_{s^{\mathrm{e}}, a^{\mathrm{e}}} m=|\mathcal{S}^{\mathrm{e}}||\mathcal{A}^{\mathrm{e}}| m, \quad m \cdot \rho \leq N\left(s^{\mathrm{e}}, a^{\mathrm{e}}\right),$$

where the last inequality follows from the fact that the environment reward is observed with probability ρ upon vising (s, a).

If we want Equations (C.1), (C.2) and (C.3) hold simultaneously with probability at $1 - \delta$ until (s, a) is visited m times, by setting $\delta = \delta^{\mathrm{m}} = \frac{\delta}{3|\mathcal{S}||\mathcal{A}|m}$ and $\delta^{\mathrm{e}} = \frac{\delta}{3|\mathcal{S}||\mathcal{A}|\rho m}$ to split the failure probability equally for rewards and transitions of all state-actions until each of them have been visited m times, it is enough ensure τ is bigger than the length of the confidence intervals:

$$\begin{aligned} m &\geq \max\left\{\frac{8\left[\ln\left(2^{|\mathcal{S}|}-2\right)-\ln\delta\right]}{\tau^2}, \frac{8\ln\left(2/\delta^{\mathrm{m}}\right)}{\tau^2}, \frac{8\ln\left(2/\delta^{\mathrm{e}}\right)}{\tau^2}\right\} \\ &\geq \max\left\{\frac{8\left[\ln\left(2^{|\mathcal{S}|}-2\right)-\ln\delta\right]}{\tau^2}, \frac{8\ln\left(2/\delta^{\mathrm{e}}\right)}{\rho\tau^2}\right\} \\ &\geq \max\left\{\frac{8\left[\ln\left(2^{|\mathcal{S}|}-2\right)+\ln\frac{3|\mathcal{S}||\mathcal{A}|m}{\delta}\right]}{\tau^2}, \frac{8\ln\frac{6|\mathcal{S}||\mathcal{A}|\rho m}{\delta}}{\rho\tau^2}\right\} \end{aligned}$$

If $\rho^{-1} \geq \mathcal{O}\left(|\mathcal{S}|\right)$, then

$$m \geq \frac{8\ln\left(\frac{6|\mathcal{S}||\mathcal{A}|\rho m}{\delta}\right)}{\rho\tau^2}$$

and by Lemma B.1, we have

$$m = \mathcal{O}\left(\frac{1}{\rho\tau^2}\ln\frac{|\mathcal{S}|\,|\mathcal{A}|}{\tau\delta}\right) = \mathcal{O}\left(\frac{1}{\rho\epsilon^2(1-\gamma)^4}\ln\frac{|\mathcal{S}|\,|\mathcal{A}|}{\epsilon(1-\gamma)^2\delta}\right) \tag{C.4}$$

If $\rho^{-1} \leq \mathcal{O}\left(|\mathcal{S}|\right)$,

$$m \geq \frac{8\left[\ln\left(2^{|\mathcal{S}|}-2\right)+\ln\frac{3|\mathcal{S}||\mathcal{A}|m}{\delta}\right]}{\tau^2},$$

which by Lemma B.1 implies

$$m = \mathcal{O}\left(\frac{|\mathcal{S}|}{\tau^2}+\frac{1}{\tau^2}\ln\frac{|\mathcal{S}|\,|\mathcal{A}|}{\tau\delta}\right) = \mathcal{O}\left(\frac{|\mathcal{S}|}{\epsilon^2(1-\gamma)^4}+\frac{1}{\epsilon^2(1-\gamma)^4}\ln\frac{|\mathcal{S}|\,|\mathcal{A}|}{\epsilon(1-\gamma)^2\delta}\right). \tag{C.5}$$

Number of samples for the unobservable set. These state-actions cannot change the sample estimate of the mean environment reward and the only quantities updated upon visits are the transition dynamics and the monitor reward. It is enough to have

$$m \geq \max\left\{\frac{8\left[\ln\left(2^{|\mathcal{S}|}-2\right)-\ln\delta\right]}{\tau^2}, \frac{8\ln\left(2/\delta^{\mathrm{m}}\right)}{\tau^2}\right\}.$$

Hence, similar to the above case when $\rho^{-1} \leq \mathcal{O}(|\mathcal{S}|)$, the dominant factor around learning the sample estimates would be the transitions and the required sample size would be

$$m = \mathcal{O}\left(\frac{|\mathcal{S}|}{\epsilon^2(1-\gamma)^4} + \frac{1}{\epsilon^2(1-\gamma)^4} \ln \frac{|\mathcal{S}|\,|\mathcal{A}|}{\epsilon(1-\gamma)^2\delta}\right). \tag{C.6}$$

The total number of samples in the worst-case is obtained when $\rho^{-1} \geq \mathcal{O}(|\mathcal{S}|)$ and

$$m = \mathcal{O}\left(\frac{1}{\rho\epsilon^2(1-\gamma)^4} \ln \frac{|\mathcal{S}|\,|\mathcal{A}|}{\epsilon(1-\gamma)^2\delta}\right). \qquad \square$$

C.5 Proof of Lemma 3.5

Proof. Fix $t \geq 0$. For some fixed partial path $P_t = S_0, A_0, R_1, \ldots, S_{t-1}, A_{t-1}$, where $R_i \coloneqq R_i^{\mathrm{e}} + R_i^{\mathrm{m}}, 1 \leq i \leq t-2$. Let $\mathcal{K}_t$ be the set of all paths P_t such that every state-action (S_i, A_i) with $0 \leq i \leq t-1$ appearing in P_t is *known*. Let $\mathbb{P}_{\pi,M}$ be the probability measure induced by executing π in M. Let $R_t(M)$ be the reward received by the agent at time step t in M, and $R_t(M, P_t)$ be the reward be the reward received by the agent at time step t in M given that P_t was the partial path generated, Now we have

$$\begin{aligned}
\left|\mathbb{E}\left[R_t\left(M'\right)\right] - \mathbb{E}\left[R_t(M)\right]\right| &= \Bigg|\sum_{P_t \in \mathcal{K}_t} \left[\mathbb{P}_{\pi,M'}(P_t) \cdot R_t(M', P_t) - \mathbb{P}_{\pi,M}(P_t) \cdot R_t(M, P_t)\right] + \\
&\sum_{P_t \notin \mathcal{K}_t} \left[\mathbb{P}_{\pi,M'}(P_t) \cdot R_t(M', P_t) - \mathbb{P}_{\pi,M}(P_t) \cdot R_t(M, P_t)\right]\Bigg| \\
&= \left|\sum_{P_t \notin \mathcal{K}_t} \left[\mathbb{P}_{\pi,M'}(P_t) \cdot R_t(M', P_t) - \mathbb{P}_{\pi,M}(P_t) \cdot R_t(M, P_t)\right]\right| \\
&\leq \left|\sum_{P_t \notin \mathcal{K}_t} \left[\mathbb{P}_{\pi,M'}(P_t) \cdot R_t(M', P_t)\right]\right| + \left|\sum_{P_t \notin \mathcal{K}_t} \left[-\mathbb{P}_{\pi,M}(P_t) \cdot R_t(M, P_t)\right]\right| \\
&\leq \left|\sum_{P_t \notin \mathcal{K}_t} \left[\mathbb{P}_{\pi,M'}(P_t) \cdot R_t(M', P_t)\right]\right| \\
&\leq \frac{\mathbb{P}_{\pi,M'}(\mathcal{E}_{M'})}{2} \\
&= \frac{\mathbb{P}_{\pi,M}(\mathcal{E}_{M})}{2} \\
&= \frac{\mathbb{P}\left(\mathcal{E}_{M}\right)}{2}.
\end{aligned}$$

The first step separates the possible paths in which the agent encounters an unknown state-action from those in which only known state-actions are reached. We eliminate the first term, because M and M' are identical on known state-actions. The third step uses triangle inequality. The fifth step uses the fact that normalized rewards are at most 1. The sixth step follows from the fact that M and M' are identical on known state-actions and the probability of the first encounter of the unknown state-action is the same. The last step follows from the definition of the induced probability measure. The results then follows:

$$\begin{aligned}
\left|V^{\pi}_{M'}(S_0)_H - V^{\pi}_{M}(S_0)_H\right| &\leq \sum_{t=0}^{H-1} \gamma^t \left|\mathbb{E}\left[R_{t+1}\left(M'\right)\right] - \mathbb{E}\left[R_{t+1}(M)\right]\right| \\
&\leq \sum_{t=0}^{\infty} \gamma^t \left|\mathbb{E}\left[R_{t+1}\left(M'\right)\right] - \mathbb{E}\left[R_{t+1}(M)\right]\right| \\
&\leq \frac{\mathbb{P}\left(\mathcal{E}_M\right)}{2(1-\gamma)}.
\end{aligned}$$

□

C.6 Proof of Lemma 3.6

Proof. Suppose m is the least number of samples required for each state-action to ensure $\bar{r}^{\mathrm{m}}$, $\bar{p}$, and $\bar{r}^{\mathrm{e}}$ are close to their true mean. To be pessimistic about the environment reward in cases that $\bar{r}^{\mathrm{e}}$ cannot be computed due to its ever-lasting unobservability, we need to investigate the optimism in two cases where $\bar{r}^{\mathrm{e}}$ can be computed and when it cannot. Consider N experiences of a joint state-action $(s,a) \equiv (s^{\mathrm{e}}, s^{\mathrm{m}}, a^{\mathrm{e}}, a^{\mathrm{m}})$ and the first N^{e} experiences of $(s^{\mathrm{e}}, a^{\mathrm{e}})$ where the environment reward has been observed. Also define $\bar{V}^*_{\downarrow}$ as:

$$\bar{V}^*_{\downarrow}(s) := \max_a \bar{Q}^*_{\downarrow}(s,a) := \sum_{s'} \bar{P}\left(s' \middle| s,a\right) V^*_{\downarrow}\left(s'\right), \quad \forall s \in \mathcal{S}.$$

Case 1. N^{e} is bigger than zero. Let X_{1i}, X_{2i}, and X_{3i} be random variables defined at the ith visit as below, where S'_i is the next state visited after the ith visit:

$$X_{1i} = R^{\mathrm{e}}_i, \quad X_{2i} = R^{\mathrm{m}}_i, \quad X_{3i} = \gamma \bar{V}^*_{\downarrow}(S'_i).$$

If (s,a) has been visited N times and $R^e(s^e, a^e)$ has been observed N^{e} times, then:

- The sequence $(X_{1i})_{i=1}^{N^{\mathrm{e}}}$ is available.
- At least the sequence $(X_{2i})_{i=1}^{N}$ is available. (At most $(X_{2i})_{i=1}^{|\mathcal{S}^{\mathrm{e}}||\mathcal{A}^{\mathrm{e}}|N}$)
- The sequence $(X_{3i})_{i=1}^{N}$ is available.

Let $(X_{1i})_{i=1}^{N^{\mathrm{e}}}, (X_{2i})_{i=1}^{N}$, and $(X_{3i})_{i=1}^{N}$ be random variables on the joint probability space. By applying the Chernoff-Hoeffding's inequality:

- For X_{1i} we have

$$\mathbb{P}\left(\mathbb{E}\left[X_{11}\right]-\frac{1}{N^{\mathrm{e}}}\sum_{i=1}^{N^{\mathrm{e}}}X_{1i}\geq B_3\right)\leq\exp\left(-\frac{N^{\mathrm{e}}B_3^2}{2}\right).$$

- For X_{2i} we have

$$\mathbb{P}\left(\mathbb{E}\left[X_{21}\right]-\frac{1}{N}\sum_{i=1}^{N}X_{2i}\geq Y_2\right)\leq\exp\left(-\frac{NY_2^2}{2}\right).$$

- For X_{3i} we have that $\gamma\frac{-2}{1-\gamma}\leq X_{3i}\leq\gamma\frac{2}{1-\gamma}$ hence

$$\mathbb{P}\left(\mathbb{E}\left[X_{31}\right]-\frac{1}{N}\sum_{i=1}^{N}X_{3i}\geq Y_1\right)\leq\exp\left(-\frac{NY_1^2(1-\gamma)^2}{8\gamma^2}\right).$$

Define the following random variables on $(\Omega,\mathcal{F},\mathbb{P})$:

$$X_1=\mathbb{E}\left[X_{11}\right]-\frac{1}{N^{\mathrm{e}}}\sum_{i=1}^{N^{\mathrm{e}}}X_{1i},\quad X_2=\mathbb{E}\left[X_{21}\right]-\frac{1}{N}\sum_{i=1}^{N}X_{2i},\quad X_3=\mathbb{E}\left[X_{31}\right]-\frac{1}{N}\sum_{i=1}^{N}X_{3i}.$$

By choosing $Y_1=\frac{\beta^{\mathrm{e}}}{\sqrt{N^{\mathrm{e}}}}, Y_2=\frac{\beta^{\mathrm{m}}}{\sqrt{N}}$, and $Y_3=\frac{\beta}{\sqrt{N}}$ where

$$\beta=\frac{2\gamma}{1-\gamma}\sqrt{2\ln\left(\frac{6\left|\mathcal{S}\right|\left|\mathcal{A}\right|m}{\delta}\right)},\quad\beta^{\mathrm{m}}=\sqrt{2\ln\left(\frac{6\left|\mathcal{S}\right|\left|\mathcal{A}\right|m}{\delta}\right)},\quad\beta^{\mathrm{e}}=\sqrt{2\ln\left(\frac{6\left|\mathcal{S}\right|\left|\mathcal{A}\right|m}{\delta}\right)}.$$

Using Corollary B.0.1, we have

$$\mathbb{P}\left(X_1+X_2+X_3\geq Y_1+Y_2+Y_3\right)\leq\exp\left(-\frac{N^{\mathrm{e}}Y_3^2}{2}\right)+\exp\left(-\frac{NY_2^2}{2}\right)+\exp\left(-\frac{NY_1^2(1-\gamma)^2}{8\gamma^2}\right).$$

Thus,

$$\mathbb{P}\left(X_1+X_2+X_3\geq\frac{\beta^{\mathrm{e}}}{\sqrt{N^{\mathrm{e}}}}+\frac{\beta^{\mathrm{m}}}{\sqrt{N}}+\frac{\beta}{\sqrt{N}}\right)\leq\frac{\delta}{2\left|\mathcal{S}\right|\left|\mathcal{A}\right|m}.\tag{C.7}$$

With probability $1-\frac{\delta}{2|\mathcal{S}||\mathcal{A}|m}$ it must hold that

$$X_1+X_2+X_3\leq\left(\frac{\beta^{\mathrm{e}}}{\sqrt{N^{\mathrm{e}}}}+\frac{\beta^{\mathrm{m}}}{\sqrt{N}}+\frac{\beta}{\sqrt{N}}\right),$$

which is equal to

$$\frac{1}{N^{\mathrm{e}}}\sum_{i=1}^{k} R_i^{\mathrm{e}} + \frac{1}{N}\sum_{j=1}^{N}\left(R_j^{\mathrm{m}} + \bar{V}_{\downarrow}^{*}(S_j')\right) + \left(\frac{\beta^{\mathrm{e}}}{\sqrt{N^{\mathrm{e}}}} + \frac{\beta^{\mathrm{m}}}{\sqrt{N}} + \frac{\beta}{\sqrt{N}}\right) \geq \mathbb{E}\left[R_1^{\mathrm{e}} + R_1^{\mathrm{m}} + \gamma\bar{V}_{\downarrow}^{*}(S_1')\right]$$
$$= Q_{\downarrow}^{*}(s,a).$$

Therefore,

$$\bar{r}^{\mathrm{e}}\left(s^{\mathrm{e}},a^{\mathrm{e}}\right) + \bar{r}^{\mathrm{m}}\left(s^{\mathrm{m}},a^{\mathrm{m}}\right) + \gamma\sum_{s'}\bar{p}\left(s'\middle|s,a\right)V_{\downarrow}^{*}\left(s'\right) + \left(\frac{\beta^{\mathrm{e}}}{\sqrt{N^{\mathrm{e}}}} + \frac{\beta^{\mathrm{m}}}{\sqrt{N}} + \frac{\beta}{\sqrt{N}}\right) \geq Q_{\downarrow}^{*}(s,a). \tag{C.8}$$

Using the union bound over $\mathcal{S}, \mathcal{A}$, and m, Equation (C.8) holds for all state-actions until they are visited m times with probability at least $1-\frac{\delta}{2}$. Instead of the left-hand side of Equation (C.8), Monitored MBIE-EB uses the following action-values to relax the lack of knowledge of V^*:

$$Q_{\downarrow}^{**}(s,a) = \bar{r}^{\mathrm{e}}\left(s^{\mathrm{e}},a^{\mathrm{e}}\right) + \bar{r}^{\mathrm{m}}\left(s^{\mathrm{m}},a^{\mathrm{m}}\right) + \gamma\sum_{s'}\bar{p}\left(s'\middle|s,a\right)V_{\downarrow}^{**}\left(s'\right) + \left(\frac{\beta^{\mathrm{e}}}{\sqrt{N^{\mathrm{e}}}} + \frac{\beta^{\mathrm{m}}}{\sqrt{N}} + \frac{\beta}{\sqrt{N}}\right). \tag{C.9}$$

Following the induction of Strehl and Littman [47, Lemma 7], we prove $Q_{\downarrow}^{**}(s,a) \geq Q_{\downarrow}^{*}(s,a)$. Let

$$C = \left(\frac{\beta}{\sqrt{N}} + \frac{\beta^{\mathrm{m}}}{\sqrt{N}} + \frac{\beta^{\mathrm{e}}}{\sqrt{N^{\mathrm{e}}}}\right).$$

Proof by induction is on the number of value iteration steps. Let $Q_{\downarrow}^{**}(s,a)_i$ be the ith iterate of the value iteration for (s,a). By the optimistic initialization we have that $Q_{\downarrow}^{**}(s,a)_0 \geq Q_{\downarrow}^{*}(s,a)$ for all state-actions. Now suppose the claim holds for $Q_{\downarrow}^{**}(s,a)_i$, we have

$$\begin{aligned}
Q_{\downarrow}^{**}(s,a)_{i+1} &= \bar{r}^{\mathrm{e}}\left(s^{\mathrm{e}},a^{\mathrm{e}}\right) + \bar{r}^{\mathrm{m}}\left(s^{\mathrm{m}},a^{\mathrm{m}}\right) + \gamma\sum_{s'}\bar{p}\left(s'\middle|s,a\right)\max_{a'}Q_{\downarrow}^{**}\left(s',a'\right)_i + C \\
&= \bar{r}^{\mathrm{e}}\left(s^{\mathrm{e}},a^{\mathrm{e}}\right) + \bar{r}^{\mathrm{m}}\left(s^{\mathrm{m}},a^{\mathrm{m}}\right) + \gamma\sum_{s'}\bar{p}\left(s'\middle|s,a\right)V_{\downarrow}^{**}\left(s'\right)_i + C \\
&\geq \bar{r}^{\mathrm{e}}\left(s^{\mathrm{e}},a^{\mathrm{e}}\right) + \bar{r}^{\mathrm{m}}\left(s^{\mathrm{m}},a^{\mathrm{m}}\right) + \gamma\sum_{s'}\bar{p}\left(s'\middle|s,a\right)V_{\downarrow}^{*}\left(s'\right) && \text{(Using induction)} \\
&\geq Q_{\downarrow}^{*}(s,a). && \text{(Equation (C.8))}
\end{aligned}$$

Case 2. N^{e} is zero. If N^{e} is zero for (s,a), then Monitored MBIE-EB assigns $-r_{\max}^{\mathrm{e}}$ to $\bar{r}\left(s^{\mathrm{e}},a^{\mathrm{e}}\right)$ deterministically. Thus, the previously random variable X_1 in Case 1, is deterministically zero and there would be no randomness around it. Consequently, Equation (C.7) is turned into

$$\mathbb{P}\left(X_2 + X_3 \geq \frac{\beta^{\mathrm{m}}}{\sqrt{N}} + \frac{\beta}{\sqrt{N}}\right) \leq \frac{\delta}{3\left|\mathcal{S}\right|\left|\mathcal{A}\right|m},$$

where β and β^{m} are as before. Then, with probability $1 - \frac{\delta}{3|\mathcal{S}||\mathcal{A}|m}$ it must hold that

$$X_2 + X_3 \leq \left(\frac{\beta^{\mathrm{m}}}{\sqrt{N}} + \frac{\beta}{\sqrt{N}} \right),$$

which is equal to

$$\frac{1}{N} \sum_{j=1}^{N} \left(R_j^{\mathrm{m}} + \gamma \bar{V}_{\downarrow}^{*}(S_j') \right) + \left(\frac{\beta^{\mathrm{m}}}{\sqrt{N}} + \frac{\beta}{\sqrt{N}} \right) \geq \mathbb{E} \left[R_1^{\mathrm{m}} + \gamma \bar{V}_{\downarrow}^{*}(S_1') \right] = Q_{\downarrow}^{*}(s,a) - (-r_{\max}^{\mathrm{e}}).$$

Therefore,

$$-r_{\max}^{\mathrm{e}} + \bar{r}^{\mathrm{m}}\left(s^{\mathrm{m}}, a^{\mathrm{m}}\right) + \gamma \sum_{s'} \bar{p}\left(s' \middle| s, a\right) V_{\downarrow}^{*}\left(s'\right) + \left(\frac{\beta}{\sqrt{N}} + \frac{\beta^{\mathrm{m}}}{\sqrt{N}} \right) \geq Q_{\downarrow}^{*}(s,a). \tag{C.10}$$

The rest of the proof is identical to the induction steps of case 1 with probability at least $1 - \frac{\delta}{3}$. Considering both cases, with probability at least $1 - \frac{\delta}{2} - \frac{\delta}{3} = 1 - \frac{5\delta}{6}$

$$Q_{\downarrow}^{**}(s,a) \geq Q_{\downarrow}^{*}(s,a). \qquad \square$$

Appendix D

Methodological Details and Results

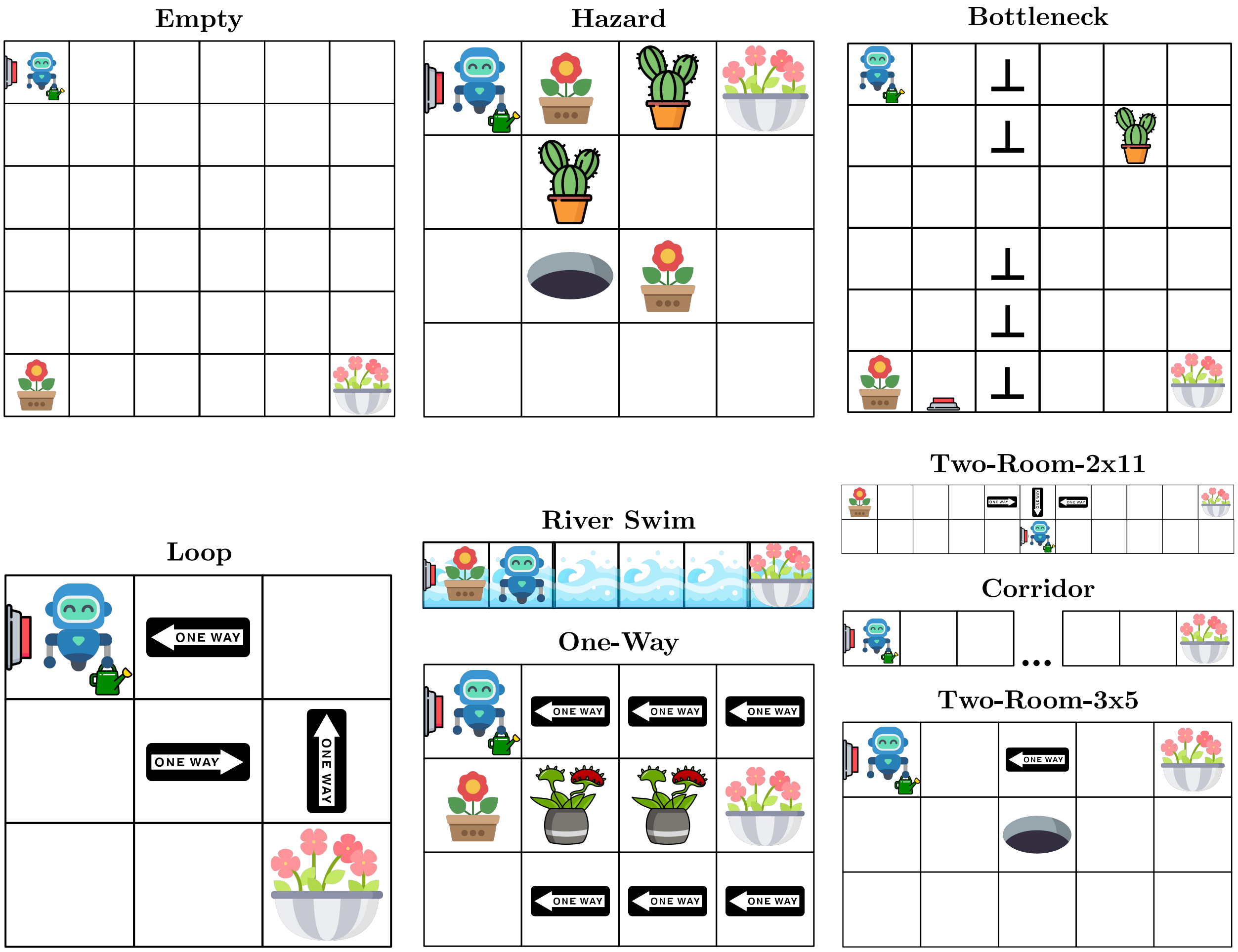
To clarify details used in our experiments, in this chapter we provide finer details around our empirical evaluation. We explain the environments' dynamics, monitors and hyperparameters. We report results of experiments containing unsolvable Mon-MDPs and when the monitor is known. Also, we report ablation studies to highlight the significance of Monitored MBIE-EB's components.

D.1 Empirical Evaluation Details

This section defines the metric used in our experiments to assess performance. We report the discounted test return, averaged over 30 random seeds with their corresponding 95% confidence intervals. To compute this metric, training is paused every 100 steps, the agent is tested over 100 episodes, and the average obtained return is recorded as a data point before training resumes.

D.1.1 Environments' Details

We provide the environments' details to clarify on what environments the algorithms are tested and what type of behavior is desirable. Environments that comprise the experiments are: **Empty**, **Hazard**, **Bottleneck**, **Loop**, **River Swim**, **One-Way**, **Corridor**, **Two-Room-3x5** and **Two-Room-2x11** shown in Figure D.1. In all of them (except River Swim) the agent, represented by the robot, has 5 actions including 4 cardinal movement {`LEFT, DOWN, RIGHT, UP`} and an extra `WATER` action. The agent's goal is get to the big flower pot as fast as possible. The agent should `WATER` the big flower pot to get a reward of +1 which also terminates the episode. The agent should not `WATER` the small flower pots as they are distractors; watering them would yield a reward of 0.1 and terminates the episode as well. Cacti should be avoided as any action leading to their states results in rewards of -10. Flytraps yield a reward of -0.1. Cells with a one-way sign transition



Empty

Hazard

Bottleneck

Two-Room-2x11

River Swim

Loop

Corridor

One-Way

Two-Room-3x5

Figure D.1: Full set of environments. Except Bottleneck, all of the environments are borrowed from Parisi et al. [39]. Cacti and Flytraps should be avoided. The goal is to water the big four flower pot. Small single flower pots are distractors. The agent gets stuck in the holes unless randomly gets pulled out. One-ways transition the agent in their own direction regardless of the action.

the agent only to their unique direction and if the agent stumbles on a hole, it would spend the whole episode in the hole, unless with 10% probability the taken action is effective and the agent gets transitioned. When a button monitor is configured on top of the environment, the location of the button is figuratively is indicated by a push button placed on a cell's border. It shows the button is pushed if agent bumps itself into that cell's border. The episode's time limit in River Swim, corridor and Two-Room-2x11 is 200 steps, and in other environments is 50 steps. In River Swim the agent has only two actions `L` ≡ `LEFT` and `R` ≡ `RIGHT`. There is no termination except the episode's time limit. In this environment small flower pot have a reward of 0.01. Rivr Swim is the only environment in our experiment suite that has stochastic transitions shown in Figure D.2.

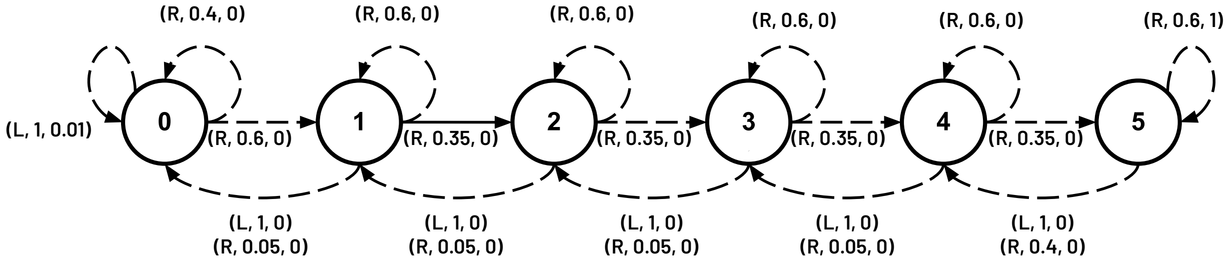


Figure D.2: Dynamics of River Swim. Each tuple represents (action, transition probability, reward).

D.1.2 Monitors' Details

In this section, we provide the monitors' details used in our experiments. These details are useful when evaluating the performance of algorithms to see how the algorithm could overcome challenges. Monitors that comprise the experiments are: **Full (MDP)**, **Semi-Random**, **Full-Random**, **Ask**, **Button**, N**-Supporters**, N**-Experts** and **Level-Up**. For any of the monitors, except **Full-Monitor**, if a cell in the environment is marked with $\perp$, then under no circumstances and at no time step, the monitor would reveal the environment reward to agent for the action that led agent to that cell. For *the rest* of the environment state-action pairs the behavior of monitors, by letting $X_t \sim \mathcal{U}[0, 1]$, where $\mathcal{U}$ is the uniform distribution and $\rho \in [0, 1]$, is as follows:

- **MDP**. This corresponds to the MDP setting:

$$\mathcal{S}^{\text{m}} \coloneqq \{\texttt{ON}\}, \qquad \mathcal{A}^{\text{m}} \coloneqq \{\texttt{NO-OP}\}, \qquad S^{\text{m}}_{t+1} \coloneqq \texttt{ON}, \qquad R^{\text{m}}_{t+1} \coloneqq 0, \qquad \widehat{R}^{\text{e}}_{t+1} \coloneqq R^{\text{e}}_{t+1}.$$

- **Semi-Random**. Like an MDP, but the monitor hides non-zero rewards half the time.:

$$\mathcal{S}^{\text{m}} \coloneqq \{\texttt{ON}\}, \quad \mathcal{A}^{\text{m}} \coloneqq \{\texttt{NO-OP}\}, \quad S^{\text{m}}_{t+1} \coloneqq \texttt{ON,} \quad R^{\text{m}}_{t+1} \coloneqq 0. \quad \widehat{R}^{\text{e}}_{t+1} \coloneqq \begin{cases} R^{\text{e}}_{t+1}, & \textbf{if } R^{\text{e}}_{t+1} = 0; \\ R^{\text{e}}_{t+1}, & \textbf{if } X_t \leq 0.5 \\ \perp, & \text{Otherwise.} \end{cases}$$

- **Full-Random**. It is similar to Semi-Random except that the monitor hides *any* environment reward with a predefined probability $1 - \rho$:

$$\mathcal{S}^{\text{m}} \coloneqq \{\texttt{ON}\}, \quad \mathcal{A}^{\text{m}} \coloneqq \{\texttt{NO-OP}\}, \quad S^{\text{m}}_{t+1} \coloneqq \texttt{ON}, \quad R^{\text{m}}_{t+1} \coloneqq 0, \quad \widehat{R}^{\text{e}}_{t+1} \coloneqq \begin{cases} R^{\text{e}}_{t+1} & \textbf{if } , X_t \leq \rho; \\ \perp, & \text{Otherwise.} \end{cases}$$

- **Ask**. The monitor state space is a singleton but its action space has two elements: $\{\texttt{ASK}, \texttt{NO-OP}\}$. The agent gets to see the environment reward with probability ρ if it ASKs. Upon asking agent pays -0.2 as the monitor reward:

$$\mathcal{S}^{\text{m}} \coloneqq \{\texttt{ON}\}, \qquad \mathcal{A}^{\text{m}} \coloneqq \{\texttt{ASK, NO-OP}\}, \qquad S^{\text{m}}_{t+1} \coloneqq \texttt{ON},$$

$$\widehat{R}^{\text{e}}_{t+1} \coloneqq \begin{cases} R^{\text{e}}_{t+1}, & \textbf{if } X_t \leq \rho \textbf{ and } A^{\text{m}}_t = \texttt{ASK}; \\ \bot, & \text{Otherwise;} \end{cases} \qquad R^{\text{m}}_{t+1} \coloneqq \begin{cases} -0.2, & \textbf{if } A^{\text{m}}_t = \texttt{ASK}; \\ 0, & \text{Otherwise.} \end{cases}$$

- **Button**. The state space is {`ON, OFF`}. The action space is a singleton. The agent sees the environment reward with probability ρ as long as the monitor is `ON`, while paying -0.2 as the monitor cost. The monitor state is flipped if the agent bumps itself to the button:

$$\mathcal{S}^{\text{m}} \coloneqq \{\texttt{OFF, ON}\}, \qquad \mathcal{A}^{\text{m}} \coloneqq \{\texttt{NO-OP}\}, \qquad \widehat{R}^{\text{e}}_{t+1} \coloneqq \begin{cases} R^{\text{e}}_{t+1}, & \textbf{if } X_t \leq \rho \textbf{ and } S^{\text{m}}_t = \texttt{ON}; \\ \bot, & \text{Otherwise;} \end{cases}$$

$$S^{\text{m}}_{t+1} \coloneqq \begin{cases} \texttt{ON}, & \textbf{if } S^{\text{m}}_t = \texttt{OFF} \textbf{ and } S^{\text{e}}_t = \texttt{"BUTTON-CELL"} \textbf{ and } A^{\text{e}}_t = \texttt{"BUMP-INTO-BUTTON"}; \\ \texttt{OFF}, & \textbf{if } S^{\text{m}}_t = \texttt{ON} \textbf{ and } S^{\text{e}}_t = \texttt{"BUTTON-CELL"} \textbf{ and } A^{\text{e}}_t = \texttt{"BUMP-INTO-BUTTON"}; \\ S^{\text{m}}_t, & \text{Otherwise;} \end{cases}$$

$$S^{\text{m}}_0 \coloneqq \text{Random uniform from } \mathcal{S}^{\text{m}}, \qquad R^{\text{m}}_{t+1} \coloneqq \begin{cases} -0.2, & \textbf{if } S^{\text{m}}_t = \texttt{ON}; \\ 0, & \text{Otherwise.} \end{cases}$$

- N**-Supporters**. The monitor state space comprises N states each representing the presence of a supporter. The action space also comprises N actions. At each time step one of the supporters is randomly present and if the agent could choose the action that matches the present supporter's index, then the agent gets to see the environment reward with probability ρ. Upon observing the environment reward, the agent pays a penalty of -0.2 as the monitor reward. However, if the agent chooses a wrong supporter, then it will be rewarded 0.001 (as a distractor):

$$\mathcal{S}^{\text{m}} \coloneqq \{0, \cdots, N-1\}, \qquad \mathcal{A}^{\text{m}} \coloneqq \{0, \cdots, N-1\}, \qquad S^{\text{m}}_{t+1} \coloneqq \text{Random uniform from } \mathcal{S}^{\text{m}},$$

$$\widehat{R}^{\text{e}}_{t+1} \coloneqq \begin{cases} R^{\text{e}}_{t+1}, & \textbf{if } X_t \leq \rho \textbf{ and } S^{\text{m}}_t = A^{\text{m}}_t; \\ \bot, & \text{Otherwise;} \end{cases} \qquad R^{\text{m}}_{t+1} \coloneqq \begin{cases} -0.2, & S^{\text{m}}_t = A^{\text{m}}_t; \\ 0.001, & \text{Otherwise.} \end{cases}$$

Parisi et al. [39] considered this monitor as challenging, due to its big spaces, for algorithms that use the successor representations. Yet, the encouraging nature of the monitor regarding the agent's mistakes makes it easy for reward-respecting algorithms, e.g., Monitored MBIE-EB.

- N**-Experts**. Similar to N-Supporter the state space has N states, each corresponding to the presence of one of the N experts. However, experts' advice is costly, hence the action space has $N+1$ action where the last action corresponds to not pinging any experts and is cost-free. At each time step, one of the experts is randomly present and if the agent selects the action that matches the present expert's index, the agent gets to see the environment reward with probability ρ. Upon observing the environment reward agent pays a penalty of -0.2 as the monitor reward. However, if the agent chooses a wrong expert it will be penalized by -0.001 as the monitor reward. Since the last action does not inquire any of the experts its monitor reward is zero:

$$\mathcal{S}^{\mathrm{m}} \coloneqq \{0, \cdots, N-1\}, \qquad \mathcal{A}^{\mathrm{m}} \coloneqq \{0, \cdots, N\}, \qquad S^{\mathrm{m}}_{t+1} \coloneqq \text{Random uniform from } \mathcal{S}^{\mathrm{m}},$$

$$\widehat{R}^{\mathrm{e}}_{t+1} \coloneqq \begin{cases} R^{\mathrm{e}}_{t+1}, & \textbf{if } X_t \leq \rho \textbf{ and } S^{\mathrm{m}}_t = A^{\mathrm{m}}_t; \\ \bot, & \text{Otherwise;} \end{cases} \qquad R^{\mathrm{m}}_{t+1} \coloneqq \begin{cases} -0.2, & \textbf{if } S^{\mathrm{m}}_t = A^{\mathrm{m}}_t; \\ 0, & \textbf{if } A^{\mathrm{m}}_t = N; \\ -0.001, & \text{Otherwise.} \end{cases}$$

- **Level-Up**. This monitor tries to test the agent's capabilities of performing deep exploration [37] in the monitor spaces. The state space has N states corresponding to N levels. The action space has $N+1$ actions. The initial state of the monitor is 0 and if at each time step the agent selects the action that matches the state of the monitor, the state increases by one. If the agent selects the wrong action the state is reset back to 0. The agent only gets to observe the environment reward with probability ρ if it takes the state of the monitor to the max level. The agent is penalized with -0.2 as the monitor reward every time it does not select the last action which does nothing and keeps the monitor state as it is:

$$\mathcal{S}^{\mathrm{m}} \coloneqq \{0, \cdots, N-1\}, \quad \mathcal{A}^{\mathrm{m}} \coloneqq \{0, \cdots, N-1, \texttt{NO-OP}\}, \quad R^{\mathrm{m}}_{t+1} \coloneqq \begin{cases} 0, & \textbf{if } A^{\mathrm{m}}_t = \texttt{NO-OP}; \\ -0.2, & \text{Otherwise;} \end{cases}$$

$$\widehat{R}^{\mathrm{e}}_{t+1} \coloneqq \begin{cases} R^{\mathrm{e}}_{t+1}, & \textbf{if } X_t \leq \rho \textbf{ and } S^{\mathrm{m}}_t = N-1; \\ \bot, & \text{Otherwise;} \end{cases}$$

$$S^{\mathrm{m}}_{t+1} \coloneqq \begin{cases} S^{\mathrm{m}}_t, & \textbf{if } A^{\mathrm{m}}_t = \texttt{NO-OP}; \\ \max\left\{S^{\mathrm{m}}_t + 1, N-1\right\}, & \textbf{if } S^{\mathrm{m}}_t = A^{\mathrm{m}}_t; \\ 0, & \text{Otherwise.} \end{cases}$$

D.2 When There Are Never-Observable Rewards

In this section, we evaluate how Monitored MBIE-EB performs in unsolvable Mon-MDPs. The importance of this section is to verify the fact that pessimism is effectively useful in unsolvable Mon-MDPs. Mon-MDPs designed by Parisi et al. [39] do not have non-ergodic monitors, which would have given rise to unsolvable Mon-MDPs. Hence, we introduce **Bottleneck** to investigate the performance of Monitored MBIE-EB compared to Directed-E^2. As noted in Footnote 1, Directed-

E^2's performance in these settings depends critically on its reward model initialization. In order to see minimax-optimal performance from that algorithm, we need to initialize it pessimistically. Using the recommended random initialization saw essentially no learning in these domains. The algorithm would believe the never-observable rewards were at their initialized value, and so seek them out rather than treat their value pessimistically.

In Bottleneck the underlying reward of cells marked by $\perp$ is the same as being a cactus (-10). In these experiments we use Full-Random monitor since it is more stochastic than Semi-Random to increase the challenge. Results are shown in Figure D.3. The location of the button is chosen to test deep exploration capabilities of the agents meaning performing a long sequence of costly actions in order to obtain the highest return. As a result the range of returns that the agent obtains with Button monitor is naturally lower than the rest of the Mon-MDPs.

One of the weaknesses of Directed-E^2 is its explicit dependence on the state space's size. Because Directed-E^2 tries to visit every joint state-action pair infinitely often without paying attention to their importance on maximizing the return, as the state space gets larger, the performance of Directed-E^2 deteriorates. To highlight this issue we use N-Experts monitor as an extension of Ask monitor; Ask is a special case when N is one. We see in Figure D.3 Directed-E^2's performance is hindered considerably when the agent faces N-Experts compared to Ask, while Monitored MBIE-EB suffers to a lesser degree.

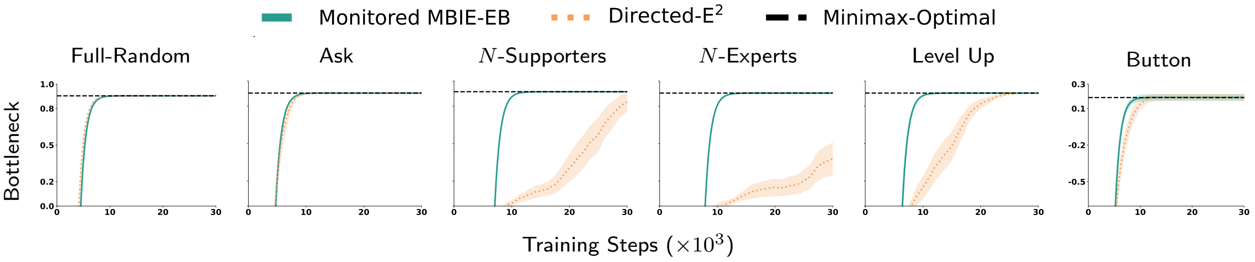


Figure D.3: Monitored MBIE-EB outperforms Directed-E^2 on Bottleneck with a non-ergodic monitor. Even though Directed-E^2's reward model was initialized pessimistically to achieve asymptotic minimax-optimality, its dependence on the state and action spaces' size makes it struggle more than Monitored MBIE-EB on N-Supporters, N-Experts and Level Up.

D.3 When There Are Stochastically-Observable Rewards

In this section, we confirm that Monitored MBIE-EB becomes pessimistic about environment rewards that are effectively never-observable. Hence, Monitored MBIE-EB should be robust against stochastic observability. In all of the previous experiments, had the agent done the action that would have revealed the environment reward, such as asking in Ask monitor, by paying the cost it would have seen the reward with 100% certainty. But even if the probability is not 100% and yet bigger than 0, then upon *enough* attempts to observe the reward and paying the cost even if a portion of them are fruitless, it is possible to observe and learn the environment reward. In Figure D.4's experiments, in addition to have environment state-action pairs that their reward is permanently unobservable, we make the monitor stochastic for other pairs such that even if the agent pays the cost, it would only observe the reward with probability ρ. In Figure D.4, it can be seen that albeit the challenge of having stochastically and permanently unobservable rewards, Monitored MBIE-EB has not become prematurely pessimistic about the rewards that can be observed, even the probability goes down as low as 5%, and still outperforms Directed-E^2.

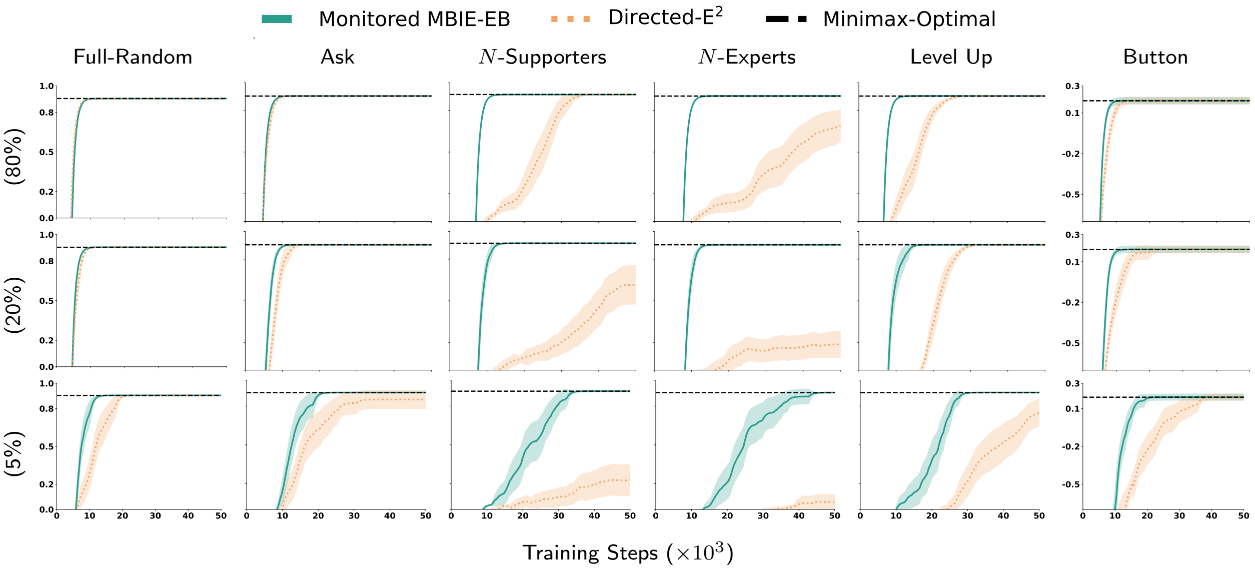


Figure D.4: Monitored MBIE-EB outperforms Directed-E^2 on Bottleneck even if environment rewards are stochastically observable. The plots also show the effect of ρ^{-1} in the Monitored MBIE-EB's sample complexity stated in Theorem 3.1. For a fixed environment and monitor, as the probability of observing the reward decreases, the more samples are required to find a minimax-optimal policy. The plots also indicates that although the sample complexity of Directed-E^2 has not been given theoretically, it must more severely depend on ρ^{-1} than Monitored MBIE-EB's.

D.4 When the Monitor is Known

In this section, we verify that knowing the monitor speeds up the Monitored MBIE-EB's learning. We want to empirically show that knowing the monitor's models is an advantage that Monitored MBIE-EB can benefit from. A trait that is not readily possible in a model-free algorithm such as Directed-E^2. We have shown the superior performance of Monitored MBIE-EB compared to Directed-E^2, now we investigate how much of the difficulty of learning in Mon-MDPs comes from the monitor being unknown. The unknown quantities of the monitor to the agent are $r^{\mathrm{m}}, p^{\mathrm{m}}$, and f^{m}, hence in the following experiments we make all of them known to the agent in advance. The only remaining unknown quantities are r^{e} and p^{e}. Hence, we replace Equation (3.5) with

$$\widetilde{Q}^{**}(s,a) = \mathbb{P}\left(\widehat{R}_{t+1} \neq \bot \mid S_t = s, A_t = a\right) + \gamma \sum_{s'} \bar{p}(s' \mid s,a) \max_{a'} \widetilde{Q}^{**}(s',a') + \beta \sqrt{\frac{g(N_v(s^{\mathrm{e}}))}{N_v(s^{\mathrm{e}},a^{\mathrm{e}})}},$$

where $N_v(s^{\mathrm{e}},a^{\mathrm{e}})$ counts the number of times $s^{\mathrm{e}},a^{\mathrm{e}}$ has been visited, $N_v(s^{\mathrm{e}}) = \sum_a N_v(s^{\mathrm{e}},a^{\mathrm{e}})$ and $g(x) = 1 + x\ln^2 x, x \geq 0$. The *intuition* behind the bonus $\beta\sqrt{\frac{g(N_v(s^{\mathrm{e}}))}{N_v(s^{\mathrm{e}},a^{\mathrm{e}})}}$ comes from the fact that $p = p^{\mathrm{e}} \otimes p^{\mathrm{m}}$ and we only need to account for the uncertainty stemming from knowing p^{e}. Since the monitor is known there is no need to use KL-UCB, as the agent already knows which environment rewards are observable (with what probability). Similarly, we replace Equation (3.2) with

$$Q^{**}_{\downarrow}(s,a) = \bar{r}^{\mathrm{e}}(s^{\mathrm{e}},a^{\mathrm{e}}) + r^{\mathrm{m}}(s^{\mathrm{m}},a^{\mathrm{m}}) + \gamma \sum_{s'} \bar{p}(s' \mid s,a) V^{**}_{\downarrow}(s') + \beta^{\mathrm{e}} \sqrt{\frac{g(N(s^{\mathrm{e}}))}{N(s^{\mathrm{e}},a^{\mathrm{e}})}} + \beta \sqrt{\frac{g(N_v(s^{\mathrm{e}}))}{N_v(s^{\mathrm{e}},a^{\mathrm{e}})}},$$

where the bonus $\beta^{\mathrm{e}}\sqrt{\frac{g(N(s^{\mathrm{e}}))}{N(s^{\mathrm{e}},a^{\mathrm{e}})}}$ is due to the environment reward model, and $\beta\sqrt{\frac{g(N_v(s^{\mathrm{e}}))}{N_v(s^{\mathrm{e}},a^{\mathrm{e}})}}$ accounts for the fact $\bar{p}^{\mathrm{e}}$ only gets more accurate by visiting insufficiently visited environment state-action pairs. Figure D.5 shows the prior knowledge of the monitor's quantities boosts the speed of Monitored MBIE-EB's learning and make it robust even in the low probability regimes.

D.5 Significance of Monitored MBIE-EB's Innovations

Throughout this thesis, we have constantly emphasizing on extending the idea of MBIE-EB to Mon-MDPs. In this section, we show how our innovations are crucial to be able to extend MBIE-EB to Mon-MDPs. We show that without all our proposed innovations, there exists at least one setting that the resulting algorithm fails.

D.5.1 Extending MBIE-EB to Mon-MDPs

MBIE-EB uses the initial action-values to assign optimistic values to state-action pairs that their counts are zero. This means that in Mon-MDPs for joint state-action pairs $(s,a) \equiv (s^{\mathrm{e}}, s^{\mathrm{m}}, a^{\mathrm{e}}, a^{\mathrm{m}})$ that any of $N(s^{\mathrm{e}},a^{\mathrm{e}}), N(s^{\mathrm{m}},a^{\mathrm{m}})$, or $N(s,a)$ is zero, $Q(s,a)$ would be shortcut to an optimistic value. This approach contrasts with the pessimism of Equation (3.1), when $N(s^{\mathrm{e}},a^{\mathrm{e}})$ is zero. Hence, we show that on the Bottleneck environment, when the reward of all $\bot$ cells are observable to the agent, MBIE-EB is effective. Because upon enough visitation resulting from the optimism, the underlying environment reward will finally be observed. The efficacy of MBIE-EB in this setting

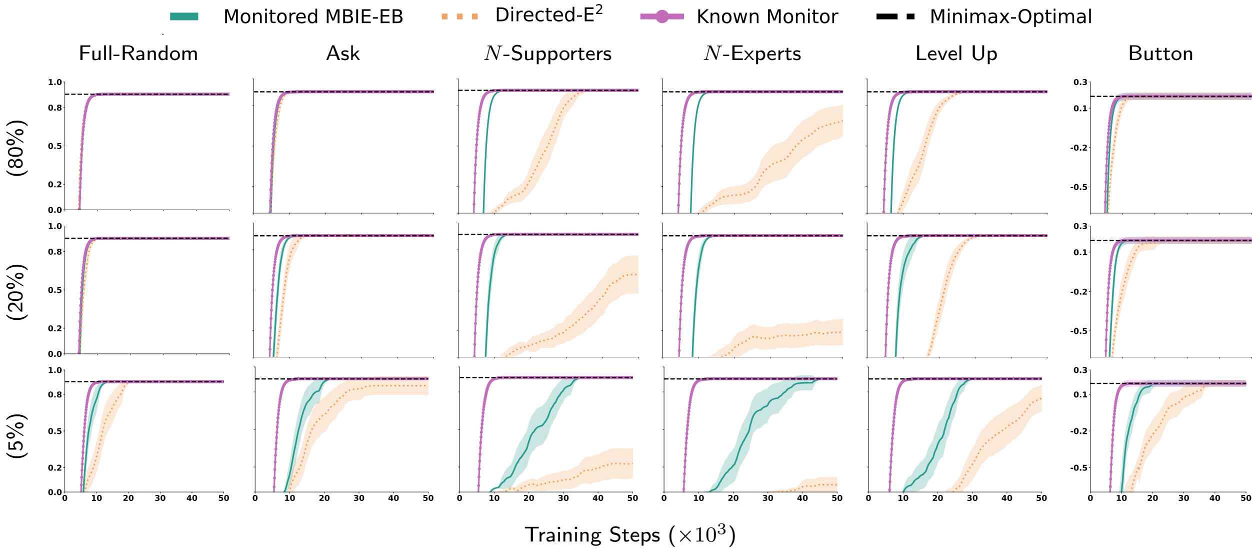
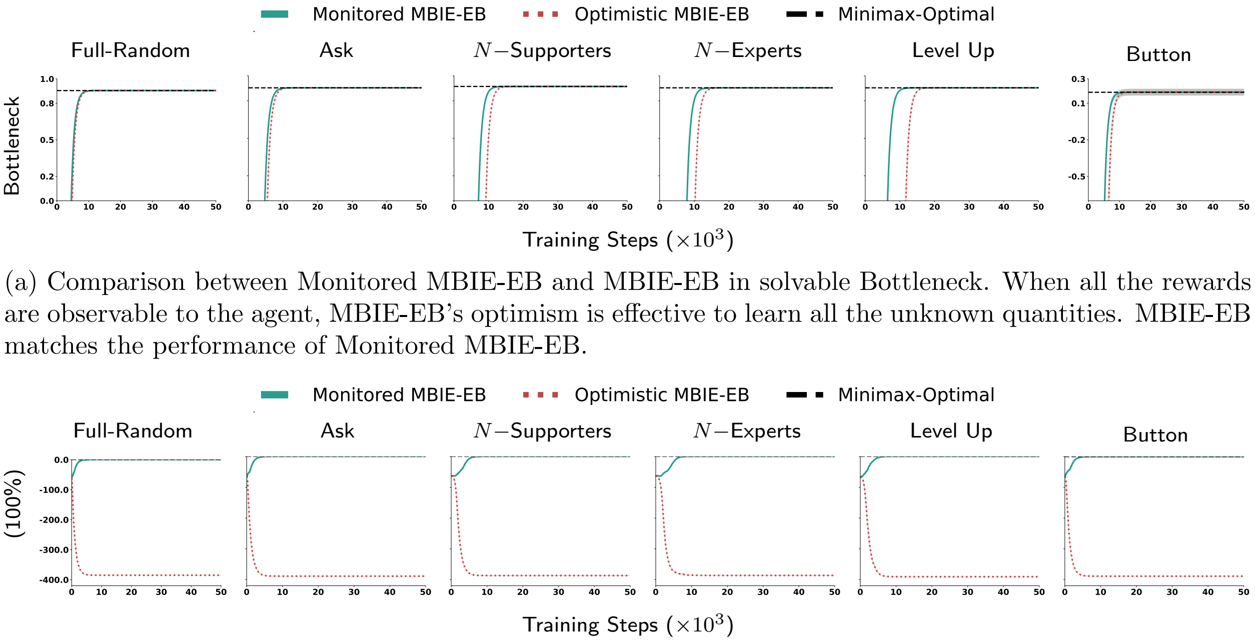


Figure D.5: Knowing the monitoring process considerably accelerates learning in Mon-MDPs. The similar learning speed in Ask and N-Experts show that the knowledge of the monitor make Monitored MBIE-EB robust against the size of the monitor spaces. Also, the similarity of learning speed for a fixed environment and monitor across experiments with high and low observability probability shows that the in-advance-given knowledge of the monitor help the agent focus its exploration on state-action pairs that their environment rewards is observable even if the probability is low.

is shown in Figure D.6a. However, the lack of necessary pessimism when the reward of all $\perp$ cells are unobservable make MBIE-EB ineffective. Because the optimism never washes out, hence the agent would visit state-action pairs with unobservable rewards for ever. The failure of MBIE in at least one unsolvable Mon-MDP is shown in the results of Figure D.6b.

D.5.2 Pessimistic MBIE-EB Without Observation Stage

In this section, we empirically highlight the importance of observation stage. In Appendix D.5.1 we showed that the excessive optimism of MBIE-EB hinders its performance in unsolvable Mon-MDPs. Now, we examine that without the observation stage, adding the pessimism with respect to the unobservable environment rewards is still prone to failure. We extend the MBIE-EB to Mon-MDPs and for all state-action pairs $(s, a) \equiv (s^{\mathrm{e}}, s^{\mathrm{m}}, a^{\mathrm{e}}, a^{\mathrm{m}})$, when $N(s^{\mathrm{e}}, a^{\mathrm{e}})$ is zero, we use pessimistic rewards. This approach is effective in Mon-MDPs with deterministic observability of the rewards ($\rho = 1$). This effectiveness is shown in Figure D.7a, where the results are obtained by running the pessimistic MBIE-EB on the solvable Bottleneck. Pessimistic MBIE-EB is also effective in unsolvable Mon-MDPs with deterministic observability, where only one visit to each state-action pair is enough to conclude whether the environment reward is observable or not. We verified this claim by running the pessimistic MBIE-EB on the unsolvable Bottleneck with 100%



(a) Comparison between Monitored MBIE-EB and MBIE-EB in solvable Bottleneck. When all the rewards are observable to the agent, MBIE-EB's optimism is effective to learn all the unknown quantities. MBIE-EB matches the performance of Monitored MBIE-EB.

(b) Comparison between Monitored MBIE-EB and MBIE-EB in unsolvable Bottleneck. If some environment rewards are unobservable to the agent, excessive MBIE-EB's optimism ineffective. While Monitored MBIE-EB is pessimistic with respect to unobservable rewards, MBIE-EB remains mistakenly optimistic about them. The ever-lasting optimism of MBIE-EB makes it visit $\perp$ cells for ever.

Figure D.6: Verifying the importance of pessimism instead of optimism in Mon-MDPs.

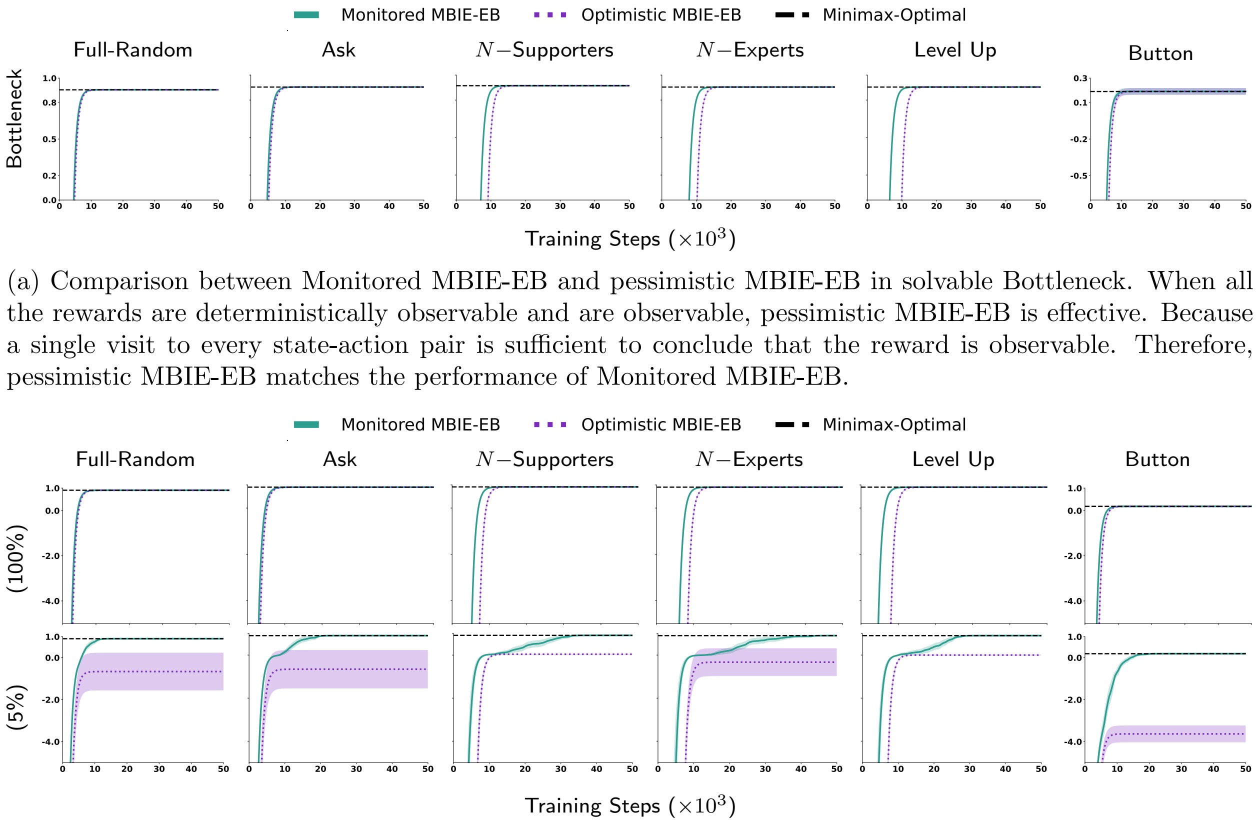
observability and the results are shown in Figure D.7b, (100%) row. However, if the observability is stochastic then premature pessimism hinders the pessimistic MBIE-EB's performance as it has become pessimistic with respect to state-action pairs that otherwise it could have observed their rewards eventually. This shortcoming of the pessimistic MBIE-EB compared to Monitored MBIE-EB that uses the observation stage is evident in Figure D.7b, (5%) row.

D.6 Considerations

In this section, we enumerate the fine details used in the experiments. These details help to make the thesis' implementation reproducible:

- In all experiments, $\rho = 1$ unless otherwise states. In experiments that include N-Supporters or N-Experts, $N = 4$ and the number of levels for Level-Up is 3.

- Hyperparameters of Directed-E^2 consist of: Q_0 the initial action-values, Ψ_0 the initial visitation-values, r_0 the initial values of the environment reward model, $\bar{\beta}$ goal-conditioned threshold specifying when a joint state-action pair should be visited through the use of visitation-values, α the learning rate to update each Q or Ψ incrementally and discount factor γ that is held fixed 0.99.



(a) Comparison between Monitored MBIE-EB and pessimistic MBIE-EB in solvable Bottleneck. When all the rewards are deterministically observable and are observable, pessimistic MBIE-EB is effective. Because a single visit to every state-action pair is sufficient to conclude that the reward is observable. Therefore, pessimistic MBIE-EB matches the performance of Monitored MBIE-EB.

(b) Comparison between Monitored MBIE-EB and pessimistic MBIE-EB in unsolvable Bottleneck. If the observability of the environment reward is deterministic, then pessimistic MBIE-EB is effective in finding the minimax-optimality. But, if the observability is stochastic, pessimistic MBIE-EB due to its premature pessimism with respect to state-actions that their reward can be observed upon enough exploration fails to find the minimax-optimal policy. On the other hand, due to exploring to observe the reward in the observation stage, Monitored MBIE-EB is robust against the stochasticity in the observability of the reward.

Figure D.7: Verifying the importance of the observation stage.

These values are directly reported from Parisi et al. [39].

- Monitored MBIE-EB's hyperparameters are set per environment and do not change across monitors. The same applies to Directed-E^2, but Parisi et al. [39] recommend to tune an ad-hoc learning rate once N-Supporters or N-Experts are used as monitors.

- KL-UCB does not have a closed form solution, we compute it using the Newton's method. The stopping condition for the Newton's method is chosen 50 iterations or the accuracy of at least 10^{-5} between successive iterative solutions, which one happens first.

- We ran all experiments on a SLURM-based cluster, using 32 Intel E5-2683 v4 Broadwell @ 2.1GHz CPUs. 30 runs took about an hour on a 32 core CPU. Runs were parallelized whenever possible.

D.7 Hyperparameters

In this section, we mention the hyperparameters used in the experiments in this thesis to make the interpretation of the results more complete. Let $\mathcal{U}$ denote the uniform distribution and $x \mapsto y$ denote the linear annealing of a quantity taking initially the value of x and ends with y.

Table D.1: Set of hyperparameters

(a) Hyperparameters of Monitored MBIE-EB across experiments.

Unknown monitor						
Experiment	Environment	Q_0	$\widetilde{Q}_0$	$\kappa^*(k)$	$\beta^{\text{KL-UCB}}$	$\beta^{\text{obs}}, \beta, \beta^{\text{m}}, \beta^{\text{e}}$
Figure 4.4	**Empty**	1	100	$\log_{1.005} k$	5×10^{-2}	5×10^{-4}
	Hazard	1	100	$\log_{1.005} k$	5×10^{-2}	5×10^{-4}
	One-Way	1	100	$\log_{1.005} k$	5×10^{-2}	5×10^{-4}
	River-Swim	30	100	$\log_{1.005} k$	5×10^{-2}	5×10^{-4}
Appendix D.2	**Bottleneck**	1	100	$\log_{1.005} k$	5×10^{-2}	5×10^{-4}
Appendix D.3	**Bottleneck**	1	100	$\log_{1.005} k$	5×10^{-2}	5×10^{-4}

Known monitor						
Experiment	Environment	Q_0	$\widetilde{Q}_0$	$\kappa^*(k)$	β^{e}	β
Appendix D.4	**Bottleneck**	1	100	$\log_{1.005} k$	5×10^{-4}	5×10^{-4}

(b) Hyperparameters of MBIE-EB across experiments.

Experiment	Environment	Q_0	$\beta, \beta^{\text{m}}, \beta^{\text{e}}$
Appendix D.5	**Bottleneck**	50	5×10^{-4}

(c) Hyperparameters of Directed-E^2 across experiments.

(Annealing for N-Supporters and N-Experts)						
Experiment	Environment	Q_0	Ψ_0	r_0	$\bar{\beta}$	α
Figure 4.4	**Empty**	-10	1	$\mathcal{U}[-0.1, 0.1]$	10^{-2}	$1(1 \mapsto 0.1)$
	One-Way	-10	1	$\mathcal{U}[-0.1, 0.1]$	10^{-2}	$1(1 \mapsto 0.1)$
	Hazard	-10	1	$\mathcal{U}[-0.1, 0.1]$	10^{-2}	$0.5(0.5 \mapsto 0.1)$
	River-Swim	-10	1	$\mathcal{U}[-0.1, 0.1]$	10^{-2}	$0.5 \mapsto 0.05$
Appendix D.3	**Bottleneck**	-10	1	-10	10^{-2}	$1(1 \mapsto 0.1)$
Appendix D.3	**Bottleneck**	-10	1	-10	10^{-2}	$1(1 \mapsto 0.1)$