Information-directed policy sampling for episodic Bayesian Markov decision processes

Victoria Diaz

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

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Archis Ghate, Chair

Chiwei Yan

Chaoyue Zhao

Program Authorized to Offer Degree:
Industrial and Systems Engineering

University of Washington

Abstract

Information-directed policy sampling for episodic Bayesian Markov decision processes

Victoria Diaz

Chair of the Supervisory Committee:
Professor Archis Ghate
Industrial & Systems Engineering

The research objective of this dissertation is to apply information theoretic methods to design provably efficient approximate solution algorithms for Markov decision processes (MDPs), partially observable MDPs (POMDPs), and hierarchical MDPs, under incomplete information. We consider these problems within an episodic Bayesian framework, where the decision-maker interacts with a stochastic system repeatedly over T episodes comprising of N stages each. The decision-maker only knows that the true parameters describing the stochastic system take values from a particular finite set. The decision-maker begins the first episode with a prior probabilistic belief about the true parameters of the system, and updates this belief at the end of each episode based on observed events. The decision-maker wishes to maximize expected total reward earned over all episodes under such incomplete information.

The challenge of balancing exploration versus exploitation is at the heart of this dissertation. The decision-maker should execute policies that provide information about the true parameters of the system (exploration), but should also exploit this acquired knowledge to implement policies that earn high rewards. Exact methods that attempt to balance this trade-off are computationally intractable due to the curse-of-dimensionality. Approximate solution methods are thus desired, but often are only available as heuristics with no or poor regret bounds.

To overcome these limitations, this dissertation proposes a framework whereby, in each

episode, the decision-maker executes a policy sampled from a probability mass function (pmf) that minimizes a so-called convex information ratio. The numerator of this information ratio equals the squared-regret incurred and the denominator equals the information gained about the true parameters of the system, by executing such a policy. Minimizing this ratio is thus a natural way to balance the exploration-exploitation trade-off. We call the resulting framework information-directed policy sampling (IDPS). This idea is motivated by the recent theoretical and computational success of a paradigm called information-directed sampling in balancing this trade-off in the special case of multi-armed bandit problems. However, the dependence of future states on current state-action pairs poses unique technical hurdles while generalizing this idea to Markovian systems. We tackle this challenge by introducing a new way to define the episodic regret and information gain using pmfs over the set of policies that are optimal under distinct system parameters, instead of the set of all policies.

We derive regret bounds that do not depend on the state-space, action-space, or observation-space cardinalities. Instead, our regret bounds scale elegantly with the number of episodes T, number of possible parameter values, number of stages N, and the entropy of prior belief. The proposed algorithms are compared computationally against a state-of-the-art approach called Posterior Sampling (PS) on three applications: queuing control, machine repair, and dynamic pricing.

The thesis is organized is as follows. The first chapter investigates MDPs where the decision-maker has incomplete information about the state transition probabilities and single-stage rewards. A regret bound for IDPS is derived, and numerical experiments show that IDPS outperforms PS on all three applications. The second chapter studies POMDPs where the decision-maker has incomplete information about the state transition probabilities and the observation probabilities. A regret bound for IDPS is derived. The third chapter relates to MDPs with a hierarchical incomplete information framework. The upper level of this hierarchy includes ambiguity about which structural model characterizes the true system-

dynamics, and the lower level corresponds to the ambiguity regarding the true parameters of these potential models. For instance, the decision-maker may not know whether the true demand model is Poisson or Binomial. Further, if the true model is Poisson, then the decision-maker may not know its mean. Three variations of IDPS are introduced, and a regret bound for one such variation is derived. Computational experiments consider a hierarchical variant of the dynamic pricing application.

Future research could focus on extending the framework and theoretical analyses in this dissertation to other settings such as indefinite-horizon MDPs, continuous-time MDPs, semi-Markov decision processes, and multi-player stochastic games.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

MARKOV DECISION PROCESSES WITH INCOMPLETE INFORMATION

1.1 Introduction

Markov decision processes (MDPs) with incomplete information are ubiquitous in applications. For example, in dynamic pricing, the decision-maker may not know how the market will respond to a product (see [16] for a survey). If the market is hot, then demand is modeled by one probability mass function (pmf) (say Poisson with a large mean), and otherwise by another pmf (say Poisson with a smaller mean). Since the demand pmf characterizes transition probabilities, the manager faces one of two possible MDPs. The manager must make pricing decisions to maximize expected total profit while simultaneously learning the market-type. Similar problems arise in queuing systems [1], mechanism design [19], and financial services [28]. As another example, a doctor may not know how patients in a clinical trial will respond to treatment [3, 4, 43]. If the treatment is effective, the patients' disease conditions will evolve with a higher likelihood of showing improvements, for the same dose, compared to a moderately effective or an ineffective treatment. The doctor thus faces one of three possible transition probability matrices. Further, evolving disease conditions impact the patient's quality of life (QOL). Several QOL metrics are available [27]. Assuming that either the Centers for Disease Control and Prevention's QOL metric or the World Health Organization's QOL metric is the true one, the doctor faces two possible reward matrices. The doctor must make treatment decisions to optimize a quantitative metric of the patient's health, while simultaneously learning treatment effectiveness and the correct QOL metric.

We focus on a Bayesian setup where the decision-maker sequentially encounters the same unknown finite-stage MDP multiple times. Each repetition in this sequence is called an

episode. The decision-maker begins with an initial belief as to which matrices represent the true transition probabilities and rewards. Based on this initial belief, the decision-maker chooses a policy, observes the initial state, and executes the chosen policy. The decision-maker observes the states and rewards induced by this policy and then updates its belief about the true MDP. This process repeats over multiple episodes. The decision-maker's aim is to choose policies to maximize the expected total reward over all episodes.

This episodic framework captures repeated interactions that occur in many applications between decision-makers and stochastic systems. In treatment planning, each episode may correspond to a distinct patient's treatment-course. The doctor begins with a prior belief about treatment effectiveness and the QOL metric, when the first patient arrives in a disease condition sampled from the population. The doctor uses a Bayesian approach to update this belief at the end of the treatment-course, based on the observed evolution of disease conditions and the corresponding QOL realizations. This updated belief is subsequently utilized while deciding the second patient's treatment-course, and the process continues similarly thereafter.

This framework has a perceived weakness. Namely, the decision-maker does not immediately alter the policy based on stage-by-stage observations. Instead, it is only updated at the end of each episode. This embodies the practice of committing to a course of action for the entire encounter of several stages, and altering the policy for the next encounter after observing events in the previous one. This is done when policy modifications based on a single state observation in one stage are not viable. For instance, in medicine, regulatory protocols prohibit mid-course policy modifications. Similarly, the pricing policy for a shipment of goods held in inventory is not altered after every unit is sold, but is instead modified only after a new shipment arrives. Thus, the episodic Bayesian setup has received attention in the literature, where it is also noted to have desirable theoretical and computational properties [20, 47, 62].

It is possible to write Bellman's equations for an episodic Bayesian problem. This is achieved by appending the initial state pmf with an "information state" that equals the be-

lief pmf at the beginning of each episode. The resulting controlled stochastic process can be viewed as a Bayes Adaptive MDP [17]. Exact solution of these Bellman's equations, however, is notoriously difficult. This is partly due to both the initial state pmf and the information state being "continuous" with values belonging to high-dimensional probability simplexes. In addition, the number of available policies in an episode can be large even when the state- and action-spaces are small. Thus, this formulation suffers from the curse-of-dimensionality and approximate solution techniques are desired. Unfortunately, the literature on computationally tractable approximation methods with provably efficient regret bounds is sparse, as demonstrated next.

1.2 Literature review

Literature on robust, distributionally robust, and percentile optimization in MDPs [29, 15] addresses the decision-maker's incomplete knowledge via *offline* methods with varying degrees of conservatism. Those works are not related to the present work, and hence are not reviewed here.

The fundamental trade-off in the above *online* learning problem is exploration-exploitation [49]. The decision-maker should explore by implementing policies that provide information about which MDP is the true one, perhaps sacrificing some amount of reward. The decision-maker should also exploit this information to execute policies that earn high rewards. This trade-off poses theoretical and computational challenges owing to the complex information-structure prevalent in applications.

A multi-armed bandit problem captures this trade-off in its simplest form. There is about 50 years of work (84,000 articles in Google Scholar) attempting to design algorithms and derive their regret bounds for this special case [21, 22, 40, 67]. The two most relevant methods for this work are Posterior Sampling (PS) [53, 54] and Information-directed Sampling (IDS) [55]. PS for (non-episodic) multi-armed bandit problems over N stages is motivated by Thompson Sampling [64]. The decision-maker maintains a belief pmf about which arm is optimal, and pulls an arm sampled from this belief at each stage n. The belief is updated

after each stage, based on the observed outcome in that stage. This simple approach performs well empirically [12] and enjoys efficient regret bounds [2, 32, 53, 54]. Nevertheless, Russo and Van Roy [55] constructed examples where PS performs poorly by failing to choose actions that provide information about the optimal arm. This motivated them to develop IDS. IDS is centered on the concept of an *information ratio*. The numerator of this ratio equals the square of the immediate regret of the reward earned upon pulling an arm, and the denominator equals the information gained about the optimal arm. This denominator is termed "mutual information" between the belief pmf and the outcome observed after pulling an arm [24]. The numerator relates to exploitation, whereas the denominator corresponds to exploration. The decision-maker prefers small values of the numerator and large values of the denominator. The decision-maker thus samples arms in stage n according to a pmf that minimizes the information ratio. This problem is convex, and the N-stage regret of IDS scales nicely with key parameters of the multi-armed bandit problem [55]. Their numerical experiments demonstrated that IDS can outperform PS and other state-of-the-art methods such as Knowledge Gradient [18] and Upper Confidence Bounds [6] for multi-armed bandits.

In MDPs, which are more general than multi-armed bandits, the exploration-exploitation trade-off has been studied in non-Bayesian Reinforcement Learning (RL) [63]. That literature includes classic approaches such as Q-learning [66] as well as many modern variants. Several algorithms therein are heuristic in that their regret bounds are either not available or believed to be of poor quality. Nevertheless, some of them have been successfully applied to massive-scale problems (see [57] for a flagship example and survey). Algorithms with provably efficient regret bounds ([9, 30, 31, 33, 65] and references therein) typically induce exploration through an approach called "optimism in the face of uncertainty" that dates back to early work on multi- armed bandits [40]. An exploration bonus is added to the reward values of state-action pairs with poorly understood transition dynamics. This incentivizes the decision-maker to execute policies that visit these state-action pairs. As the decision- maker progressively learns more about these state-action pairs, the effect of the exploration bonus is gradually reduced. Exact implementation of these algorithms, however, is often computationally intractable; the

literature typically includes simulations on toy examples.

Several heuristics are also available for Bayesian RL (for example, [14, 25, 59]). These methods can sometimes settle for higher short term rewards by favoring more exploitation. Kolter and Ng [36] proposed an optimistic algorithm that added an exploration bonus to infrequently visited state-action pairs. This enabled them to utilize analytical techniques from the optimistic non-Bayesian RL methods to establish theoretical complexity bounds. Gopalan and Mannor [23] noted that such optimistic RL methods are not effective when the transition probabilities are governed by lower-dimensional underlying parameters that induce complex dependencies in system-dynamics across several state-action pairs. As described in Section 1.1, this is ubiquitous in applications because system-dynamics are characterized by parameters such as customer arrival/service/demand rate. Strens [62] proposed an alternative PS heuristic for episodic Bayesian RL. At the beginning of each episode $t = 1 : T$, the decision-maker samples a single MDP from the belief pmf b_t, and executes a policy that is optimal for that MDP. Osband et al. [45] derived an efficient regret bound for PS thirteen years after Strens originally envisioned that algorithm, and provided computational results where it drastically outperformed state-of-the-art optimistic algorithms. Osband and Van Roy [46] made rigorous the precise sense in which PS is better than optimistic algorithms. Since then, PS has remained the cornerstone of episodic Bayesian RL.

Unfortunately, PS suffers from the same limitation in MDPs as it did in bandits — it chooses policies without any regard to the amount of information they provide about the true MDP. This can be a disadvantage when system-dynamics across different state-action pairs are coupled. In fact, recall that this weakness of PS previously motivated the development of IDS for the special case of multi-armed bandits. Kumar [38] constructed example MDPs that highlighted this issue.

On the other hand, IDS in the MDP setting takes into account the amount of information policies can provide about the true MDP, potentially explaining why encouraging preliminary computational performance has been reported [38, 44, 51, 68]. However, theoretical regret analyses have proven challenging. Kumar and Ghate [39] and Kumar [38]

explored an extension of IDS to the *non-episodic Bayesian* MDP and POMDP settings but faced analytical difficulties induced by the dependence of future states on the current state-action pair — a feature fundamental to MDPs but not relevant in bandits. This hurdle in information-theoretic regret analyses for *non-episodic* and for *non-Bayesian* MDPs has been noted elsewhere [23, 44].

The challenge in the *episodic Bayesian* MDP setting has been in proposing algorithms which both solve a tractable information ratio minimization problem and allow for a theoretical regret analysis. Two course-projects at Stanford University that attempted to extend IDS to episodic Bayesian MDPs suffered from this dilemma [51, 68]. An exact version called for solving an intractable information ratio minimization problem wherein the decision variable corresponded to a pmf over *all* possible policies. This is a vector of dimension exponential in both N and the cardinality of the state-space $|S|$. A theoretical regret analysis was provided, but only when the transition probabilities were *known* and the rewards were uncertain. An approximate version that restricted search to a subset of policies resulted in a tractable information ratio minimization problem, but was not amenable to regret analysis. Lu [41] also considered a version of IDS for episodic Bayesian MDPs with an intractable minimization problem over all possible policies. A regret bound was established when the decision-maker has a Beta belief over Bernoulli rewards and a Dirichlet belief over the next system-state, after a simplifying assumption about the parameters of the Dirichlet belief was made.

As such, an extension of IDS to episodic Bayesian MDPs that does not call for solving an intractable minimization problem *and* also allows for a theoretical regret analysis, is missing. We address this gap in the literature using the problem formulation described in the following section. Our framework lies between two extremes present in the current literature. On one end of the spectrum lies the multi-model MDP frameworks that attempt to find a policy which maximizes a convex combination of expected total rewards from M different MDPs [60, 61, 11]. There is no dynamic learning. On the other end are the contextual MDP framework of [26] and the multi-task RL framework of [10]. There, the decision-maker does not know the possible transition matrices it could be facing and thus, in a sense, must learn

all of them by repeated trajectory sampling. Thus, our framework achieves a good balance between a sufficiently rich modeling approach, and computational and analytical tractability.

1.3 Problem formulation

An MDP with complete information over time-stages $n = 1 : N$ is described as follows [50]. A system is in state $s_n \in S$ at the beginning of stage n, where S is a finite set. After observing s_n, a decision-maker chooses an action $a_n \in A$, where A is a finite set. The system then transitions into a state $s_{n+1} \in S$ with probability $p(s_{n+1}|s_n, a_n)$ and the decision-maker earns a reward $r(s_{n+1}|s_n, a_n)$. For simplicity of notation, we assume that the terminal reward is 0 regardless of the state at the end of the Nth stage. A policy $\pi = (\pi_1, \cdots, \pi_N)$ is an ordered tuple of mappings such that $\pi_n(s_n) \in A$ is the action prescribed in state $s_n \in S$ in stage n. The finite set of all such policies is denoted by Π. The initial state s_1 is presented to the decision-maker according to a pmf ρ_1 over S. The decision-maker wishes to choose a policy in Π that maximizes the expected total reward earned. Such an optimal policy can be found via backward recursive solution of Bellman's equations [7].

Transition probability values are stored in matrices $P(a)$ of size $|S| \times |S|$, for $a \in A$. The entry in the ith row and jth column of $P(a)$ equals $p(j|i, a)$, for $i, j \in S$. The 3-dimensional matrix of size $|S| \times |S| \times |A|$ formed by stacking matrices $P(a)$, for $a \in A$, is denoted by P. Reward matrices R are constructed similarly.

We consider the problem where the decision-maker has incomplete information about the transition probability matrix P and the reward matrix R. Specifically, the decision-maker only knows that the true transition probability matrix takes one of M possible values $\{P^1, \cdots, P^M\}$. Similarly, the true reward matrix takes one of L possible values $\{R^1, \cdots, R^L\}$. As such, the true MDP is one among ML possible options. The decision-maker needs to learn which of these ML MDPs is the true one, while simultaneously choosing actions that earn high rewards in an *online* fashion.

As a classic example of how this problem could arise, suppose that stochastic demand is modeled using a Poisson pmf in an inventory control problem. Thus, the transition proba-

bility matrices are completely characterized by the mean λ of this pmf. Consequently, the decision-maker's uncertainty about λ is directly translated into distinct known options for the transition probability matrices. Specifically, low, medium, and high values of λ translate to 3 known transition probability matrices, that is, $M = 3$. Even if the uncertainty in λ were a continuous interval, it can be arbitrarily well-approximated via an M-point discretization.

In episodic Bayesian framework, the decision-maker repeatedly and sequentially encounters the same but unknown N-stage MDP T times in an online fashion. The tth encounter is called the tth episode, for $t = 1 : T$. The state at the beginning of each episode is reset, independently of everything else, according to the pmf ρ_1. At the beginning of the first episode, the decision-maker believes that $b_1(m, \ell)$ is the probability that the MDP with transition probability matrix P^m and reward matrix R^ℓ is the true one. We refer to this MDP as the (m, ℓ)th MDP. Based on this initial belief, the decision-maker commits to a policy $\pi^1 \stackrel{\text{def}}{=} (\pi_1^1, \cdots, \pi_N^1) \in \Pi$ for the first episode. A state sampled according to ρ_1 is then revealed to the decision-maker. The decision-maker then executes policy π^1 until the end of the first episode. The decision-maker then employs Bayes' theorem to update its belief about the true MDP it faces, based on the states $s^1 = (s_1^1, \cdots, s_{N+1}^1)$ and the rewards $r^1 = (r_1^1, \cdots, r_{N+1}^1)$ that were observed in the first episode induced by π^1. This updated belief is denoted b_2. Based on b_2, the decision-maker commits to a policy $\pi^2 \in \Pi$ for the second episode; the process repeats until the end of episode T. The decision-maker wants to maximize expected total reward over T episodes.

We apply information-theoretic methods to design a provably efficient algorithm for approximate solution of such problems and derive asymptotically optimal regret bounds that scale nicely with problem-parameters in Section 1.4. This methodology is validated in Section 1.5 via numerical simulations against PS on three applications: machine repair, queuing control, and dynamic pricing. These applications possess a complex information structure that is known to render traditional sampling methods ineffective [23].

1.4 Information-Directed Policy Sampling

Motivated by the observations made in Section 1.2, Information-Directed Policy Sampling (IDPS) combines the benefits of IDS and PS. At the beginning of episode t, a policy is sampled from the set $\{\pi^{*11}, \cdots, \pi^{*ML}\}$ of policies optimal to the ML potential MDPs as in PS; here, policy $\pi^{*m\ell}$ is optimal to the (m, ℓ)th MDP. But, instead of sampling this policy from the ML-dimensional posterior belief b_t over those policies as in PS, it will be sampled from an ML-dimensional pmf that minimizes the information ratio as in IDS. Specifically, let $\mathcal{P}(ML)$ denote the probability simplex in $\mathbb{R}^{ML}$. Then, the information ratio for any pmf $f \in \mathcal{P}(ML)$ is defined as the square of the expected regret suffered in episode t by a policy sampled according to f, divided by the expected information gained in episode t by such a policy. Algorithm 1 conveys this idea schematically, using minimal other notation.

Algorithm 1 IDPS for episodic Bayesian MDPs with incomplete information

1: Start with initial belief pmf $b_1 \in \mathcal{P}(ML)$.
2: **for** episodes $t = 1 : T$ **do**
3: Let f_t^* be an optimal solution of the information ratio minimization problem

$$\min_{f \in \mathcal{P}(ML)} \ \phi_t(f) \stackrel{\text{def}}{=} \frac{(\text{Expected regret of } f \text{ in episode } t \text{ based on belief pmf } b_t)^2}{\text{Expected info. gain of } f \text{ in episode } t \text{ based on belief pmf } b_t}. \tag{1.1}$$

4: Observe initial state s_1^t sampled from $\rho_1 \in \mathcal{P}(|S|)$.
5: Execute π^t sampled from $\{\pi^{*11}, \cdots, \pi^{*ML}\}$ according to f_t^*; observe states $s_2^t, \cdots, s_{N+1}^t$ and rewards $r_1^t, \cdots, r_N^t$.
6: Update belief pmf as

$$b_{t+1}(m, \ell) \propto b_t(m, \ell)\rho_1(s_1^t)p^m(s_2^t|s_1^t, \pi_1^t(s_1^t)) \cdots p^m(s_{N+1}^t|s_N^t, \pi_N^t(s_N^t)) \cdots \\ \cdots \prod_{n=1}^{N} \mathbb{1}(r^\ell(s_{n+1}^t|s_n^t, \pi_n^t(s_n^t)) = r_n^t). \tag{1.2}$$

7: **end for**

Since the state at the beginning of each episode is reset per the initial distribution ρ_1 over S, independently of everything else, successive episodes are coupled only via the posterior

beliefs $b_2, \cdots, b_T$. This renders the episodic Bayesian MDP structurally similar to multi-armed bandit problems in that an episode can be viewed as a single stage of a multi-armed bandit problem. This correspondence leads to intuitive definitions of episodic regret and information gain that deliver efficient bounds. Importantly, the extent of the decision-maker's uncertainty about the true MDP in the episodic Bayesian framework is characterized by ML, just as in multi-armed bandit problems it is characterized by the number of arms. As such, ML is an exogenous quantity that equals the number of possible values of parameters that characterize the MDPs.

We make the following assumption.

Assumption 1.4.1. *For each fixed $\ell \in \{1, \ldots, L\}$, the difference between any two entries of the reward matrix R^ℓ is at most* 1. *Mathematically,*

$$\max_{j \in S, i \in S, a \in A} r^\ell(j|i,a) - \min_{j \in S, i \in S, a \in A} r^\ell(j|i,a) \leq 1, \text{ for each } \ell \in \{1, \ldots, L\}. \tag{1.3}$$

This assumption is without loss of generality as the single-stage rewards can always be normalized. Moreover, it implies that the difference between the total rewards earned in one episode by any two policies in MDP (m, ℓ) is at most N. This holds because, for any two policies $\mu, \nu \in \Pi$, we have,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mu,m,\rho_1}\left[\sum_{n=1}^{N} r^\ell(\mathbf{s}_{n+1}|\mathbf{s}_n, \mu_n(\mathbf{s}_n))\right] - \mathop{\mathbb{E}}_{\nu,m,\rho_1}\left[\sum_{n=1}^{N} r^\ell(\mathbf{s}_{n+1}|\mathbf{s}_n, \nu_n(\mathbf{s}_n))\right] \\
&\leq \mathop{\mathbb{E}}_{\mu,m,\rho_1}\left[\sum_{n=1}^{N} \left(\max_{j \in S, i \in S, a \in A} r^\ell(j|i,a)\right)\right] - \mathop{\mathbb{E}}_{\nu,m,\rho_1}\left[\sum_{n=1}^{N} \left(\min_{j \in S, i \in S, a \in A} r^\ell(j|i,a)\right)\right] \\
&= N\left(\max_{j \in S, i \in S, a \in A} r^\ell(j|i,a) - \min_{j \in S, i \in S, a \in A} r^\ell(j|i,a)\right) \leq N.
\end{aligned} \tag{1.4}$$

Here, the subscript μ, m, ρ_1 emphasizes that the expectation is taken with respect to the uncertainty in the stochastic state trajectory $\mathbf{s}_1, \ldots, \mathbf{s}_{N+1}$ induced by policy μ, transition probability matrix P^m, and initial state pmf ρ_1. Similarly for the subscript ν, m, ρ_1. The

final inequality above follows from (1.3).

Under Assumption 1.4.1, we show the powerful conclusion that regret is at most $N\sqrt{TML\mathcal{E}(b_1)} \leq N\sqrt{TML\log(ML)}$ *regardless* of $|S|$ and $|A|$. Here, $\mathcal{E}(b_1)$ is the Shannon entropy of the decision-maker's initial joint belief about the true transition probability and reward matrices. In particular, this bound elegantly captures the dependence of regret on prior belief — a hallmark of the Bayesian viewpoint — just as the regret analysis of IDS did for multi-armed bandit problems. This bound also implies that the regret-per-episode asymptotically vanishes at the rate $1/\sqrt{T}$. Similar bounds hold for the special case where either the transition probability matrix or reward matrix is known. Again, these bounds will only depend on a quantity determining the number of possible MDPs, the Shannon entropy of the initial belief, and the difference in total rewards between any two policies. To make these ideas precise, we begin by defining the episodic regret and episodic information gain that appear in the numerator and denominator of (1.1).

1.4.1 Episodic regret and information gain

Let V^{*ij} denote the expected total reward earned by the decision-maker over stages $n = 1 : N+1$ in MDP (i, j) upon executing policy π^{*ij} that is optimal in MDP (i, j), given that the initial state is sampled according to pmf ρ_1. Similarly, $V^{ijm\ell}$ is the expected total reward earned by the decision-maker over stages $n = 1 : N+1$ in MDP (m, ℓ) upon executing policy π^{*ij} that is optimal for MDP (i, j), given that the initial state is sampled according to pmf ρ_1. Note, therefore, that $V^{*m\ell} = V^{m\ell m\ell}$. Applying the law of iterated expectation [52], the expected regret in the numerator of optimization problem (1.1) equals

$$\sum_{i=1}^{M}\sum_{j=1}^{L} f(i,j)\left(\text{Expected regret in episode } t \text{ of policy } \pi^{*ij}\right)$$
$$= \sum_{i=1}^{M}\sum_{j=1}^{L} f(i,j)$$

$$\times \left(\sum_{m=1}^{M} \sum_{\ell=1}^{L} b_t(m,\ell) \left[\text{Regret in episode } t \text{ of policy } \pi^{*ij} \middle| \text{MDP } (m,\ell) \text{ is the true MDP} \right] \right)$$
$$= \sum_{i=1}^{M} \sum_{j=1}^{L} f(i,j) \left(\sum_{m=1}^{M} \sum_{\ell=1}^{L} b_t(m,\ell) \left[V^{*m\ell} - V^{ijm\ell} \right] \right) = \sum_{i=1}^{M} \sum_{j=1}^{L} f(i,j) \Delta_t(i,j) = f \bullet \Delta_t.$$

Here, $\Delta_t(i,j) \stackrel{\text{def}}{=} \sum_{m=1}^{M} \sum_{\ell=1}^{L} b_t(m,\ell) \left[V^{*m\ell} - V^{ijm\ell} \right]$ is called the regret of policy π^{*ij} in episode t. Moreover, $\bullet$ above denotes the dot product between ML-dimensional vectors f and Δ_t whose (i,j)th components are $f(i,j)$ and $\Delta_t(i,j)$, respectively. Quantities $V^{*m\ell}$ and $V^{ijm\ell}$ are known to the decision-maker, as they can be computed offline *a priori* via the usual backward recursion of dynamic programming [7]. As such, in episode t, the numerator on the righthand side in (1.1) is the square of a linear function of pmf f.

Let $g_t(i,j)$ denote the mutual information in episode t between the belief pmf b_t and the state-reward pairs (s^t, r^t) induced by π^{*ij}. The definition of mutual information between two discrete random variables [24] yields

$$g_t(i,j) = \left(\sum_{m=1}^{M} \sum_{\ell=1}^{L} b_t(m,\ell) \sum_{s^t, r^t} \mathbb{P}(s^t, r^t | \pi^{*ij},\ \text{MDP } (m,\ell) \text{ is the true MDP}) \right.$$
$$\left. \log \left(\frac{\mathbb{P}\left(s^t, r^t | \pi^{*ij},\ \text{MDP } (m,\ell) \text{ is the true MDP}\right)}{\mathbb{P}\left(s^t, r^t | \pi^{*ij}\right)} \right) \right).$$

The denominator of (1.1) then equals $f \bullet g_t$ where g_t is the ML-dimensional vector whose (i,j)th component is $g_t(i,j)$. Thus, the denominator of (1.1) is a linear function of pmf f. The probability $\mathbb{P}(s^t, r^t | \pi^{*ij},\ \text{MDP } (m,\ell) \text{ is the true MDP})$ in the definition of $g_t(i,j)$ can be computed offline as

$$\begin{aligned}
&\mathbb{P}(s^t, r^t | \pi^{*ij},\ \text{MDP } (m,\ell) \text{ is the true MDP}) \\
&= \mathbb{P}(s^t | \pi^{*ij},\ \text{MDP } (m,\ell) \text{ is the true MDP})\, \mathbb{P}(r^t | s^t, \pi^{*ij},\ \text{MDP } (m,\ell) \text{ is the true MDP}) \\
&= \mathbb{P}(s^t | \pi^{*ij}, m)\, \mathbb{P}(r^t | s^t, \pi^{*ij}, \ell), \qquad (1.5)
\end{aligned}$$

where the conditional on "m" stands for the event "P^m is the true transition probability matrix", and the conditional on "ℓ" stands for the event "R^ℓ is the true reward matrix". The two requisite probabilities on the RHS of (1.5) are derived later in formulas (1.13) and (1.15).

Applying known results about operations that preserve convexity [8], the objective function in (1.1) can be shown to be convex. As such, it is a convex optimization problem in $\mathbb{R}^{ML}$.

Both the numerator and denominator of (1.1) incorporate past data via the (posterior) belief pmf b_t. The posterior belief pmf b_{t+1} can be derived after implementing policy π^t, observing state trajectory s^t, and observing reward trajectory r^t using the (prior) belief b_t as

$$b_{t+1}(m,\ell) = b_t(m,\ell|\pi^t, s^t, r^t) \tag{1.6}$$

$$= \frac{b_t(m,\ell)\,\mathbb{P}^{\text{IDPS}}(\pi^t, s^t, r^t|m,\ell)}{\mathbb{P}^{\text{IDPS}}(\pi^t, s^t, r^t)} \tag{1.7}$$

$$= \frac{b_t(m,\ell)\,\mathbb{P}(r^t|\pi^t, s^t, m,\ell)\,\mathbb{P}(s^t|\pi^t, m,\ell)\,\mathbb{P}^{\text{IDPS}}(\pi^t|m,\ell)}{\mathbb{P}(s^t, r^t|\pi^t)\,\mathbb{P}^{\text{IDPS}}(\pi^t)} \tag{1.8}$$

$$= \frac{b_t(m,\ell)\,\mathbb{P}(r^t|\pi^t, s^t, \ell)\,\mathbb{P}(s^t|\pi^t, m)\,\mathbb{P}^{\text{IDPS}}(\pi^t)}{\mathbb{P}(s^t, r^t|\pi^t)\,\mathbb{P}^{\text{IDPS}}(\pi^t)} \tag{1.9}$$

$$= \frac{b_t(m,\ell)\,\mathbb{P}(r^t|\pi^t, s^t, \ell)\,\mathbb{P}(s^t|\pi^t, m)}{\mathbb{P}(s^t, r^t|\pi^t)} \tag{1.10}$$

$$\propto b_t(m,\ell)\,\mathbb{P}(r^t|\pi^t, s^t, \ell)\,\mathbb{P}(s^t|\pi^t, m) \tag{1.11}$$

$$\propto b_t(m,\ell)\rho_1(s_1^t)p^m(s_2^t|s_1^t, \pi_1^t(s_1^t))\cdots p^m(s_{N+1}^t|s_N^t, \pi_N^t(s_N^t))\prod_{n=1}^{N}\mathbb{1}(r^\ell(s_{n+1}^t|s_n^t, \pi_n^t(s_n^t)) = r_n^t). \tag{1.12}$$

The superscript IDPS in (1.7-1.9) emphasizes that the probability depends on IDPS. Specifically, it is the probability that policy π^t is chosen by the IDPS algorithm for episode t. The equality in (1.6) follows from the definition of $b_{t+1}(m,\ell)$, (1.7) uses Bayes' theorem, and (1.8) factors the rightmost probability in the numerator and denominator. The equality in

(1.9) holds by eliminating irrelevant quantities in several of the conditional probabilities. In particular, the probability of a reward trajectory only depends on the MDP's reward matrix, the state trajectory, and the actions chosen, implying $\mathbb{P}(r^t|\pi^t, s^t, m, \ell) = \mathbb{P}(r^t|\pi^t, s^t, \ell)$; the probability of a state trajectory only depends on the actions chosen and the underlying transition probability matrix, implying that $\mathbb{P}(s^t|\pi^t, m, \ell) = \mathbb{P}(s^t|\pi^t, m)$; and the probability of a policy chosen by IDPS does not depend on the true transition or reward matrix as the algorithm is not privy to this information, implying $\mathbb{P}^{\text{IDPS}}(\pi^t|m, \ell) = \mathbb{P}^{\text{IDPS}}(\pi^t)$. Cancelling $\mathbb{P}^{\text{IDPS}}(\pi^t)$ in the numerator and denominator of (1.9) justifies (1.10). Moreover, (1.10) shows how updating the posterior belief pmf does *not* depend on the algorithm used, as all probabilities depending on IDPS have been eliminated. Meaning, the posterior belief does *not* depend on how the data was acquired. The proportionality sign in (1.11) is commonly used when determining the posterior belief pmf as the denominator is typically difficult to compute while normalizing a pmf is straightforward, and "$\mathbb{1}$" in the RHS of (1.12) is the indicator function. Finally, (1.12) holds true due to the following two observations.

First, the probability $\mathbb{P}(s^t|\pi^t, m)$ of observing state trajectory $s^t = (s_1^t, \cdots, s_{N+1}^t)$ due to implementing policy $\pi^t = (\pi_1^t, \cdots, \pi_N^t)$ when the true MDP has transition probability matrix P^m can be written using the initial state pmf and the Markov property as

$$\mathbb{P}(s^t|\pi^t, P^m \text{ is the true transition probability matrix}) \tag{1.13}$$

$$= \rho_1(s_1^t)p^m(s_2^t|s_1^t, \pi_1^t(s_1^t)) \cdots p^m(s_{N+1}^t|s_N^t, \pi_N^t(s_N^t)). \tag{1.14}$$

Second, we have

$$\mathbb{P}(r^t|s^t, \pi^t, R^\ell \text{ is the true reward matrix}) \propto \prod_{n=1}^{N} \mathbb{1}(r^\ell(s_{n+1}^t|s_n^t, \pi_n^t(s_n^t)) = r_n^t). \tag{1.15}$$

This follows from the fact that $\mathbb{P}(\boldsymbol{r}^t|s^t, \pi^t, r^\ell \text{ is the true reward matrix})$ is a discrete uniform pmf over the set of all reward trajectories $r^t = (r_1^t, \ldots, r_N^t)$ such that $r^\ell(s_{n+1}^t|s_n^t, \pi_n^t(s_n^t)) = r_n^t$ for all $1 \leq n \leq N$.

Finally, the observations in (1.13) and (1.15) justify the proportionality sign in (1.12), yielding the belief update formula in (1.2) of Algorithm 1.

These expressions for the expected regret and expected information gain of pmf f in episode t based on belief pmf b_t render the regret analysis of IDPS similar to that of IDS.

1.4.2 Regret analysis

We now define the total T-episode regret of IDPS in the episodic Bayesian setting. Let π^t denote the policy executed by IDPS in episode t, and $V^{\pi^t m\ell}$ denote the expected total reward earned by this policy over stages $n = 1 : N+1$ of episode t given that the initial state was sampled according to pmf ρ_1, if the (m,ℓ)th MDP is the true one. Thus, the expected total reward earned by the IDPS algorithm in T episodes equals $\overset{\text{IDPS}}{\mathbb{E}}\left[\sum_{t=1}^{T} V^{\boldsymbol{\pi}^t m\ell}\right]$. The randomness in π^t arises from two sources. First, π^t is sampled from pmf f_t^*. Second, f_t^* itself depends (through the posterior b_t) on the policies $\pi^1, \cdots, \pi^{t-1}$ implemented as well as the state trajectories $s^1, \cdots, s^{t-1}$ and reward trajectories $r^1, \cdots, r^{t-1}$ observed in episodes $1 : t-1$. Recall that $V^{*m\ell}$ is the optimal expected total reward earned by the (optimal) policy $\pi^{*m\ell}$ in stages $n = 1 : N+1$ of MDP (m,ℓ). Thus, if the decision-maker knew that MDP (m,ℓ) was the true MDP, it would execute policy $\pi^{*m\ell}$ in each episode and earn the optimal reward of $TV^{*m\ell}$. That is, the regret of IDPS would be $TV^{*m\ell} - \overset{\text{IDPS}}{\mathbb{E}}\left[\sum_{t=1}^{T} V^{\boldsymbol{\pi}^t m\ell}\right]$. Since $b_1(m,\ell)$ denotes the probability that MDP (m,ℓ) is the true one, the regret of IDPS is given by

$$\begin{aligned}\text{Regret}(\text{IDPS}, T) &\overset{\text{def}}{=} \sum_{m=1}^{M}\sum_{\ell=1}^{L} b_1(m,\ell)\left(TV^{*m\ell} - \overset{\text{IDPS}}{\mathbb{E}}\left[\sum_{t=1}^{T} V^{\boldsymbol{\pi}^t m\ell}\right]\right)\\ &= \mathbb{E}\left[TV^{*\boldsymbol{m\ell}} - \overset{\text{IDPS}}{\mathbb{E}}\left[\sum_{t=1}^{T} V^{\boldsymbol{\pi}^t \boldsymbol{m\ell}}\right]\right].\end{aligned}$$

Our regret analysis will use a slightly different notation than that in the statement of Algorithm 1. When the algorithm is implemented, it is convenient to maintain and update a posterior belief b_t at the end of each episode based on policy π^t, state trajectory s^t, and

reward trajectory r^t and then disregard π^t, s^t, and r^t throughout the rest of the algorithm's implementation. However, the regret analysis will rely on an expectation with respect to the history $h^t = (\pi^1, s^1, r^1, \cdots, \pi^{t-1}, s^{t-1}, r^{t-1})$, which is a tuple containing all policies implemented over episodes $1 : t-1$ and their resulting state and reward trajectories. Removing the dependence of the posterior belief b_t on the history h^t introduces ambiguities when writing expectations necessary for the regret analysis. Thus, all quantities dependent on a particular history will be expressed as such. Specifically, the posterior belief pmf due to observing history h^t will be denoted by the ML-dimensional probability vector $b(h^t)$ where the (m, ℓ)th element, $b(m, \ell|h^t)$, is the belief that MDP (m, ℓ) is the true MDP after observing history h^t. We simplify $b(h^1)$ to b, as there is no history at the start of the first episode. Moreover, the expected regret suffered, expected information gain, and information ratio of a policy sampled according to a pmf $f \in \mathcal{P}(ML)$ after observing history h^t will be denoted by $f \bullet \Delta(h^t)$, $f \bullet g(h^t)$, and $\phi(f|h^t)$, respectively, to emphasize the dependence of these quantities on history h^t. Here, the (m, ℓ)th coordinates of $\Delta(h^t)$ and $g(h^t)$, representing the regret and information gain of policy $\pi^{*m\ell}$ after incorporating history h^t into the posterior belief pmf, are denoted by $\Delta(m, \ell|h^t)$ and $g(m, \ell|h^t)$, while

$$f^*(h^t) \in \underset{f \in \mathcal{P}(ML)}{\operatorname{argmin}} \ \phi(f|h^t) \stackrel{\text{def}}{=} \frac{(f \bullet \Delta(h^t))^2}{f \bullet g(h^t)} \tag{1.16}$$

is a pmf minimizing the information ratio after observing history h^t. For simplicity, we let $\phi(f^*|h^t) \stackrel{\text{def}}{=} \phi(f^*(h^t)|h^t)$ be the information ratio of an optimal solution to the problem in (1.16).

Now, our regret analysis can follow an approach similar to that of multi-armed bandit problems at a high level. This approach was originally used to derive a regret bound for PS by Russo and Van Roy [54] and has recently been used elsewhere [34, 35, 42, 48] in the multi-armed bandit setting. Lu [41] used it in the episodic Bayesian MDP setting. For our setting, the regret is established in the following main theorem. This theorem has several dependencies which we illustrate in Figure 1.1.

Theorem 1.4.2. *The regret of IDPS is bounded above by $N\sqrt{TML\mathcal{E}(b)}$, where N is the number of time-stages, T is the number of episodes, M is the number of possible transition probability matrices, L is the number of possible reward matrices, and $\mathcal{E}(b)$ is the Shannon entropy of the initial belief pmf b. That is, Regret$(IDPS, T) \leq N\sqrt{TML\mathcal{E}(b)}$.*

Proof. We have

$$\text{Regret}(\text{IDPS}, T) = \sum_{t=1}^{T} \underset{\boldsymbol{h}^t}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^t) \bullet \Delta(\boldsymbol{h}^t)) = \sum_{t=1}^{T} \underset{\boldsymbol{h}^t}{\overset{\text{IDPS}}{\mathbb{E}}} \left(\sqrt{\phi(f^*|\boldsymbol{h}^t) \left[f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t) \right]} \right) \tag{1.17}$$

$$\leq \sum_{t=1}^{T} \sqrt{\underset{\boldsymbol{h}^t}{\overset{\text{IDPS}}{\mathbb{E}}} (\phi(f^*|\boldsymbol{h}^t))} \sqrt{\underset{\boldsymbol{h}^t}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t))} \tag{1.18}$$

$$\leq \sqrt{\sum_{t=1}^{T} \underset{\boldsymbol{h}^t}{\overset{\text{IDPS}}{\mathbb{E}}} (\phi(f^*|\boldsymbol{h}^t))} \sqrt{\sum_{t=1}^{T} \underset{\boldsymbol{h}^t}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t))} \tag{1.19}$$

$$\leq \sqrt{\sum_{t=1}^{T} \underset{\boldsymbol{h}^t}{\overset{\text{IDPS}}{\mathbb{E}}} (\phi(f^*|\boldsymbol{h}^t))} \sqrt{\mathcal{E}(b)} \tag{1.20}$$

$$\leq \sqrt{TN^2ML}\sqrt{\mathcal{E}(b)} = N\sqrt{TML\mathcal{E}(b)}. \tag{1.21}$$

The first equality showing the regret is additively separable across episodes in (1.17) is due to Proposition 1.4.3, and the second equality follows from the definition of the information ratio $\phi(f^*|\boldsymbol{h}^t)$. The inequality in (1.18) is an application of Hölder's inequality $\mathbb{E}(|\boldsymbol{XY}|) \leq \mathbb{E}(|\boldsymbol{X}|^p)^{1/p}\, \mathbb{E}(|\boldsymbol{Y}|^q)^{1/q}$ with $\boldsymbol{X} \equiv \sqrt{\phi(f^*|\boldsymbol{h}^t)}, \boldsymbol{Y} \equiv \sqrt{f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)}$, and $p = q = 1/2$. The inequality in (1.19) follows from applying the square root of the Cauchy-Schwarz inequality $\left(\sum_t x_t y_t \right)^2 \leq \sum_t x_t^2 \sum_t y_t^2$ with $x_t \equiv \sqrt{\phi(f^*|\boldsymbol{h}^t)}$ and $y_t \equiv \sqrt{f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)}$. The inequality in (1.20) bounding the cumulative expected information gain of the pmfs chosen by IDPS over episodes $t = 1 : T$ by $\mathcal{E}(b)$ is due to Proposition 1.4.5. Finally, the inequality in (1.21), which bounds the cumulative expected information ratio of the pmf $f^*(h^t)$ chosen by IDPS by N^2ML, is due to Proposition 1.4.8. □

Proposition 1.4.8 uses an approach similar to multi-armed bandit problems to establish

that the information ratio of PS is at most N^2ML and then notes that the information ratio of IDPS is no larger than that of PS to establish the bound. This last step is different from that of the approach in [41], which uses an UCB algorithm to bound the information ratio.

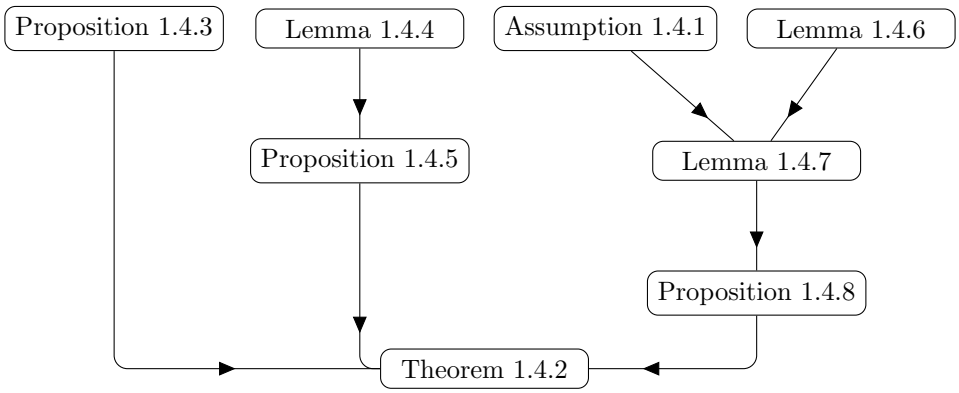


Figure 1.1: A schematic representation of the proof of Theorem 1.4.2.

The following proposition proves the equality in (1.17) of Theorem 1.4.2. It shows the regret can be written in an additively separable form across episodes.

Proposition 1.4.3. *The total regret of IDPS over T episodes can be expressed as the sum of its expected episodic regrets. That is, Regret$(IDPS, T) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{IDPS} (f^*(\boldsymbol{h}^t) \bullet \Delta(\boldsymbol{h}^t))$.*

Proof. The proof is by induction, and the base case is trivial. Probabilities of the form $\mathbb{P}^{\text{IDPS}}(\cdot|m,\ell)$ are conditional on the event "MDP (m,ℓ) is the true MDP".

Assume the proposition is true for $T = t$. We show it is true for $T = t+1$ by regrouping the regret of IDPS in terms of episodes, updating the posterior belief pmf, and simplifying the resulting quantities. That is,

$$\text{Regret}(\text{IDPS}, t+1) = \sum_{m=1}^{M}\sum_{\ell=1}^{L} b(m,\ell)\left[(t+1)V^{*m\ell} - \mathop{\mathbb{E}}^{\text{IDPS}}\left[\sum_{k=1}^{t+1} V^{\boldsymbol{\pi}^k m\ell}\right]\right] \tag{1.22}$$

$$= (t+1)\sum_{m=1}^{M}\sum_{\ell=1}^{L} b(m,\ell)V^{*m\ell} - \sum_{m=1}^{M}\sum_{\ell=1}^{L} b(m,\ell)\sum_{\pi^1,\cdots,\pi^{t+1}} \mathbb{P}^{\text{IDPS}}(\pi^1,\cdots,\pi^{t+1}|m,\ell)\sum_{k=1}^{t+1} V^{\pi^k m\ell} \tag{1.23}$$

$$= \left[t \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) V^{*m\ell} - \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) \sum_{\pi^1,\cdots,\pi^t} \sum_{\pi^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(\pi^1,\cdots,\pi^{t+1}|m,\ell) \sum_{k=1}^{t} V^{\pi^k m\ell} \right] \tag{1.24}$$

$$+ \left[\sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) V^{*m\ell} - \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) \sum_{\pi^1,\cdots,\pi^t} \sum_{\pi^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(\pi^1,\cdots,\pi^{t+1}|m,\ell) V^{\pi^{t+1} m\ell} \right]. \tag{1.25}$$

Lines (1.22-1.25) expand the definition of regret and reorganize terms in order to invoke the inductive hypothesis. The term $\mathbb{P}^{\text{IDPS}}(\pi^1,\cdots,\pi^{t+1}|m,\ell)$, first appearing in (1.23), is the probability of IDPS choosing policies $\pi^1,\cdots,\pi^{t+1}$ given that the (m,ℓ)th MDP is the true one. This joint probability depends on the true MDP through the posterior belief pmf. Line (1.24) can be rewritten as $\sum_{k=1}^{t} \overset{\text{IDPS}}{\underset{\boldsymbol{h}^k}{\mathbb{E}}}(f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k))$ using the inductive hypothesis as

$$\left[t \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) V^{*m\ell} - \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) \sum_{\pi^1,\cdots,\pi^t} \sum_{\pi^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(\pi^1,\cdots,\pi^{t+1}|m,\ell) \sum_{k=1}^{t} V^{\pi^k m\ell} \right] \tag{1.26}$$

$$= t \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) V^{*m\ell} \tag{1.27}$$

$$- \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) \sum_{\pi^1,\cdots,\pi^t} \overset{\text{IDPS}}{\mathbb{P}}(\pi^1,\cdots,\pi^t|m,\ell) \left[\sum_{\pi^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(\pi^{t+1}|\cancel{m,\ell},\pi^1,\cdots,\pi^t) \right]^{\nearrow 1} \sum_{l=1}^{t} V^{\pi^k m\ell} \tag{1.28}$$

$$= t \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) V^{*m\ell} - \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) \sum_{\pi^1,\cdots,\pi^t} \overset{\text{IDPS}}{\mathbb{P}}(\pi^1,\cdots,\pi^t|m,\ell) \sum_{l=1}^{t} V^{\pi^k m\ell} \tag{1.29}$$

$$= \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) \left[tV^{*m\ell} - \sum_{\pi^1,\cdots,\pi^t} \overset{\text{IDPS}}{\mathbb{P}}(\pi^1,\cdots,\pi^t|m,\ell) \sum_{l=1}^{t} V^{\pi^k m\ell} \right] \tag{1.30}$$

$$= \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) \left[tV^{*m\ell} - \overset{\text{IDPS}}{\mathbb{E}} \left[\sum_{k=1}^{t} V^{\boldsymbol{\pi}^k m\ell} \right] \right] \tag{1.31}$$

$$= \text{Regret}(\text{IDPS}, t) = \sum_{k=1}^{t} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k)). \tag{1.32}$$

Above, (1.28) factors probability $\mathbb{P}^{\text{IDPS}}(\pi^1, \cdots, \pi^{t+1}|m, \ell)$ in (1.26) and regroups quantities in order to remove a factor of 1 in (1.28), while (1.30) factors out the initial belief probabilities. The first equality in (1.32) uses the definition of regret and the second uses the inductive hypothesis. Substituting $\sum_{k=1}^{t} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k))$ for the quantity in (1.24) implies

$$\text{Regret}(\text{IDPS}, t+1) \tag{1.33}$$

$$= \sum_{k=1}^{t} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k)) + \left[\sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) V^{*m\ell} \right. \tag{1.34}$$

$$\left. - \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) \sum_{\pi^1,\cdots,\pi^t} \sum_{\pi^{t+1}} \mathbb{P}^{\text{IDPS}}(\pi^1, \cdots, \pi^{t+1}|m,\ell) V^{\pi^{t+1}m\ell} \right] \tag{1.35}$$

$$= \sum_{k=1}^{t} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k)) + \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) V^{*m\ell} \tag{1.36}$$

$$- \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) \sum_{\pi^1,\cdots,\pi^t} \sum_{\pi^{t+1}} \left[\sum_{s^1,r^1,\cdots,s^t,r^t} \mathbb{P}^{\text{IDPS}}(\pi^1, s^1, r^1, \cdots, \pi^t, s^t, r^t, \pi^{t+1}|m,\ell) \right] V^{\pi^{t+1}m\ell} \tag{1.37}$$

$$= \sum_{k=1}^{t} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k)) + \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) V^{*m\ell} \tag{1.38}$$

$$- \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) \sum_{\substack{\pi^1,s^1,r^1,\cdots,\\ \pi^t,s^t,r^t}} \sum_{\pi^{t+1}} \mathbb{P}^{\text{IDPS}}(\pi^1, s^1, r^1, \cdots, \pi^t, s^t, r^t, \pi^{t+1}|m,\ell) V^{\pi^{t+1}m\ell} \tag{1.39}$$

$$= \sum_{k=1}^{t} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k)) + \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) V^{*m\ell} \tag{1.40}$$

$$- \sum_{m=1}^{M} \sum_{\ell=1}^{L} b(m,\ell) \sum_{\substack{\pi^1,s^1,r^1,\cdots,\\ \pi^t,s^t,r^t}} \mathbb{P}^{\text{IDPS}}(\pi^1, s^1, r^1, \cdots, \pi^t, s^t, r^t|m,\ell) \tag{1.41}$$

$$\times \sum_{\pi^{t+1}} \mathbb{P}^{\text{IDPS}}(\pi^{t+1}|m,\ell,\pi^1, s^1, r^1, \cdots, \pi^t, s^t, r^t) V^{\pi^{t+1}m\ell} \tag{1.42}$$

$$= \sum_{k=1}^{t} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k)) + \sum_{m=1}^{M}\sum_{\ell=1}^{L} b(m,\ell) V^{*m\ell} \tag{1.43}$$

$$- \sum_{m=1}^{M}\sum_{\ell=1}^{L} b(m,\ell) \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}|m,\ell) \sum_{\pi^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(\pi^{t+1}|h^{t+1}) V^{\pi^{t+1}m\ell} \tag{1.44}$$

$$= \sum_{k=1}^{t} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k)) + \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}) \sum_{m=1}^{M}\sum_{\ell=1}^{L} b(m,\ell|h^{t+1}) \underbrace{\sum_{i=1}^{M}\sum_{j=1}^{L} f^*(i,j|h^{t+1})}_{1} V^{*m\ell} \tag{1.45}$$

$$- \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}) \sum_{m=1}^{M}\sum_{\ell=1}^{L} b(m,\ell|h^{t+1}) \sum_{i=1}^{M}\sum_{j=1}^{L} f^*(i,j|h^{t+1}) V^{ijm\ell} \tag{1.46}$$

$$= \sum_{k=1}^{t} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k)) \tag{1.47}$$

$$+ \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}) \sum_{m=1}^{M}\sum_{\ell=1}^{L} b(m,\ell|h^{t+1}) \sum_{i=1}^{M}\sum_{j=1}^{L} f^*(i,j|h^{t+1}) \left[V^{*m\ell} - V^{ijm\ell}\right] \tag{1.48}$$

$$= \sum_{k=1}^{t} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k)) \tag{1.49}$$

$$+ \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}) \sum_{i=1}^{M}\sum_{j=1}^{L} f^*(i,j|h^{t+1}) \sum_{m=1}^{M}\sum_{\ell=1}^{L} b(m,\ell|h^{t+1}) \left[V^{*m\ell} - V^{ijm\ell}\right] \tag{1.50}$$

$$= \sum_{k=1}^{t} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k)) + \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}) \sum_{i=1}^{M}\sum_{j=1}^{L} f^*(i,j|h^{t+1}) \Delta(i,j|h^{t+1}) \tag{1.51}$$

$$= \sum_{k=1}^{t} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k)) + \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}) (f^*(h^{t+1}) \bullet \Delta(h^{t+1})) \tag{1.52}$$

$$= \sum_{k=1}^{t} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k)) + \underset{\boldsymbol{h}^{t+1}}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^{t+1}) \bullet \Delta(\boldsymbol{h}^{t+1})) = \sum_{k=1}^{t+1} \underset{\boldsymbol{h}^k}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^k) \bullet \Delta(\boldsymbol{h}^k)). \tag{1.53}$$

In (1.37), the probability of IDPS choosing policies $\pi^1, \cdots, \pi^{t+1}$ given the true MDP (m,ℓ) is rewritten as the probability of IDPS choosing policies $\pi^1, \cdots, \pi^{t+1}$, observing state trajectories $s^1, \cdots, s^t$, and observing reward trajectories $r^1, \cdots, r^t$ given true MDP (m,ℓ) while summing over $s^1, r^1, \cdots, s^t, r^t$. This allowsf us to write a summation over

histories up to episode $t+1$ in (1.44). Specifically, we factor the probability of implementing $\pi^1, \cdots, \pi^{t+1}$ and observing $s^1, r^1, \cdots, s^t, r^t$ given the (m, ℓ)th MDP is the true one in (1.39) to obtain the desired summation over histories up to episode $t+1$ in (1.44). Moreover, $\mathbb{P}^{\text{IDPS}}(\pi^{t+1}|m, \ell, \pi^1, s^1, r^1, \cdots, \pi^t, s^t, r^t) = \mathbb{P}^{\text{IDPS}}(\pi^{t+1}|m, \ell, h^t)$ in (1.42) reduces to $\mathbb{P}^{\text{IDPS}}(\pi^{t+1}|h^t)$ in (1.44) as the probability of IDPS choosing policy π^{t+1} is independent of the true transition and reward matrices. Now, we wish to combine the second term in (1.43) with the term in (1.44). Since the second term in (1.43) does not include a summation over the history h^{t+1} or over the policy π^{t+1} to implement in episode $t+1$, we artificially introduce these summations in (1.45) to naturally combine quantities in (1.48). In (1.46), we use the definition of $f^*(i, j|h^{t+1})$. Rearranging quantities in (1.48) and (1.50) allows us to use the definition of $\Delta(i, j|h^{t+1})$ in (1.51). Finally, simple manipulations yield the expected result in (1.53). □

The next step is to prove the cumulative expected information gain due to sampling policies according to pmfs $f^*(h^1), \cdots, f^*(h^T)$ is at most the Shannon entropy of the initial belief pmf b; this corresponds to the inequality in (1.21) of Theorem 1.4.2.

We begin by first defining several common information-theoretic quantities. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be finite random variables. The Shannon entropy of random variable $\boldsymbol{X}$ measures the amount of information in random variable $\boldsymbol{X}$ and is defined by

$$\mathcal{E}(\boldsymbol{X}) \stackrel{\text{def}}{=} -\sum_x \mathbb{P}(x) \log(\mathbb{P}(x)).$$

The conditional Shannon entropy of random variable $\boldsymbol{X}$ given $\boldsymbol{Y} = y$ measures the amount of information in random variable $\boldsymbol{X}|\boldsymbol{Y} = y$ and is defined by

$$\mathcal{E}(\boldsymbol{X}|\boldsymbol{Y} = y) \stackrel{\text{def}}{=} -\sum_x \mathbb{P}(x|\boldsymbol{Y} = y) \log(\mathbb{P}(x|\boldsymbol{Y} = y)).$$

The conditional Shannon entropy of $\boldsymbol{X}$ given random variable Y measures the expected

amount of information in random variable $\boldsymbol{X}$ given $\boldsymbol{Y}$ and is defined by

$$\mathcal{E}(\boldsymbol{X}|\boldsymbol{Y}) \stackrel{\text{def}}{=} \sum_{y} \mathbb{P}(y)\mathcal{E}(\boldsymbol{X}|\boldsymbol{Y}=y).$$

We use the Shannon entropy of a random variable interchangeably with the Shannon entropy of the random variable's pmf.

Lemma 1.4.4 derives an alternate expression for the expected information gain of a policy chosen according to the pmf over optimal policies as indicated by IDPS. It will be used to establish the claim in Proposition 1.4.5.

Lemma 1.4.4. *The expected information gain due to sampling a policy according to the pmf indicated by IDPS is equal to the difference in Shannon entropy between the prior belief pmf at the beginning of episode t and the prior belief pmf at the beginning of episode $t+1$. That is,*

$$\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{IDPS}(f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)) = \mathcal{E}(b|\boldsymbol{h}^{t,IDPS}) - \mathcal{E}(b|\boldsymbol{h}^{t+1,IDPS}), \tag{1.54}$$

where random variables $\boldsymbol{h}^{t,IDPS}$ and $\boldsymbol{h}^{t+1,IDPS}$ represent the (random) histories generated by IDPS up to episodes t and $t+1$, respectively.

Proof. We will expand the LHS of (1.54) and express the mutual information between discrete random variables as the difference of Shannon entropies.

The LHS of (1.54) is

$$f^*(h^t) \bullet g(h^t) = \sum_{m=1}^{M}\sum_{\ell=1}^{L} f^*(m,\ell|h^t)g(m,\ell|h^t) \tag{1.55}$$

$$= \sum_{m=1}^{M}\sum_{\ell=1}^{L} \mathbb{P}^{\text{IDPS}}(\boldsymbol{\pi}^t = \pi^{*m\ell}|\boldsymbol{h}^t = h^t) I(b|\boldsymbol{h}^t = h^t; (\boldsymbol{s}^t, \boldsymbol{r}^t)|\boldsymbol{\pi}^t = \pi^{*m\ell}) \tag{1.56}$$

$$= \sum_{m=1}^{M}\sum_{\ell=1}^{L} \mathbb{P}^{\text{IDPS}}(\boldsymbol{\pi}^t = \pi^{*m\ell}|\boldsymbol{h}^t = h^t) \left[\mathcal{E}(b|\boldsymbol{h}^t = h^t) - \mathcal{E}(b|\boldsymbol{h}^t = h^t, \boldsymbol{s}^t, \boldsymbol{r}^t, \boldsymbol{\pi}^t = \pi^{*m\ell})\right] \tag{1.57}$$

$$= \sum_{m=1}^{M}\sum_{\ell=1}^{L} \mathbb{P}^{\text{IDPS}}(\boldsymbol{\pi}^t = \pi^{*m\ell}|\boldsymbol{h}^t = h^t) \left[\, \mathcal{E}(b|\boldsymbol{h}^t = h^t)\right. \tag{1.58}$$

$$- \sum_{s^t} \sum_{r^t} \mathbb{P}(\boldsymbol{s}^t = s^t, \boldsymbol{r}^t = r^t | \boldsymbol{h}^t = h^t, \boldsymbol{\pi}^t = \pi^{*m\ell}) \mathcal{E}(b | \boldsymbol{h}^t = h^t, \boldsymbol{\pi}^t = \pi^{*m\ell}, \boldsymbol{s}^t = s^t, \boldsymbol{r}^t = r^t) \;] \tag{1.59}$$

$$= \mathcal{E}(b | \boldsymbol{h}^t = h^t) \tag{1.60}$$

$$- \sum_{m=1}^{M} \sum_{\ell=1}^{L} \sum_{s^t} \sum_{r^t} \overset{\text{IDPS}}{\mathbb{P}} (\boldsymbol{\pi}^t = \pi^{*m\ell}, \boldsymbol{s}^t = s^t, \boldsymbol{r}^t = r^t | \boldsymbol{h}^t = h^t) \tag{1.61}$$

$$\times \mathcal{E}(b | \boldsymbol{h}^t = h^t, \boldsymbol{\pi}^t = \pi^{*m\ell}, \boldsymbol{s}^t = s^t, \boldsymbol{r}^t = r^t). \tag{1.62}$$

Above, (1.55) holds by the definition of the expected information gain due to implementing the pmf $f^*(h^t)$ indicated by IDPS after incorporating history h^t in the posterior belief pmf. The superscript IDPS in (1.56) emphasizes the probability of $\boldsymbol{\pi}^t = \pi^{*m\ell}$ is determined by IDPS. The equality in (1.57) follows from the well-known fact that the mutual information can be expressed as the difference of Shannon entropies: $I(\boldsymbol{X}; \boldsymbol{Y}) = \mathcal{E}(\boldsymbol{X}) - \mathcal{E}(\boldsymbol{X}|\boldsymbol{Y})$ for discrete random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$. Lines (1.58-1.59) expand the conditional Shannon entropy in (1.57). Finally, (1.60-1.62) separates the difference in the previous line's summand and expresses the probability of a particular policy, state trajectory, and reward trajectory as a joint probability.

Taking the outer expectation of (1.55) yields

$$\overset{\text{IDPS}}{\underset{\boldsymbol{h}^t}{\mathbb{E}}} (f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)) \tag{1.63}$$

$$= \sum_{h^t} \overset{\text{IDPS}}{\mathbb{P}} (\boldsymbol{h}^t = h^t) \Bigg[\mathcal{E}(b | \boldsymbol{h}^t = h^t) \tag{1.64}$$

$$- \sum_{m=1}^{M} \sum_{\ell=1}^{L} \sum_{s^t} \sum_{r^t} \overset{\text{IDPS}}{\mathbb{P}} (\boldsymbol{\pi}^t = \pi^{*m\ell}, \boldsymbol{s}^t = s^t, \boldsymbol{r}^t = r^t | \boldsymbol{h}^t = h^t) \tag{1.65}$$

$$\times \mathcal{E}(b | \boldsymbol{h}^t = h^t, \boldsymbol{\pi}^t = \pi^{*m\ell}, \boldsymbol{s}^t = s^t, \boldsymbol{r}^t = r^t) \Bigg] \tag{1.66}$$

$$= \mathcal{E}(b|\boldsymbol{h}^{t,\text{IDPS}}) - \sum_{h^t} \overset{\text{IDPS}}{\mathbb{P}}(\boldsymbol{h}^t = h^t) \sum_{m=1}^{M} \sum_{\ell=1}^{L} \sum_{s^t} \sum_{r^t} \overset{\text{IDPS}}{\mathbb{P}}(\boldsymbol{\pi}^t = \pi^{*m\ell}, \boldsymbol{s}^t = s^t, \boldsymbol{r}^t = r^t | \boldsymbol{h}^t = h^t) \tag{1.67}$$

$$\times \mathcal{E}(b|\boldsymbol{h}^t = h^t, \boldsymbol{\pi}^t = \pi^{*m\ell}, \boldsymbol{s}^t = s^t, \boldsymbol{r}^t = r^t) \tag{1.68}$$

$$= \mathcal{E}(b|\boldsymbol{h}^{t,\text{IDPS}}) - \sum_{h^t} \sum_{m=1}^{M} \sum_{\ell=1}^{L} \sum_{s^t} \sum_{r^t} \mathbb{P}(\boldsymbol{h}^{t,\text{IDPS}} = h^t, \boldsymbol{\pi}^{t,\text{IDPS}} = \pi^{*m\ell}, \boldsymbol{s}^t = s^t, \boldsymbol{r}^t = r^t) \tag{1.69}$$

$$\times \mathcal{E}(b|\boldsymbol{h}^{t,\text{IDPS}} = h^t, \boldsymbol{\pi}^{t,\text{IDPS}} = \pi^{*m\ell}, \boldsymbol{s}^t = s^t, \boldsymbol{r}^t = r^t) \tag{1.70}$$

$$= \mathcal{E}(b|\boldsymbol{h}^{t,\text{IDPS}}) - \mathcal{E}(b|\boldsymbol{h}^{t+1,\text{IDPS}}). \tag{1.71}$$

Above, the right side of the equality in (1.64) is due to substituting (1.60-1.61) for $f^*(h^t) \bullet g(h^t)$ in (1.63). Line (1.67) separates the difference in (1.66) and rewrites the first term as the Shannon entropy of the belief pmf given a random history up to episode t where policies $\pi^1, \cdots, \pi^{t-1}$ in the history were indicated by IDPS. Lines (1.69-1.70) combine the product of probabilities in (1.67) into a joint probability where the superscript IDPS is moved to the random history and policy as to emphasize that these are the quantities which depend on IDPS; the state and reward trajectories s^t, r^t only depend on IDPS indirectly through the policy chosen by IDPS. Finally, the last line follows from the definition of conditional Shannon entropy and $\boldsymbol{h}^{t+1,\text{IDPS}}$.

Thus, the expected information gain in episode t due to implementing a policy sampled from a pmf $f^*(h^t)$ minimizing the information ratio after observing history $h^{t,\text{IDPS}}$ can be expressed as the difference between the Shannon entropy of the current posterior belief pmf and that of the posterior belief pmf one episode into the future. Thus, the lemma holds. □

Proposition 1.4.5 is the last step to prove the inequality in (1.17) of Theorem 1.4.2.

Proposition 1.4.5. *The cumulative expected information gain due to sampling policies according to pmfs indicated by IDPS over all episodes $t = 1 : T$ is bounded above by the Shannon*

entropy of the initial belief pmf. That is,

$$\sum_{t=1}^{T} \underset{\boldsymbol{h}^t}{\overset{IDPS}{\mathbb{E}}} (f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)) \leq \mathcal{E}(b).$$

Proof. From Lemma 1.4.4, we know

$$\underset{\boldsymbol{h}^t}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)) = \mathcal{E}(b|\boldsymbol{h}^{t,\text{IDPS}}) - \mathcal{E}(b|\boldsymbol{h}^{t+1,\text{IDPS}}) \text{ for } t = 1:T. \tag{1.72}$$

Summing (1.72) over all episodes $t = 1:T$ yields

$$\sum_{t=1}^{T} \underset{\boldsymbol{h}^t}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)) = \sum_{t=1}^{T} \mathcal{E}(b|\boldsymbol{h}^{t,\text{IDPS}}) - \mathcal{E}(b|\boldsymbol{h}^{t+1,\text{IDPS}}) \tag{1.73}$$

$$= \mathcal{E}(b|\boldsymbol{h}^{1,\text{IDPS}}) - \mathcal{E}(b|\boldsymbol{h}^{T+1,\text{IDPS}}) \tag{1.74}$$

$$\leq \mathcal{E}(b|\boldsymbol{h}^{1,\text{IDPS}}) = \mathcal{E}(b). \tag{1.75}$$

The equality in (1.74) holds since the sum in (1.73) telescopes, canceling all terms except the first and the last. The inequality in (1.75) is due to the fact that Shannon entropy is nonnegative for discrete random variables. Finally, the first quantity in (1.75) is the conditional Shannon entropy of the initial belief pmf given the random history produced by IDPS at the beginning of episode 1; this is simply the Shannon entropy of initial belief pmf b since there is no history prior to episode 1. Thus, we have the equality in (1.75). □

The next step necessary to prove the regret bound is to prove the inequality in (1.21) of Theorem 1.4.2. This inequality states that the expected value of the information ratio of a pmf obtained by IDPS is at most N^2ML. It is proved with the help of Lemmas 1.4.6 and 1.4.7. The first of which proves a general statement, and the second of which proves an inequality to be used in Proposition 1.4.8. Both of these lemmas will use the Kullback-Leibler (KL) divergence between two pmfs. The KL divergence between pmfs p and q defined over the same finite sample space $\mathcal{X}$ is defined as $D_{KL}(p||q) \overset{\text{def}}{=} \sum_{x\in\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right)$.

Lemma 1.4.6. *Let P and Q be pmfs defined on the finite set Ω, and let $X : \Omega \to \mathbb{R}$ be a random variable. If $\psi : \mathbb{R} \to \mathbb{R}$ is a function such that $\max_{\omega' \in \Omega} \psi(X(\omega')) - \min_{\omega''} \psi(X(\omega'')) \leq c$ for any constant $c > 0$, then*

$$D_{KL}(P||Q) \geq \frac{2}{c^2} \left(\mathop{\mathbb{E}}_{P} [\psi(\mathbf{X})] - \mathop{\mathbb{E}}_{Q} [\psi(\mathbf{X})] \right)^2 .$$

Proof. This proof is a straightforward generalization of the argument justifying "Fact 9" in [54, Appendix A.1]. Let $\omega_* \in \operatorname{argmin}_{\omega \in \Omega} \psi(X(\omega))$ and $\omega^* \in \operatorname{argmax}_{\omega \in \Omega} \psi(X(\omega))$. Let $\sigma(\omega) \stackrel{\text{def}}{=} \psi(X(\omega)) - \psi(X(\omega_*)) - \frac{c}{2}$. This definition of σ implies $\sigma : \Omega \to \left[-\frac{c}{2}, \frac{c}{2}\right]$, as the smallest value of σ is

$$\psi(X(\omega_*)) - \psi(X(\omega_*)) - \frac{c}{2} = 0 - \frac{c}{2} = -\frac{c}{2},$$

and the largest value of f is

$$\psi(X(\omega^*)) - \psi(X(\omega_*)) - \frac{c}{2} \leq c - \frac{c}{2} = \frac{c}{2}.$$

Moreover, since $\sigma : \Omega \to \left[-\frac{c}{2}, \frac{c}{2}\right]$, we know that $\left|\frac{2}{c}\sigma\right| : \Omega \to [0, 1]$, which is a fact that we will use in (1.77) below.

Now observe that

$$\begin{aligned}
c\sqrt{\frac{1}{2} D_{KL}(P||Q)} \geq c||P - Q||_{TV} &= \frac{c}{2} \sum_{\omega \in \Omega} |P(\omega) - Q(\omega)| && (1.76)\\
&\geq \frac{c}{2} \sum_{\omega \in \Omega} \left| \frac{2}{c}\sigma(\omega) \right| |P(\omega) - Q(\omega)| \geq \sum_{\omega \in \Omega} \sigma(\omega)(P(\omega) - Q(\omega)) && (1.77)\\
&= \sum_{\omega \in \Omega} \sigma(\omega) P(\omega) - \sum_{\omega \in \Omega} \sigma(\omega) Q(\omega) = \mathop{\mathbb{E}}_{P} [\sigma(\mathbf{X})] - \mathop{\mathbb{E}}_{Q} [\sigma(\mathbf{X})] && (1.78)\\
&= \mathop{\mathbb{E}}_{P} \left[\psi(\mathbf{X}) - \psi(\omega_*) - \frac{c}{2} \right] - \mathop{\mathbb{E}}_{Q} \left[\psi(\mathbf{X}) - \psi(\omega_*) - \frac{c}{2} \right] && (1.79)\\
&= \mathop{\mathbb{E}}_{P} [\psi(\mathbf{X})] - \mathop{\mathbb{E}}_{Q} [\psi(\mathbf{X})] . && (1.80)
\end{aligned}$$

The inequality in (1.76) holds by multiplying both sides of Pinsker's inequality [24] by

$c > 0$, while the equality in (1.76) holds from the definition of the total variation distance of probability measures P and Q: $||P - Q||_{TV} \stackrel{\text{def}}{=} \frac{1}{2} \sum_{\omega \in \Omega} |P(\omega) - Q(\omega)|$. The first inequality in (1.77) uses the fact that $y \geq xy$ when $x \in [0, 1]$ and $y \geq 0$ with $x \equiv \left|\frac{2}{c}\sigma(\omega)\right|$ and $y \equiv |P(\omega) - Q(\omega)|$. The second inequality in (1.77) holds by simplifying factors and noting that the LHS only contains nonnegative terms, while the RHS may contain negative terms. The first equality in (1.78) rewrites the sum of a difference on the LHS as the difference of sums in the RHS. The second equality in (1.78) uses the definition of the expectation. The definition of f is used in (1.79), and constants are reduced in (1.80). Thus, steps (1.76-1.80) show that $c\sqrt{\frac{1}{2}D_{KL}(P||Q)} \geq \mathop{\mathbb{E}}_{P}[\psi(\mathbf{X})] - \mathop{\mathbb{E}}_{Q}[\psi(\mathbf{X})]$. Rearranging terms yields the result $D_{KL}(P||Q) \geq \frac{2}{c^2}\left(\mathop{\mathbb{E}}_{P}[\psi(\mathbf{X})] - \mathop{\mathbb{E}}_{Q}[\psi(\mathbf{X})]\right)^2$. This completes the proof. □

Lemma 1.4.7 proves an inequality which will be used in Proposition 1.4.8.

Lemma 1.4.7. *The information gained by implementing policy π^{*ij} satisfies*

$$N^2 g(i, j|h^t) \geq \sum_{m=1}^{M}\sum_{\ell=1}^{L} b(m, \ell|\boldsymbol{h}^t = h^t)\left[V^{ijm\ell} - \sum_{m'=1}^{M}\sum_{\ell'=1}^{L} b(m', \ell'|\boldsymbol{h}^t = h^t)V^{ijm'\ell'}\right]^2. \quad (1.81)$$

Proof. We have,

$$g(i, j|h^t) = I(b|\boldsymbol{h}^t = h^t; (\boldsymbol{s}^t, \boldsymbol{r}^t)|\boldsymbol{\pi}^t = \pi^{*ij}) \quad (1.82)$$

$$= \sum_{m=1}^{M}\sum_{\ell=1}^{L} b(m, \ell|\boldsymbol{h}^t = h) D_{\text{KL}}\left(\mathbb{P}\left(\boldsymbol{s}^t, \boldsymbol{r}^t|\boldsymbol{\pi}^t = \pi^{*ij}, (m, \ell) \text{ is the true MDP}\right) || \, \mathbb{P}\left(\boldsymbol{s}^t, \boldsymbol{r}^t|\boldsymbol{\pi}^t = \pi^{*ij}\right)\right) \quad (1.83)$$

$$\geq \frac{2}{N^2}\sum_{m=1}^{M}\sum_{\ell=1}^{L} b(m, \ell|\boldsymbol{h}^t = h^t) \times \left[\mathbb{E}\left[U(\boldsymbol{s}^t, \boldsymbol{r}^t)|\boldsymbol{\pi}^t = \pi^{*ij}, (m, \ell) \text{ is the true MDP}\right] - \mathbb{E}\left[U(\boldsymbol{s}^t, \boldsymbol{r}^t)|\boldsymbol{\pi}^t = \pi^{*ij}\right]\right]^2 \quad (1.84)$$

$$\geq \frac{1}{N^2}\sum_{m=1}^{M}\sum_{\ell=1}^{L} b(m, \ell|\boldsymbol{h}^t = h^t)\left[V^{ijm\ell} - \sum_{m'=1}^{M}\sum_{\ell'=1}^{L} b(m', \ell'|\boldsymbol{h}^t = h^t)V^{ijm'\ell'}\right]^2. \quad (1.85)$$

The equality in (1.82) holds by the definition of $g(i,j|h^t)$, and the one in (1.83) holds by the well-known fact that $I(\boldsymbol{X};\boldsymbol{Y}) = \underset{\boldsymbol{X}}{\mathbb{E}}(D_{KL}(\mathbb{P}(\boldsymbol{Y}|\boldsymbol{X})||\,\mathbb{P}(\boldsymbol{Y})))$ for random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$. To obtain the inequality in (1.84), we recall from (1.4) the implication of Assumption 1.4.1 that the difference in the total rewards earned in one episode by any two policies is at most N. Then, we apply Lemma 1.4.6 with $\psi \equiv U$, where $U(\mathbf{s}^t, \mathbf{r}^t)$ is the total reward earned in episode t when the state and reward trajectories are, respectively, $\mathbf{s}^t$ and $\mathbf{r}^t$. Continuing on, the first expectation in (1.84) is the expected reward given policy $\boldsymbol{\pi}^t = \pi^{*ij}$ is implemented in the (m,ℓ)th MDP, and the second expectation is the expected reward in episode t given policy $\boldsymbol{\pi}^t = \pi^{*ij}$ is implemented. The aforementioned expectations are, respectively, written parsimoniously as $V^{ijm\ell}$ and $\sum_{m'=1}^{M}\sum_{\ell'=1}^{L} b(m',\ell'|\boldsymbol{h}^t = h^t)V^{ijm'\ell'}$ in (1.85). Multiplying (1.82-1.85) by N^2 yields the desired inequality. □

Finally, Proposition 1.4.8 shows the expected value of the information ratio of a pmf obtained by IDPS is at most N^2ML.

Proposition 1.4.8. *The expected value of the information ratio of an optimal solution $f^*(h^t)$ to (1.16) is at most N^2ML. That is, $\underset{\boldsymbol{h}^t}{\overset{IDPS}{\mathbb{E}}}(\phi(f^*|\boldsymbol{h}^t)) \leq N^2ML$.*

Proof. First, we prove the expected value of the information ratio due to sampling a policy according to PS is at most N^2ML. Given history $\boldsymbol{h}^t = h^t$, define probability vector $f(h^t) \in \mathcal{P}(ML)$ such that $f(m,\ell|h^t) \overset{\text{def}}{=} b(m,\ell|\boldsymbol{h}^t = h^t)$ for $m = 1,\cdots,M$ and $\ell = 1,\cdots,L$. Meaning, policy $\pi^{*m\ell}$ will be chosen with the same probability as the decision-maker's belief that the true MDP is the one with transition probability matrix P^m and reward matrix R^ℓ. Observe that $(f(h^t) \bullet \Delta(h^t))^2 \leq N^2ML(f(h^t) \bullet g(h^t))$ as

$$[f(h^t) \bullet \Delta(h^t)]^2 = \left[\sum_{m=1}^{M}\sum_{\ell=1}^{L} f(m,\ell|h^t)\Delta(m,\ell|h^t)\right]^2 \tag{1.86}$$

$$= \left[\sum_{m=1}^{M}\sum_{\ell=1}^{L} f(m,\ell|h^t)\sum_{i=1}^{M}\sum_{j=1}^{L} f(i,j|h^t)(V^{*ij} - V^{m\ell ij})\right]^2 \tag{1.87}$$

$$= \left[\cancelto{1}{\sum_{m=1}^{M}\sum_{\ell=1}^{L} f(m,\ell|h^t)} \sum_{i=1}^{M}\sum_{j=1}^{L} f(i,j|h^t)V^{*ij} - \sum_{m=1}^{M}\sum_{\ell=1}^{L} f(m,\ell|h^t) \sum_{i=1}^{M}\sum_{j=1}^{L} f(i,j|h^t)V^{m\ell ij} \right]^2 \tag{1.88}$$

$$= \left[\sum_{i=1}^{M}\sum_{j=1}^{L} f(i,j|h^t)V^{*ij} - \sum_{m=1}^{M}\sum_{\ell=1}^{L} f(m,\ell|h^t) \sum_{i=1}^{M}\sum_{j=1}^{L} f(i,j|h^t)V^{m\ell ij} \right]^2 \tag{1.89}$$

$$= \left[\sum_{m=1}^{M}\sum_{\ell=1}^{L} f(m,\ell|h^t)V^{*m\ell} - \sum_{m=1}^{M}\sum_{\ell=1}^{L} f(m,\ell|h^t) \sum_{i=1}^{M}\sum_{j=1}^{L} f(i,j|h^t)V^{m\ell ij} \right]^2 \tag{1.90}$$

$$= \left[\sum_{m=1}^{M}\sum_{\ell=1}^{L} f(m,\ell|h^t) \left(V^{*m\ell} - \sum_{i=1}^{M}\sum_{j=1}^{L} f(i,j|h^t)V^{m\ell ij} \right) \right]^2 \tag{1.91}$$

$$\leq \sum_{m=1}^{M}\sum_{\ell=1}^{L} (1)^2 \sum_{m=1}^{M}\sum_{\ell=1}^{L} \left(f(m,\ell|h^t) \left(V^{*m\ell} - \sum_{i=1}^{M}\sum_{j=1}^{L} f(i,j|h^t)V^{m\ell ij} \right) \right)^2 \tag{1.92}$$

$$= ML \sum_{m=1}^{M}\sum_{\ell=1}^{L} \left(f(m,\ell|h^t) \right)^2 \left(V^{*m\ell} - \sum_{i=1}^{M}\sum_{j=1}^{L} f(i,j|h^t)V^{m\ell ij} \right)^2 \tag{1.93}$$

$$\leq ML \sum_{m=1}^{M}\sum_{\ell=1}^{L} f(m,\ell|h^t) \sum_{m'=1}^{M}\sum_{\ell'=1}^{L} f(m',\ell'|h^t) \left(V^{m\ell m'\ell'} - \sum_{i=1}^{M}\sum_{j=1}^{L} f(i,j|h^t)V^{m\ell ij} \right)^2 \tag{1.94}$$

$$\leq N^2 ML \sum_{m=1}^{M}\sum_{\ell=1}^{L} f(m,\ell|h^t) g(m,\ell|h^t) = N^2 ML (f(h^t) \bullet g(h^t)). \tag{1.95}$$

The equalities in (1.86) and (1.87) hold by definition. Line (1.88) rewrites the difference found in the previous line as the difference of two terms; (1.89) recognizes $\sum_{m=1}^{M}\sum_{\ell=1}^{L} f(m,\ell|h^t) = 1$; (1.90) changes the index of the first two summations from i, j to m, ℓ; and (1.91) groups similar terms. Moving forward, (1.92) applies the Cauchy-Schwarz inequality $\left(\sum_{m,\ell} x_{m\ell}\, y_{m\ell} \right)^2 \leq \sum_{m,\ell} x_{m\ell}^2 \sum_{m,\ell} y_{m\ell}^2$ with $x_{m\ell} \equiv 1$ and $y_{m\ell} \equiv f(m,\ell|h^t) \left(V^{*m\ell} - \sum_{i=1}^{M}\sum_{j=1}^{L} f(i,j|h^t)V^{m\ell ij} \right)$. Line (1.93) simplifies the previous line, and (1.94) bounds the quantity in (1.93) by adding additional nonnegative terms, doing so allows us to use Lemma 1.4.4 in the following line. The last equality holds by definition.

Collectively, (1.86-1.95) show $\left(f(h^t) \bullet \Delta(h^t)\right)^2 \leq N^2ML(f(h^t) \bullet g(h^t))$, or equivalently, the information ratio of pmf $f(h^t)$ chosen by PS after observing history h^t is bounded by N^2ML. That is,

$$\phi(f(h^t)|h^t) = \left(f(h^t) \bullet \Delta(h^t)\right)^2 / (f(h^t) \bullet g(h^t)) \leq N^2ML.$$

Since $f^*(h^t)$ minimizes the information ratio, we have that

$$\phi(f^*|h^t) \leq \phi(f(h^t)|h^t) \leq N^2ML, \tag{1.96}$$

which reduces to $\phi(f^*|h^t) \leq N^2ML$ and explains why

$$\underset{\boldsymbol{h}^t}{\overset{\text{IDPS}}{\mathbb{E}}}(\phi(f^*|\boldsymbol{h}^t)) \leq \underset{\boldsymbol{h}^t}{\overset{\text{IDPS}}{\mathbb{E}}}(N^2ML) = N^2ML.$$

□

Now, Theorem 1.4.2 has officially been established as all of its dependencies have now been proven.

The regret bound of IDPS in Theorem 1.4.2 incorporates prior information in the form an initial belief pmf — a fundamental quantity of Bayesian frameworks. While Corollary 1.4.9 removes this dependency and provides a bound independent of the prior belief.

Corollary 1.4.9. *The regret of IDPS is bounded by $N\sqrt{TML\log(ML)}$, where N is the number of time-stages, T is the number of episodes, M is the number of possible transition probability matrices, and L is the number of possible reward matrices. That is, Regret$(IDPS, T) \leq N\sqrt{TML\log(ML)}$.*

Proof. By use of Theorem 1.4.2 and the well-known fact that $\mathcal{E}(\boldsymbol{X}) \leq \log(|\mathcal{X}|)$ for any discrete random variable $\boldsymbol{X}$ with support $\mathcal{X}$, we have, Regret$(\text{IDPS}, T) \leq N\sqrt{TML\mathcal{E}(b)} \leq N\sqrt{TML\log(ML)}$. □

Theorem 1.4.2 and Corollary 1.4.9 imply Corollary 1.4.10 below, which states our results for the specific case where either the transition probability matrix or reward matrix is known.

Corollary 1.4.10. *The following two statements are true.*

(a) Suppose the reward matrix R is known, and b is a belief pmf over the set of M possible transition probability matrices $\{P^1, \cdots, P^M\}$ so that $b(m)$ is the decision-maker's belief that transition probability matrix P^m is the true one. Then, the regret of IDPS is at most $N\sqrt{TM\mathcal{E}(b)} \leq N\sqrt{TM\log(M)}$.

(b) Suppose the transition probability matrix P is known, and b is a belief pmf over the set of L possible reward matrices $\{R^1, \cdots, R^L\}$ so that $b(\ell)$ is the decision-maker's belief that reward matrix R^ℓ is the true one. Then, the regret of IDPS is at most $N\sqrt{TL\mathcal{E}(b)} \leq N\sqrt{TL\log(L)}$.

The previous theorem and two corollaries also hold for PS as the chain of inequalities in (1.17-1.21) of Theorem 1.4.2 was originally introduced for PS and then used for IDS in the multi-armed bandit setting. Unfortunately, this method of analyzing the regret does not yield distinct bounds for IDPS and PS, similarly to how it did not yield distinct bounds for IDS and PS. However, (1.96) in the proof of Proposition 1.4.8, which bounds the information ratio's value of pmf f by that of pmf f^* where pmf f is determined by PS and f^* is determined by IDPS, suggests that IDPS may perform better than PS. Section 1.5 explores this empirically in the setting of Corollary 1.4.10.

1.5 Computational results

The performance of IDPS is compared to that of PS via numerical experiments where only the transition probability matrices are uncertain in the following three applications: machine repair, queuing control, and dynamic pricing [7]. These applications are well-known and their MDP models can be easily solved using Bellman's equations in the complete information case.

1. **Machine repair**. Consider the problem of maintaining a machine over N time-stages at minimal cost. States $s = 1 : S$ describe the condition of the machine. The cost of operating a machine with condition s is $c(s)$. We assume $c(1) \leq \cdots \leq c(S)$ so that condition s is better than condition $s + 1$ and $s = 1$ is the best possible condition. At each time-stage, the decision-maker can choose to repair the machine before operating. Repairing the machine moves its condition to state $s = 1$ at a cost of R. The machine's condition deteriorates probabilistically if it is not repaired.

   We consider two models for probabilistic deterioration: Bernoulli and truncated geometric. In the Bernoulli model, the machine's condition remains the same with probability p and deteriorates to condition $s + 1$ with probability $1 - p$ for every state $s < S$. The machine's condition remains in state $s = S$, indicating the worst possible condition, with probability 1. In the truncated geometric model, we assume the machine's condition deteriorates from condition s to condition $s + i$ with a probability proportional to p^i for $i = 0 : S - s$.

2. **Queuing control.** Consider the problem of determining the speed of service for customers in a finite-capacity queuing system over N time-stages. Customers arrive to the queue at the beginning of each time-stage. The probability of $i = 1, 2, \cdots$ customers arriving at the queue is p_i. Customers who find the queue full, meaning the system already has S customers, leave without service. The number of customers arriving across different time-stages is independent. Service also begins (ends) at the beginning (end) of a time-stage. The probability that a customer completes service during one time-stage is q_f when the service is fast and q_s when the service is slow, with $q_f > q_s$. The speed of service can be changed at the beginning of each time-stage. The cost of operating the system with a fast (slow) service is c_f (c_s). An additional per stage cost of $v(i)$ is incurred when there are i customers in the queuing system, and a terminal cost of $b(i)$ is incurred when there are i customers left in the queuing system at the end of time-stage N. The decision-maker chooses which type of service

(fast/slow) to offer during each time-stage in order to minimize the expected total cost.

We consider two different stochastic arrival models: the Poisson model and the Binomial model. In the Poisson model, we assume customers arrive according to a Poisson pmf with parameter λ. In the Binomial model, we assume customers arrive according to a Binomial pmf with parameters (B, p).

3. **Dynamic pricing.** Consider the problem of operating an inventory system over N time-stages. At the beginning of each time-stage, the seller observes inventory level $s \in \{0, 1, \cdots, S\}$ and sets the price per unit at $p \in A$, where A is a finite set. This price induces a demand d with probability $\mathbb{P}(d|p)$. The seller sells $\min(s, d)$ units each at price p. A holding cost of h is incurred per unit that remains in inventory after this sale. This costs $(s - \min(s, d))h$. Excess demand is lost at a penalty cost of c per unit; that is, lost demand costs $(d - \min(s, d))c$. Inventory that remains at the end of time-stage N is worthless. The decision-maker chooses prices to maximize expected total profit.

   We consider two probabilistic price-demand functions. The first is Poisson with mean $\lambda e^{-\alpha p}$ for constants $\lambda > 0$ and $\alpha > 0$. The second is Binomial$(B, e^{-\alpha p})$ with constants $B > 0$ (integer) and $\alpha > 0$.

1.5.1 Experimental design

The computational results for each application and model will follow the same format, which we will now motivate and describe.

Recall the regret of IDPS is defined as

$$\text{Regret}(\text{IDPS}, T) = \sum_{m=1}^{M} b_1(m) \left(TV^{*m} - \overset{\text{IDPS}}{\mathbb{E}} \left[\sum_{t=1}^{T} V^{\boldsymbol{\pi}^t m} \right] \right). \tag{1.97}$$

Similarly, the regret of PS is defined as

$$\text{Regret}(\text{PS}, T) = \sum_{m=1}^{M} b_1(m) \left(TV^{*m} - \overset{\text{PS}}{\mathbb{E}} \left[\sum_{t=1}^{T} V^{\boldsymbol{\pi}^t m} \right] \right). \tag{1.98}$$

In (1.98), $\boldsymbol{\pi}^t$ is the random policy chosen by PS for episode t, and the expectation depends on PS similarly to how the expectation in (1.97) depends on IDPS. To take a closer look at the regrets, we will focus on the parenthesized factors in (1.97) and (1.98). These quantities are the regrets of each algorithm given the true transition probability matrix is the mth one. We will assume the decision-maker's prior belief b_1 and initial state pmf ρ_1 are discrete uniform. We will compute TV^{*m} exactly and estimate the expectations by averaging the random cumulative reward from implementing random policy $\boldsymbol{\pi}^t$ over 50 independent replications.

For every model, we will provide a plot displaying the average cumulative regret as a function of episode t. Each one of these plots will have $M \times 2$ lines, one for each parenthesized factor in (1.97-1.98). The line style will indicate which transition probability matrix is assumed to be the true one for all episodes, and the line's color will indicate which algorithm was used. Blue lines will indicate that IDPS was used, while green lines will be used for PS. Then, we will use statistical hypothesis testing to further examine the plots. Specifically, we will test whether the expected cumulative regret at episode T when implementing IDPS is smaller than that when implementing PS for each possible true transition probability matrix. Pictorially, this corresponds to testing whether the rightmost point on a blue (IDPS) line is statistically smaller than the rightmost point on a green (PS) line with the same line style. Thus, we will conduct a total of M hypothesis tests using the data sets used to create the aforementioned rightmost points. The results of all tests will be provided in a table.

Finally, we checked that the regret-per-episode of each scenario approaches zero as T increases. This should be the case as our bound implies this quantity asymptotically vanishes at the rate $1/\sqrt{T}$ for both IDPS and PS. Plots showing these results are not provided for the sake of brevity. The cvxpy package in Python 3.8 was used to solve all convex optimization

Table 1.1: Unknown and known parameter values for the machine repair application.

	Unknown parameter values	Known parameter values
Bernoulli model	$p \in \{0.2, 0.4, 0.6, 0.8\}$	$S = 4, N = 5, c(s) = s,$ $R = 4.5, T = 40$
Truncated geometric model	$p \in \{0.15, 0.30, 0.45, 0.60,$ $0.75, 0.9\}$	$S = 4, N = 5, c(s) = s,$ $R = 8, T = 80$

problems required by IDPS.

1.5.2 Machine repair

Probability p fully characterizes the transition probability matrices in this application, making the choice of a transition probability matrix equivalent to the choice of a value for p. Table 1.1 summarizes the unknown and known parameter values for both of the considered models.

Bernoulli model. The choice of parameter values imply $\pi^{*1}, \cdots, \pi^{*4}$ are all distinct. Figure 1.2 compares the cumulative regret of IDPS averaged over 50 independent replications with that of PS averaged over 50 independent replications. Visually, IDPS outperforms PS in every scenario. Table 1.2 shows the outperformance by IDPS is statistically significant in every case, except for when $p = 0.8$. The performance of the two algorithms is most similar when the probability p of a machine remaining in the same condition is large ($p = 0.8$) and most distinct when p is small ($p = 0.2$).

Truncated geometric model. The choice of parameter values yield four distinct optimal policies. Namely, the optimal policies $\pi^{*1}, \pi^{*2}, \pi^{*4}$, and π^{*5} are distinct, while $\pi^{*3} = \pi^{*2}$ and $\pi^{*6} = \pi^{*5}$. The average cumulative regret of IDPS and PS plotted against the number of episodes is shown in Figure 1.3 below.

In contrast to the Bernoulli model, Table 1.3 shows the performance difference between IDPS and PS is, generally, larger when the true probability p is larger. Further, recall that

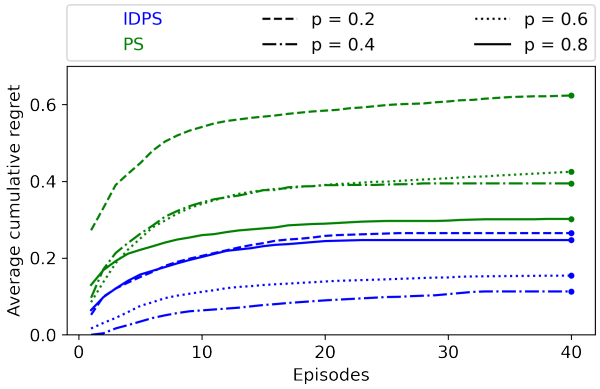


Figure 1.2: Cumulative regret of IDPS and PS for the machine repair problem with the Bernoulli model.

Table 1.2: Results from statistically testing whether the mean cumulative regret of IDPS at episode $T = 40$ is less than that of PS using data from the 50 independent replications.

p	Statistical test result	Average cumulative regret: IDPS, PS	Standard error: IDPS, PS
0.2	significant	0.265, 0.624	0.038, 0.080
0.4	significant	0.113, 0.395	0.040, 0.032
0.6	significant	0.154, 0.425	0.022, 0.042
0.8	not significant	0.247, 0.302	0.032, 0.039

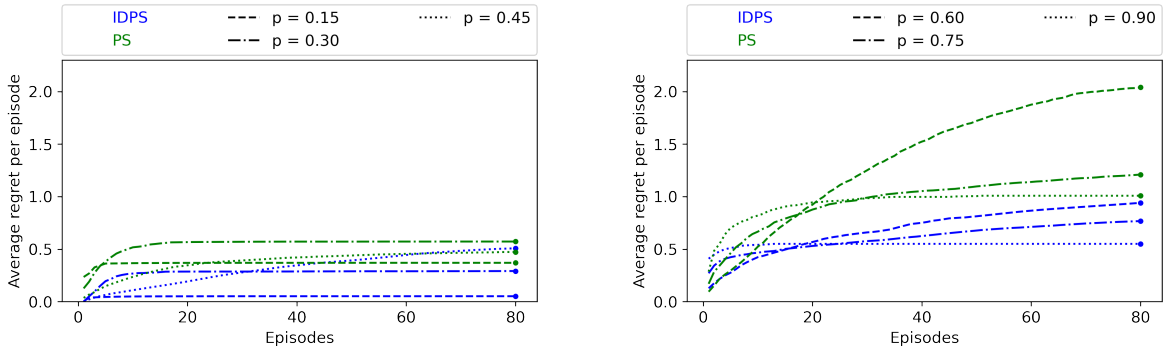


Figure 1.3: Average cumulative regret of IDPS and PS for the machine repair problem with the truncated geometric model.

Table 1.3: Results from statistically testing whether the mean cumulative regret of IDPS at episode $T = 80$ is less than that of PS using data from the 50 independent replications.

p	Statistical test result	Average cumulative regret: IDPS, PS	Standard error: IDPS, PS
0.15	significant	0.052, 0.370	0.015, 0.047
0.30	significant	0.291, 0.572	0.054, 0.066
0.45	not significant	0.507, 0.473	0.079, 0.050
0.60	significant	0.940, 2.04	0.183, 0.212
0.75	significant	0.767, 1.210	0.115, 0.124
0.90	significant	0.550, 1.008	0.044, 0.078

$\pi^{*2} = \pi^{*3}$ and $\pi^{*5} = \pi^{*6}$ is an artifact of this application given the chosen parameters. Interestingly, the cumulative regret when $p = 0.30$ and $p = 0.45$ is similar, while that when $p = 0.75$ and $p = 0.90$ is not; this remark highlights the importance of knowing the true transition probability matrix instead of simply the best policy for the decision-maker to choose. The only instance where IDPS did not statistically outperform PS was when $p = 0.45$.

1.5.3 Queue control

We consider the Poisson and binomial stochastic arrival models as well as a scenario where the service probabilities q_s and q_f are uncertain. In the aforementioned scenario, the decision-maker does not know the true value of the pair (q_s, q_f). Instead, the decision-maker only knows that the pair takes one of M possible values $\{(q_s^1, q_f^1), \cdots, (q_s^M, q_f^M)\}$. Customers are assumed to arrive stochastically according to a Poisson pmf.

The transition probability matrices are fully characterized, respectively, by parameters $\lambda, (B, p)$, and (q_s, q_f) in the three scenarios. Table 1.4 summarizes the unknown and parameter values.

Poisson model. The optimal policies when $\lambda = 0.2, 0.4$, and 0.8 are distinct, and the optimal policy for the MDP whose transition probability matrix is characterized by $\lambda = 0.4$

Table 1.4: Unknown and known parameter values for the queue control application.

	Unknown parameter values	Known parameter values
Poisson model	$\lambda \in \{0.2, 0.4, 0.6, 0.8\}$	$S = 4, N = 5, v(s) = 2s, b(s) = s,$ $q_f = 0.8, q_s = 0.3, c_f = 4, c_s = 3, T = 60$
Binomial model	$(B, p) \in \{(3, 0.25), (3, 0.75),$ $(5, 0.30), (5, 0.90)\}$	$S = 4, N = 5, v(s) = 3s, b(s) = s,$ $q_f = 0.8, q_s = 0.3, c_f = 3, c_s = 1, T = 50$
Uncertain service probabilities	$(q_s, q_f) \in \{(0.3, 0.8), (0.2, 0.9),$ $(0.4, 0.8)\}$	$S = 4, N = 5, v(s) = s, b(s) = s,$ $c_f = 5, c_s = 4, T = 100$, and customers arrive according to Poisson(1)

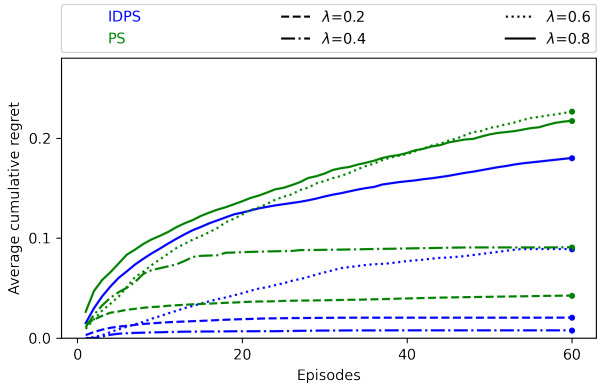


Figure 1.4: Average cumulative regret of IDPS and PS for the queue control problem with the Poisson model.

is the same as that for the MDP whose transition probability matrix is characterized by $\lambda = 0.6$. The average cumulative regret of IDPS and PS is shown in Figure 1.4 below. Looking at Table 1.5 below, we see that both the average cumulative regret and standard error are (generally) smaller when λ is smaller, and the difference in performance is also smaller for such values of λ. The outperformance by IDPS is significant for all values of λ, except for $\lambda = 0.8$.

Binomial model. Recall from Table 1.4 that the decision-maker is faced with four possible choices for the pair (B, p): $(3, 0.25)$, $(3, 0.75)$, $(5, 0.30)$, $(5, 0.90)$. The first two choices

Table 1.5: Results from statistically testing whether the mean cumulative regret of IDPS at episode $T = 60$ is less than that of PS using data from the 50 independent replications.

λ	Statistical test result	Average cumulative regret: IDPS, PS	Standard error: IDPS, PS
0.2	significant	0.021, 0.043	0.002, 0.005
0.4	significant	0.008, 0.091	0.002, 0.009
0.6	significant	0.089, 0.226	0.023, 0.029
0.8	not significant	0.18, 0.217	0.023, 0.019

of (B, p) assume that there are *fewer* customers than the maximum number of customers allowed in the system, while the last two choices of (B, p) assume there are *more* customers than the previously mentioned maximum number.

Figure 1.5 shows the pmf for each possible value of (B, p). We see that it is more probable that less customers will arrive in one time-stage when $(B, p) = (3, 0.25), (5, 0.30)$; the optimal policies for these values of (B, p) are identical. On the other hand, the probability increases as the number of customers arriving in one time-stage increases when $(B, p) = (5, 0.90)$, though this is not the case when $(B, p) = (3, 0.75)$. The optimal policies when $(B, p) = (5, 0.90)$ and $(B, p) = (3, 0.75)$ are distinct. Thus, there are a total of three unique optimal policies. Figure 1.6 shows the average cumulative regret of IDPS and PS. This regret is significantly larger for PS when $(B, p) = (3, 0.25)$ and $(B, p) = (5, 0.30)$.

Table 1.6 shows the expected cumulative regret of IDPS is statistically less than that of PS for $(B, p) = (3, 0.25), (5, 0.30)$; recall that the optimal policies for these parameter values were the same. The comparison for $(B, p) = (3, 0.75), (5, 0.90)$ was not statistically significant.

Uncertain service probabilities. The chosen parameter values for this scenario yield three unique optimal policies. The average cumulative regret of IDPS and PS over $T = 100$ episodes is shown in Figure 1.7 below. The results in Table 1.7 show that the expected regret of IDPS is statistically significantly less than that of PS when $(q_s, q_f) = (0.3, 0.8), (0.2, 0.9)$; this is not the case when $(q_s, q_f) = (0.4, 0.8)$.

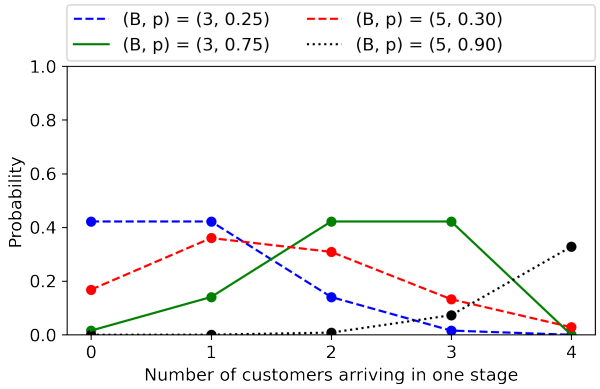


Figure 1.5: Probability of $s = 0 : 4$ customers arriving in one stage given different values of (B, p). Distinct line styles indicate the corresponding optimal policies are distinct.

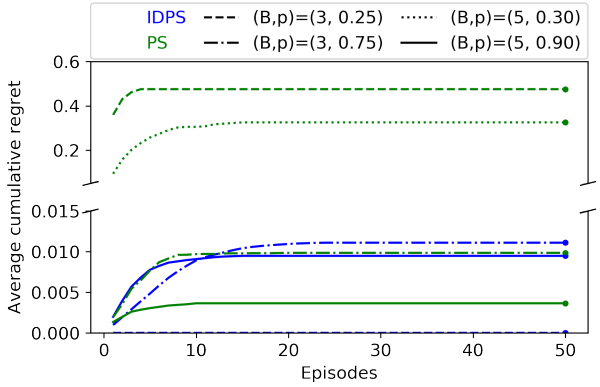


Figure 1.6: Average cumulative regret of IDPS and PS for the queue control problem with the binomial model.

Table 1.6: Results from statistically testing whether the mean cumulative regret of IDPS at episode $T = 50$ is less than that of PS using data from the 50 independent replications.

(B, p)	Statistical test result	Average cumulative regret: IDPS, PS	Standard error: IDPS, PS
(3, 0.25)	significant	0.000, 0.476	0.000, 0.062
(3, 0.75)	not significant	0.011, 0.010	0.001, 0.001
(5, 0.30)	significant	0.000, 0.326	0.000, 0.053
(5, 0.90)	not significant	0.009, 0.004	0.001, 0.000

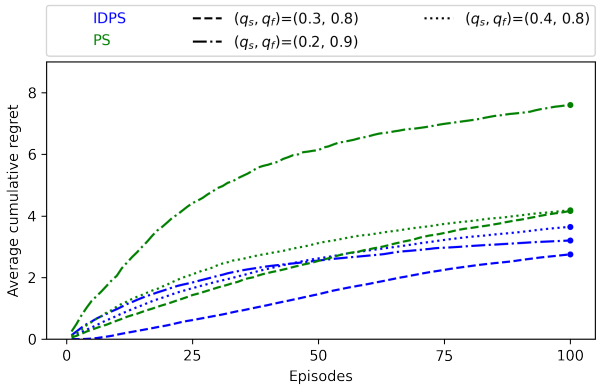


Figure 1.7: Average cumulative regret of IDPS and PS for the queue control problem with uncertain probabilities (q_s, q_f).

Table 1.7: Results from statistically testing whether the mean cumulative regret of IDPS at episode $T = 100$ is less than that of PS using data from the 50 independent replications.

(q_s, q_f)	Statistical test result	Average cumulative regret: IDPS, PS	Standard error: IDPS, PS
(0.3, 0.8)	significant	2.752, 4.165	0.362, 0.194
(0.2, 0.9)	significant	3.201, 7.608	0.534, 0.696
(0.4, 0.8)	not significant	3.645, 4.192	0.397, 0.363

Table 1.8: Unknown and known parameter values for the dynamic pricing application.

	Unknown parameter values	Known parameter values
Poisson model	$(\lambda, \alpha) \in \{(9, 0.15), (10, 0.20), (9, 0.25), (10, 0.30)\}$	$S = 4, N = 5, h = 0.5,$ $c = 1, T = 60$
Binomial model	$(B, \alpha) \in \{(10, 0.10), (9, 0.20), (10, 0.30), (9, 0.40)\}$	

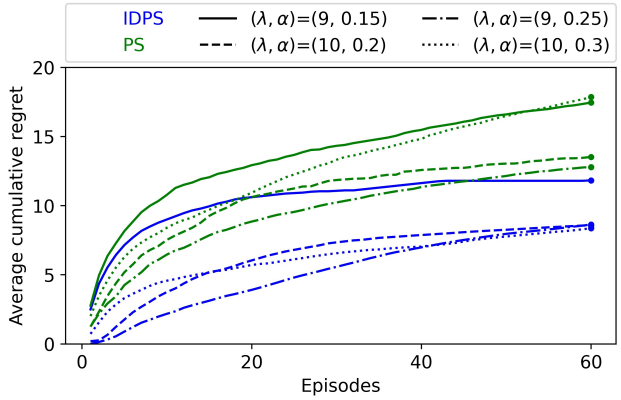


Figure 1.8: Cumulative regret of IDPS and PS for the dynamic pricing problem with the Poisson model.

1.5.4 Dynamic pricing

The transition probability matrices are fully characterized by the parameters of the price-demand function. Table 1.8 summarizes the unknown and known parameter values for the considered functions. Further, the decision-maker chooses prices from the set $A = \{5, 10, 15, 20\}$.

Poisson model. Optimal policies $\pi^{*1}, \cdots, \pi^{*4}$ are distinct. Figure 1.8 compares the cumulative regret of IDPS averaged over 50 independent replications with that of PS averaged over 50 independent replications. Visually, IDPS outperforms PS in every scenario. Table 1.9 shows the outperformance by IDPS is statistically significant in every case.

Binomial model. The choice of parameter values yield four distinct optimal policies. The average cumulative regret of IDPS and PS plotted against the number of episodes is

Table 1.9: Results from statistically testing whether the mean cumulative regret of IDPS at episode $T = 60$ is less than that of PS using data from the 50 independent replications.

(λ, α)	Statistical test result	Average cumulative regret: IDPS, PS	Standard error: IDPS, PS
(9, 0.15)	significant	11.830, 17.453	1.208, 2.887
(10, 0.20)	significant	8.560, 13.521	1.637, 1.35
(9, 0.25)	significant	8.637, 12.794	1.054, 1.118
(10, 0.30)	significant	8.366, 17.853	1.569, 1.243

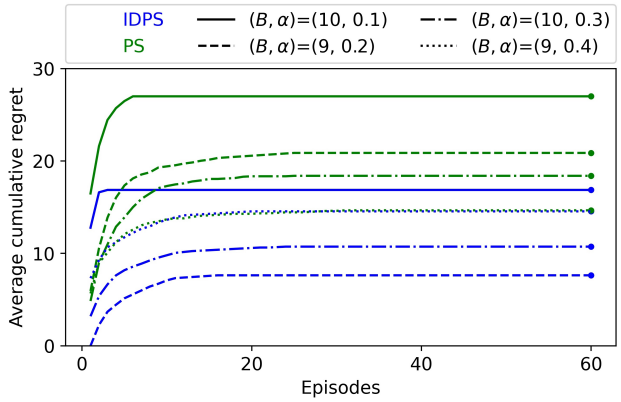


Figure 1.9: Cumulative regret of IDPS and PS for the dynamic pricing problem with the Bernoulli model.

shown in Figure 1.9 below. Table 1.10 shows IDPS statistically outperforms PS in every case. Moreover, both the average cumulative regret and standard error of IDPS is lower than that of PS in every case.

1.6 Conclusions

We proposed IDPS as an extension of IDS for episodic Bayesian MDPs with incomplete information. The episodic Bayesian framework enabled the derivation of a regret bound that depends explicitly on the amount of information in the prior belief pmf and *not* on

Table 1.10: Results from statistically testing whether the mean cumulative regret of IDPS at episode $T = 60$ is less than that of PS using data from the 50 independent replications.

(B, α)	Statistical test result	Average cumulative regret: IDPS, PS	Standard error: IDPS, PS
(10, 0.10)	significant	16.863, 26.993	0.926, 3.124
(9, 0.20)	significant	7.627, 20.870	1.305, 2.453
(10, 0.30)	significant	10.728, 18.391	1.293, 1.482
(9, 0.40)	significant	14.547, 14.657	1.469, 1.874

the sizes of state- or action-spaces. We compared the performance of IDPS against PS on machine repair, queueing control, and dynamic pricing problems. Information-directed policy sampling statistically outperformed PS in most scenarios. In the next chapter, we will extend these results to partially observable MPDs (POMDPs).

Chapter 2

PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES WITH INCOMPLETE INFORMATION

2.1 Problem formulation

Consider the following partially observable extension of the N-stage MDP from Section 1.3. At the beginning of stage n, the decision-maker cannot observe the system-state s_n. Instead, it measures the state, observes a signal o_n, and then chooses an action a_n. The finite set of possible signal observations is denoted O. As before, the system evolves into state s_{n+1} with probability $p(s_{n+1}|s_n, a_n)$ stored in the transition probability matrix P. Given that action a_n transformed the system into state s_{n+1}, the next measurement produces a signal o_{n+1} with probability $\xi(o_{n+1}|s_{n+1}, a_n)$. These probabilities are stored in $|S| \times |O|$ matrices $\Xi(a)$, for each $a \in A$. The 3-dimensional matrix of size $|S| \times |O| \times |A|$ formed by stacking $\Xi(a)$, across $a \in A$, is denoted Ξ. The reward structure is the same as before, and the decision-maker's goal is to maximize expected total reward. As before, the initial system state is drawn from pmf ρ_1 known to the decision-maker.

The partially observable problem can be reformulated as an MDP whose "information state" at the beginning of stage n is $\rho_n \in \mathcal{P}(|S|)$ [37]. Here, $\rho_n(s_n)$ is the decision-maker's probabilistic belief that the system-state is s_n. Knowing this information state, the decision-maker chooses action a_n, observes signals o_{n+1} upon measuring the unobservable state s_{n+1}, and then calculates the new information state $\rho_{n+1} \in \mathcal{P}(|S|)$. This reformulation relies on standard probability formulas for $\mathbb{P}(o_{n+1}|\rho_n, a_n)$ and $\mathbb{P}(s_{n+1}|\rho_n, a_n, o_{n+1})$. The optimal value function for this reformulated MDP is positively homogeneous and convex polyhedral. This can be utilized in designing a tailored algorithm to solve the resulting Bellman's equations offline via backward recursion [58]. Optimal actions can then be recovered and executed

online as the information state is calculated based on signal observations over stages. An optimal policy for this MDP reformulation is denoted $\mu^* \stackrel{\text{def}}{=} (\mu_1^*, \cdots, \mu_N^*)$, where μ_n^* is a function that assigns actions from A to information pmfs $\rho_n(s_n)$ from $\mathcal{P}(|S|)$, for each stage $n = 1 : N$.

We consider an episodic Bayesian version of the above POMDP, where the decision-maker only knows that the true P is one of M possibilities $\{P^1, \cdots, P^M\}$ *and* that the true Ξ is one of K possibilities $\{\Xi^1, \cdots, \Xi^K\}$. Similarly to the problem of Markov decision processes with incomplete information from Chapter 1, the decision-maker maintains and updates a joint belief pmf $b_t(m, k)$, for $m = 1 : M$ and $k = 1 : K$. The decision-maker uses this belief to learn which pair (P^m, Ξ^k) describes the true POMDP, while choosing actions yielding high rewards in an *online* fashion. Here, we assume the rewards $r(s_{n+1}|s_n, a_s)$ are known for simplicity. The case of unknown rewards is discussed in Remark 2.2.10 of Section 2.2.2. The IDPS algorithm in this setting is described in Section 2.2 and an analysis of its regret is provided in Section 2.2.2.

2.2 Information-Directed Policy Sampling

Let μ^{*mk} be an optimal policy for the (m, k)th POMDP with transition probability matrix P^m and measurement matrix Ξ^k. At the beginning of episode t, IDPS samples and implements a policy μ^t from the set of potential optimal policies $\{\mu^{*11}, \cdots, \mu^{*MK}\}$ according to the pmf $f \in \mathcal{P}(MK)$ which minimizes the information ratio in (2.1) of Algorithm 2. The policy μ^t and stochastic signals o^t observed while implementing policy μ^t are then used to update the decision-maker's belief about the true transition probability matrix P^m and measurement matrix Ξ^k. This process continues over episodes $t = 1 : T$, and, at a high-level, is the same as that from Chapter 1.

However, unlike in Chapter 1, the decision-maker here is not privy to the initial system-state. Instead, the decision-maker only knows the initial information state ρ_1^t. The information state changes each time the decision-maker chooses an action, measures the state, and observes the resulting signal. It is updated N times during one episode as the decision-maker

will take N actions and observe N measurements during one episode. Algorithm 2 below summarizes this process.

Algorithm 2 IDPS for episodic Bayesian POMDPs with incomplete information

Start with initial belief pmf $b_1 \in \mathcal{P}(MK)$.
for episodes $t = 1 : T$ **do**
  Let f_t^* be an optimal solution of the information ratio minimization problem

$$\min_{f \in \mathcal{P}(MK)} \phi_t(f) \stackrel{\text{def}}{=} \frac{(\text{Expected regret of } f \text{ in episode } t \text{ based on belief pmf } b_t)^2}{\text{Expected info. gain of } f \text{ in episode } t \text{ based on belief pmf } b_t}. \quad (2.1)$$

  Choose μ^t sampled from $\{\mu^{*11}, \cdots, \mu^{*MK}\}$ according to f_t^*.
  for stages $n = 1 : N$ **do**
    Execute action $\mu_n^t(\rho_n^t)$
    Observe measurement o_{n+1}^t
    Calculate new information state ρ_{n+1}^t as

$$\rho_{n+1}^t(s_{n+1}^t) = \sum_{m=1}^{M} \sum_{k=1}^{K} b_t(m,k) \frac{\left[\sum_{s_n^t} p^m(s_{n+1}^t | s_n^t, \mu_n^t(\rho_n^t)) \rho_n^t(s_n^t)\right] \xi^k(o_{n+1}^t | s_{n+1}^t, \mu_n^t(\rho_n^t))}{\sum_{s_{n+1}^t} \xi^k(o_{n+1}^t | s_{n+1}^t, \mu_n^t(\rho_n^t)) \sum_{s_n^t} p^m(s_{n+1}^t | s_n^t, \mu_n^t(\rho_n^t)) \rho_n^t(s_n^t)}. \quad (2.2)$$

  end for
  Update belief pmf $b_{t+1} \in \mathcal{P}(MK)$ as

$$b_{t+1}(m,k) \propto b_t(m,k) \prod_{n=1}^{N} \sum_{s_{n+1}^t} \xi^k(o_{n+1}^t | s_{n+1}^t, \mu_n^t(\rho_n^t)) \cdots \sum_{s_n^t} p^m(s_{n+1}^t | s_n^t, \mu_n^t(\rho_n^t)) \rho_n^t(s_n^t). \quad (2.3)$$

end for

We supply precise definitions for each quantity of Algorithm 2 in the next section. Working under Assumption 1.4.1, these definitions allow us to show the regret is at most $N\sqrt{TMK\mathcal{E}(b_1)} \leq N\sqrt{TMK\log(MK)}$ — a bound independent of $|S|$, $|A|$, and $|O|$.

2.2.1 Episodic regret and information gain

The decision variable in (2.1) is the joint pmf $f \in \mathcal{P}(MK)$ over the MK possible optimal policies corresponding to POMDPs with a transition probability matrix from the set $\{P^1, \cdots, P^M\}$ and a measurement matrix from the set $\{\Xi^1, \cdots, \Xi^K\}$. The expected regret of f in episode t based on belief pmf b_t in the numerator of (2.1) equals $f \bullet \Delta_t$. The components of Δ_t are $\Delta_t(i,j) = \sum_{m=1}^{M} \sum_{k=1}^{K} b_t(m,k)[V^{*mk} - V^{ijmk}]$. Here, V^{ijmk} is the expected total reward earned by the decision-maker over stages $n = 1 : N+1$ of the POMDP with transition probability matrix P^m and measurement matrix Ξ^k upon executing the policy that is optimal for the POMDP with transition probability matrix P^i and measurement matrix Ξ^j, and $V^{*mk} = V^{mkmk}$. Similarly, the expected information gain of f in episode t based on belief pmf b_t in the denominator of (2.1) equals $f \bullet g_t$, where $g_t(i,j)$ is the mutual information between the decision-maker's joint belief pmf b_t and the signal observations $o^t \stackrel{\text{def}}{=} (o_2^t, \cdots, o_{N+1}^t)$ induced by the policy that is optimal for the POMDP with transition probability matrix P^i and measurement matrix Ξ^j. Thus, the denominator in problem (2.1) equals

$$\sum_{i=1}^{M} \sum_{j=1}^{K} f(i,j) g_t(i,j) = \sum_{i=1}^{M} \sum_{j=1}^{K} f(i,j) \Bigg(\sum_{m=1}^{M} \sum_{k=1}^{K} b_t(m,k) \sum_{o^t} \mathbb{P}\left(o^t | \mu^{*ij}, \text{POMDP } (m,k) \text{ is the true POMDP}\right) \log\left(\frac{\mathbb{P}\left(o^t | \mu^{*ij}, \text{POMDP } (m,k) \text{ is the true POMDP}\right)}{\mathbb{P}\left(o^t | \mu^{*ij}\right)} \right) \Bigg). \quad (2.4)$$

Let ρ_n^t denote the information state determined by $\mu_1^{*ij}, o_2^t, \ldots, \mu_{n-1}^{*ij}, o_n^t$. This quantity can be derived recursively using Proposition 2.2.1 and allows us to express the first probability in (2.4) as

$$\begin{aligned} &\mathbb{P}\left(o^t | \mu^{*ij}, \text{POMDP } (m,k) \text{ is the true POMDP}\right) \\ &= \mathbb{P}\left(o_2^t, \ldots, o_{N+1}^t | \mu^{*ij}, \text{POMDP } (m,k) \text{ is the true POMDP}\right) \end{aligned}$$

$$= \prod_{n=1}^{N} \mathbb{P}(o_{n+1}^t | o_2^t, \ldots, o_n^t, \mu^{*ij}, \text{POMDP } (m,k) \text{ is the true POMDP}) \tag{2.5}$$

$$= \prod_{n=1}^{N} \mathbb{P}(o_{n+1}^t | \mu_n^{*ij}(\rho_n^t), \text{POMDP } (m,k) \text{ is the true POMDP})$$

$$= \prod_{n=1}^{N} \sum_{s_{n+1}^t} \mathbb{P}(o_{n+1}^t | s_{n+1}^t, \mu_n^{*ij}(\rho_n^t), \text{POMDP } (m,k) \text{ is the true POMDP}) \tag{2.6}$$

$$\mathbb{P}(s_{n+1}^t | \mu_n^{*ij}(\rho_n^t), \text{POMDP } (m,k) \text{ is the true POMDP})$$

$$= \prod_{n=1}^{N} \sum_{s_{n+1}^t} \mathbb{P}(o_{n+1}^t | s_{n+1}^t, \mu_n^{*ij}(\rho_n^t), k) \sum_{s_n^t} \mathbb{P}(s_{n+1}^t | s_n^t, \mu_n^{*ij}(\rho_n^t), m)\, \mathbb{P}(s_n^t | \mu_n^{*ij}(\rho_n^t), m) \tag{2.7}$$

$$= \prod_{n=1}^{N} \sum_{s_{n+1}^t} \xi^k(o_{n+1}^t | s_{n+1}^t, \mu_n^{*ij}(\rho_n^t)) \sum_{s_n^t} p^m(s_{n+1}^t | s_n^t, \mu_n^{*ij}(\rho_n^t)) \rho_n^t(s_n^t). \tag{2.8}$$

Above, the equality in (2.5) uses the chain rule to express the joint probability as a product of conditional probabilities. The equality in (2.6) holds by conditioning on the future system-state s_{n+1}^t and the one in (2.7) holds by conditioning on the current system-state s_n^t. Similarly to as in Chapter 1, the "m" in the conditional represents the event "P^m is the true transition probability matrix", and the "k" in the conditional represents the event "Ξ^k is the true measurement matrix." Finally, (2.8) uses the notation from our problem formulation.

The second probability in (2.4) is

$$\mathbb{P}\left(o^t | \mu^{*ij}\right) = \sum_{m=1}^{M} \sum_{k=1}^{K} b_t(m,k)\, \mathbb{P}\left(o^t | \mu^{*ij}, \text{POMDP } (m,k) \text{ is the true POMDP}\right).$$

As before, (2.1) is convex in the joint pmf f, and its solution will be denoted by f_t^*. Finally, Proposition 2.2.1 and 2.2.2 below justify the way in which the belief state ρ_{n+1}^t and belief pmf b_{t+1} are, respectively, written in (2.2) and (2.3).

Proposition 2.2.1. *This proposition justifies (2.2) in Algorithm 2. It shows that the belief state ρ_{n+1}^t can be calculated using the previous belief state ρ_n^t, the current belief pmf $b_t(m,k)$,*

signal measurement o^t, and policy μ^t as

$$\rho^t_{n+1}(s^t_{n+1}) = \sum_{m=1}^{M}\sum_{k=1}^{K} b_t(m,k)\frac{\left[\sum\limits_{s^t_n} p^m(s^t_{n+1}|s^t_n,\mu^t_n(\rho^t_n))\rho^t_n(s^t_n)\right]\xi^k(o^t_{n+1}|s^t_{n+1},\mu^t_n(\rho^t_n))}{\sum\limits_{s^t_{n+1}}\xi^k(o^t_{n+1}|s^t_{n+1},\mu^t_n(\rho^t_n))\sum\limits_{s^t_n}p^m(s^t_{n+1}|s^t_n,\mu^t_n(\rho^t_n))\rho^t_n(s^t_n)}.$$

Proof. The decision-maker's belief that the system-state is s^t_{n+1} in stage n during episode t can be expressed by conditioning on the true transition probability matrix P^m and the true measurement matrix Ξ^k as

$$\begin{aligned}
&\rho^t_{n+1}(s^t_{n+1}) = \mathbb{P}(s^t_{n+1}|\rho^t_n,\mu^t_n(\rho^t_n),o^t_{n+1})\\
&= \sum_{m=1}^{M}\sum_{k=1}^{K} b_t(m,k)\,\mathbb{P}(s^t_{n+1}|\rho^t_n,\mu^t_n(\rho^t_n),o^t_{n+1},m,k).\\
&= \sum_{m=1}^{M}\sum_{k=1}^{K} b_t(m,k)\frac{\mathbb{P}(s^t_{n+1}|\rho^t_n,\mu^t_n(\rho^t_n),m,k)\,\mathbb{P}(o^t_{n+1}|\rho^t_n,\mu^t_n(\rho^t_n),s^t_{n+1},m,k)}{\mathbb{P}(o^t_{n+1}|\rho^t_n,\mu^t_n(\rho^t_n),m,k)} \qquad (2.9)\\
&= \sum_{m=1}^{M}\sum_{k=1}^{K} b_t(m,k)\frac{\mathbb{P}(s^t_{n+1}|\rho^t_n,\mu^t_n(\rho^t_n),m,k)\,\mathbb{P}(o^t_{n+1}|\mu^t_n(\rho^t_n),s^t_{n+1},k)}{\mathbb{P}(o^t_{n+1}|\rho^t_n,\mu^t_n(\rho^t_n),m,k)} \qquad (2.10)\\
&= \sum_{m=1}^{M}\sum_{k=1}^{K} b_t(m,k)\\
&\times\frac{\left[\sum\limits_{s^t_n}\mathbb{P}(s^t_{n+1}|s^t_n,\rho^t_n,\mu^t_n(\rho^t_n),m,k)\,\mathbb{P}(s^t_n|\rho^t_n,\mu^t_n(\rho^t_n),m,k)\right]\mathbb{P}(o^t_{n+1}|\mu^t_n(\rho^t_n),s^t_{n+1},k)}{\mathbb{P}(o^t_{n+1}|\rho^t_n,\mu^t_n(\rho^t_n),k)} \qquad (2.11)\\
&= \sum_{m=1}^{M}\sum_{k=1}^{K} b_t(m,k)\frac{\left[\sum\limits_{s^t_n} p^m(s^t_{n+1}|s^t_n,\mu^t_n(\rho^t_n))\rho^t_n(s^t_n)\right]\xi^k(o^t_{n+1}|s^t_{n+1},\mu^t_n(\rho^t_n))}{\sum\limits_{s^t_{n+1}}\xi^k(o^t_{n+1}|s^t_{n+1},\mu^t_n(\rho^t_n))\sum\limits_{s^t_n}p^m(s^t_{n+1}|s^t_n,\mu^t_n(\rho^t_n))\rho^t_n(s^t_n)}. \qquad (2.12)
\end{aligned}$$

Line (2.9) follows from Bayes' theorem, (2.10) holds because observations are Markovian and the probability of a signal measurement is independent of m, (2.11) is due to conditioning on s^t_n, and (2.12) follows since

$\mathbb{P}(o^t_{n+1}|\rho^t_n, \mu^t_n(\rho^t_n), m, k) = \sum_{s^t_{n+1}} \xi^k(o^t_{n+1}|s^t_{n+1}, \mu^t_n(\rho^t_n)) \sum_{s^t_n} p^m(s^t_{n+1}|s^t_n, \mu^t_n(\rho^t_n))\rho^t_n(s^t_n)$. This last equality follows from a standard probability formula for POMDPs. Thus, the proposition holds. □

Proposition 2.2.2. *This proposition justifies (2.3). It shows that the belief pmf at the beginning of episode $t+1$ is calculated using*

$$b_{t+1}(m,k) \propto b_t(m,k) \sum_{s^t_2} \xi^k(o^t_2|s^t_2, \mu^t_1(\rho^t_1)) \sum_{s^t_1} p^m(s^t_2|s^t_1, \mu^t_1(\rho^t_1))\rho^t_1(s^t_1) \cdots$$
$$\cdots \sum_{s^t_{N+1}} \xi^k(o^t_{N+1}|s^t_{N+1}, \mu^t_n(\rho^t_n)) \sum_{s^t_N} p^m(s^t_{N+1}|s^t_N, \mu^t_n(\rho^t_n))\rho^t_n(s^t_N).$$

Proof. This derivation is similar to the derivation for Markov decision processes with incomplete information in (1.6-1.12) as

$$\begin{aligned}
b_{t+1}(m,k) &= b_t(m,k|\mu^t, o^t) \\
&= \frac{b_t(m,k)\,\mathbb{P}^{\text{IDPS}}(\mu^t, o^t|m,k)}{\mathbb{P}^{\text{IDPS}}(\mu^t, o^t)} && (2.13)\\
&= \frac{b_t(m,k)\,\mathbb{P}(o^t|\mu^t, m, k)\,\mathbb{P}^{\text{IDPS}}(\mu^t|m,k)}{\mathbb{P}(o^t|\mu^t)\,\mathbb{P}^{\text{IDPS}}(\mu^t)} && (2.14)\\
&= \frac{b_t(m,k)\,\mathbb{P}(o^t|\mu^t, m, k)}{\mathbb{P}(o^t|\mu^t)} && (2.15)\\
&\propto b_t(m,k)\,\mathbb{P}(o^t|\mu^t, m, k) && (2.16)\\
&\propto b_t(m,k) \prod_{n=1}^{N} \sum_{s^t_{n+1}} \xi^k(o^t_{n+1}|s^t_{n+1}, \mu^t_n(\rho^t_n)) \cdots \sum_{s^t_n} p^m(s^t_{n+1}|s^t_n, \mu^t_n(\rho^t_n))\rho^t_n(s^t_n). && (2.17)
\end{aligned}$$

The equality in (2.13) is due to Bayes' theorem, (2.14) expands the joint pmfs in order to cancel $\mathbb{P}^{\text{IDPS}}(\mu^t|m,k)$ with $\mathbb{P}^{\text{IDPS}}(\mu^t)$ in (2.15), and (2.17) uses (2.5-2.8) with $\mu^{*ij} \equiv \mu^t$. □

2.2.2 Regret analysis

We begin by defining the total T-episode regret of IDPS in the episodic Bayesian setting for POMDPs. Let μ^t denote the policy executed by IDPS in episode t, and $V^{\mu^t mk}$ denote the expected total reward earned by this policy over stages $n = 1 : N+1$ of episode t given that the initial information state is pmf ρ_1, if the (m,k)th POMDP is the true one. Thus, the expected total reward earned by the IDPS algorithm in T episodes equals $\overset{\text{IDPS}}{\mathbb{E}}\left[\sum_{t=1}^{T} V^{\boldsymbol{\mu}^t mk}\right]$. The expectation outside this sum is with respect to the randomness in the choice of μ^t. The randomness in μ^t comes from two sources. First, μ^t is sampled from pmf f_t^*. Second, f_t^* itself depends (through the posterior b_t) on policies $\mu^1, \cdots, \mu^{t-1}$ implemented as well as signal measurements $o^1, \cdots, o^{t-1}$ observed in episodes $1 : t-1$. Recall that V^{*mk} is the optimal expected total reward earned by the (optimal) policy μ^{*mk} in stages $n = 1 : N+1$ of POMDP (m,k). Thus, if the decision-maker knew that POMDP (m,k) was the true POMDP, then it would execute policy μ^{*mk} in each episode and earn the optimal reward of TV^{*mk}. That is, the regret of IDPS would be $TV^{*mk} - \overset{\text{IDPS}}{\mathbb{E}}\left[\sum_{t=1}^{T} V^{\boldsymbol{\mu}^t mk}\right]$. Since $b_1(m,k)$ denotes the probability that POMDP (m,k) is the true POMDP, the regret of IDPS is given by

$$\begin{aligned}\text{Regret}(\text{IDPS}, T) &\overset{\text{def}}{=} \sum_{m=1}^{M}\sum_{k=1}^{K} b_1(m,k)\left(TV^{*mk} - \overset{\text{IDPS}}{\mathbb{E}}\left[\sum_{t=1}^{T} V^{\boldsymbol{\mu}^t mk}\right]\right)\\ &= \mathbb{E}\left[TV^{*\boldsymbol{mk}} - \overset{\text{IDPS}}{\mathbb{E}}\left[\sum_{t=1}^{T} V^{\boldsymbol{\mu}^t \boldsymbol{mk}}\right]\right].\end{aligned}$$

An analysis of this regret will include expectations with respect to the policies $\mu^1, \ldots, \mu^{t-1}$ chosen and signal measurements $o^1, \ldots, o^{t-1}$ observed in episodes $1, \ldots, t-1$. Such information will be referred to as the history $h^t = (\mu^1, o^1, \ldots, \mu^{t-1}, o^{t-1})$ up to episode t. As in Section 1.4.2, we will express all quantities dependent on a particular history as such in order to carefully analyze the regret. Specifically, $f \bullet \Delta(h^t)$, $f \bullet g(h^t)$, and $\phi(f|h^t)$ are, respectively, the expected regret suffered, expected information gain, and information ratio of a policy

sampled according to a pmf $f \in \mathcal{P}(MK)$ after observing history h^t. A pmf which minimizes the information ratio after observing history h^t will be denoted by

$$f^*(h^t) \in \underset{f \in \mathcal{P}(MK)}{\operatorname{argmin}} \ \phi(f|h^t) \stackrel{\text{def}}{=} \frac{(f \bullet \Delta(h^t))^2}{f \bullet g(h^t)}. \tag{2.18}$$

For simplicity, we let $\phi(f^*|h^t) \stackrel{\text{def}}{=} \phi(f^*(h^t)|h^t)$ be the information ratio of an optimal solution to the problem in (2.18).

Our analysis of IDPS in this section will follow the same chain of inequalities as in Chapter 1:

$$\text{Regret}(\text{IDPS}, T) \stackrel{(a)}{=} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{IDPS}} [f^*(\boldsymbol{h}^t) \bullet \Delta(\boldsymbol{h}^t)] \stackrel{(b)}{\leq} \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{IDPS}} [\phi(f^*|\boldsymbol{h}^t)]} \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{IDPS}} [f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)]} \tag{2.19}$$

$$\stackrel{(c)}{\leq} \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{IDPS}} [\phi(f^*|\boldsymbol{h}^t)]} \sqrt{\mathcal{E}(b)} \stackrel{(d)}{\leq} \sqrt{\sum_{t=1}^{T} N^2 M L} \sqrt{\mathcal{E}(b)} \stackrel{(e)}{=} N\sqrt{TML\mathcal{E}(b)}. \tag{2.20}$$

However, quantities will be defined as in Section 2.2.1, and the random history at the beginning of episode t will be $\boldsymbol{h}^t = (\boldsymbol{\mu}^1, \boldsymbol{o}^1, \ldots, \boldsymbol{\mu}^{t-1}, \boldsymbol{o}^{t-1})$. As a result, the proofs of this chain of inequalities will follow the template established in Chapter 1 with signal measurements o^t in place of state and reward trajectories (s^t, r^t) and policies μ^t in place of policies π^t.

The following proposition proves equality "(a)" in (2.19).

Proposition 2.2.3. *The regret of IDPS can be expressed as the cumulative sum of the expected episodic regret of each episode when policies are chosen according to IDPS. Namely, we have* $\textit{Regret}(\textit{IDPS}, T) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\textit{IDPS}} (f^*(\boldsymbol{h}^t) \bullet \Delta(\boldsymbol{h}^t))$.

Proof. This proof is similar to that of Proposition 1.4.3 in Section 1.4.2. Specifically, all summations continue to be discrete as there are a discrete number of possible transition probability matrices, observation matrices, observation trajectories, and policies. For in-

stance, $\mu^t \in \{\mu^{*11}, \ldots, \mu^{*MK}\}$ and $o^t \in \underbrace{O \times \ldots \times O}_{((N+1)-2)+1\text{times}}$, for $1 \leq t \leq T$.

Thus, analogously to (1.22-1.53), we have

$$\text{Regret}(\text{IDPS}, t+1) = \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)\left[(t+1)V^{*mk} - \overset{\text{IDPS}}{\mathbb{E}}\left[\sum_{w=1}^{t+1} V^{\boldsymbol{\mu}^w mk}\right]\right]$$

$$= (t+1)\sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)V^{*mk} - \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)\sum_{\mu^1,\cdots,\mu^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(\mu^1,\cdots,\mu^{t+1}|m,k)\sum_{w=1}^{t+1} V^{\mu^w mk} \tag{2.21}$$

$$= \left[t\sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)V^{*mk} - \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)\sum_{\mu^1,\cdots,\mu^t}\sum_{\mu^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(\mu^1,\cdots,\mu^{t+1}|m,k)\sum_{w=1}^{t} V^{\mu^w mk}\right] \tag{2.22}$$

$$+ \left[\sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)V^{*mk} - \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)\sum_{\mu^1,\cdots,\mu^t}\sum_{\mu^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(\mu^1,\cdots,\mu^{t+1}|m,k)V^{\mu^{t+1}mk}\right],$$

where the quantity in (2.21) can be expressed as

$$\left[t\sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)V^{*mk} - \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)\sum_{\mu^1,\cdots,\mu^t}\sum_{\mu^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(\mu^1,\cdots,\mu^{t+1}|m,k)\sum_{w=1}^{t} V^{\mu^w mk}\right]$$

$$= t\sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)V^{*mk}$$

$$- \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)\sum_{\mu^1,\cdots,\mu^t} \overset{\text{IDPS}}{\mathbb{P}}(\mu^1,\cdots,\mu^t|m,k)\left[\sum_{\mu^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(\mu^{t+1}|m,k,\mu^1,\cdots,\mu^t)\right]\overset{1}{\rightarrow}\sum_{l=1}^{t} V^{\mu^w mk}$$

$$= t\sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)V^{*mk} - \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)\sum_{\mu^1,\cdots,\mu^t} \overset{\text{IDPS}}{\mathbb{P}}(\mu^1,\cdots,\mu^t|m,k)\sum_{l=1}^{t} V^{\mu^w mk}$$

$$= \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)\left[tV^{*mk} - \sum_{\mu^1,\cdots,\mu^t} \overset{\text{IDPS}}{\mathbb{P}}(\mu^1,\cdots,\mu^t|m,k)\sum_{l=1}^{t} V^{\mu^w mk}\right]$$

$$= \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k)\left[tV^{*mk} - \overset{\text{IDPS}}{\mathbb{E}}\left[\sum_{w=1}^{t} V^{\boldsymbol{\mu}^w mk}\right]\right]$$

$$= \text{Regret}(\text{IDPS}, t) = \sum_{w=1}^{t} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w)).$$

Substituting $\sum_{w=1}^{t} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w))$ for the quantity in (2.22) yields

$$\begin{aligned}
&\text{Regret}(\text{IDPS}, t+1) \\
&= \sum_{w=1}^{t} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w)) + \left[\sum_{m=1}^{M} \sum_{k=1}^{K} b(m,k) V^{*mk} \right. \\
&\quad \left. - \sum_{m=1}^{M} \sum_{k=1}^{K} b(m,k) \sum_{\mu^1, \cdots, \mu^t} \sum_{\mu^{t+1}} \overset{\text{IDPS}}{\mathbb{P}} (\mu^1, \cdots, \mu^{t+1} | m, k) V^{\mu^{t+1} mk} \right] \\
&= \sum_{w=1}^{t} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w)) + \sum_{m=1}^{M} \sum_{k=1}^{K} b(m,k) V^{*mk} \\
&\quad - \sum_{m=1}^{M} \sum_{k=1}^{K} b(m,k) \sum_{\mu^1, \cdots, \mu^t} \sum_{\mu^{t+1}} \left[\sum_{o^1, \cdots, o^t} \overset{\text{IDPS}}{\mathbb{P}} (\mu^1, o^1, \cdots, \mu^t, o^t, \mu^{t+1} | m, k) \right] V^{\mu^{t+1} mk} \\
&= \sum_{w=1}^{t} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w)) + \sum_{m=1}^{M} \sum_{k=1}^{K} b(m,k) V^{*mk} \\
&\quad - \sum_{m=1}^{M} \sum_{k=1}^{K} b(m,k) \sum_{\substack{\mu^1, o^1, \cdots, \\ \mu^t, o^t}} \sum_{\mu^{t+1}} \overset{\text{IDPS}}{\mathbb{P}} (\mu^1, o^1, \cdots, \mu^t, o^t, \mu^{t+1} | m, k) V^{\mu^{t+1} mk} \\
&= \sum_{w=1}^{t} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w)) + \sum_{m=1}^{M} \sum_{k=1}^{K} b(m,k) V^{*mk} \\
&\quad - \sum_{m=1}^{M} \sum_{k=1}^{K} b(m,k) \sum_{\substack{\mu^1, o^1, \cdots, \\ \mu^t, o^t}} \overset{\text{IDPS}}{\mathbb{P}} (\mu^1, o^1, \cdots, \mu^t, o^t | m, k) \\
&\quad \times \sum_{\mu^{t+1}} \overset{\text{IDPS}}{\mathbb{P}} (\mu^{t+1} | m, k, \mu^1, o^1, \cdots, \mu^t, o^t) V^{\mu^{t+1} mk} \\
&= \sum_{w=1}^{t} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w)) + \sum_{m=1}^{M} \sum_{k=1}^{K} b(m,k) V^{*mk}
\end{aligned}$$

$$
\begin{aligned}
&\quad - \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k) \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}|m,k) \sum_{\mu^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(\mu^{t+1}|h^{t+1}) V^{\mu^{t+1}mk} \\
&= \sum_{w=1}^{t} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}}(f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w)) + \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}) \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k|h^{t+1}) \underbrace{\sum_{i=1}^{M}\sum_{j=1}^{K} f^*(i,j|h^{t+1})}_{1} V^{*mk} \\
&\quad - \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}) \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k|h^{t+1}) \sum_{i=1}^{M}\sum_{j=1}^{K} f^*(i,j|h^{t+1}) V^{ijmk} \\
&= \sum_{w=1}^{t} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}}(f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w)) \\
&\quad + \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}) \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k|h^{t+1}) \sum_{i=1}^{M}\sum_{j=1}^{K} f^*(i,j|h^{t+1}) \left[V^{*mk} - V^{ijmk}\right] \\
&= \sum_{w=1}^{t} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}}(f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w)) \\
&\quad + \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}) \sum_{i=1}^{M}\sum_{j=1}^{K} f^*(i,j|h^{t+1}) \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k|h^{t+1}) \left[V^{*mk} - V^{ijmk}\right] \\
&= \sum_{w=1}^{t} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}}(f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w)) + \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}) \sum_{i=1}^{M}\sum_{j=1}^{K} f^*(i,j|h^{t+1}) \Delta(i,j|h^{t+1}) \\
&= \sum_{w=1}^{t} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}}(f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w)) + \sum_{h^{t+1}} \overset{\text{IDPS}}{\mathbb{P}}(h^{t+1}) (f^*(h^{t+1}) \bullet \Delta(h^{t+1})) \\
&= \sum_{w=1}^{t} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}}(f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w)) + \underset{\boldsymbol{h}^{t+1}}{\overset{\text{IDPS}}{\mathbb{E}}}(f^*(\boldsymbol{h}^{t+1}) \bullet \Delta(\boldsymbol{h}^{t+1})) = \sum_{w=1}^{t+1} \underset{\boldsymbol{h}^w}{\overset{\text{IDPS}}{\mathbb{E}}}(f^*(\boldsymbol{h}^w) \bullet \Delta(\boldsymbol{h}^w)).
\end{aligned}
$$

Hence, the proposition holds. □

The following lemma will be used by Proposition 2.2.5 to prove inequality "(c)" in (2.19).

Lemma 2.2.4. *The expected information gain due to sampling a policy according to the pmf indicated by IDPS is equal to the difference in Shannon entropy between the prior belief pmf at the beginning of episode t and the prior belief pmf at the beginning of episode $t+1$. That is, $\underset{\boldsymbol{h}^t}{\overset{IDPS}{\mathbb{E}}}(f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)) = \mathcal{E}(b|\boldsymbol{h}^{t,IDPS}) - \mathcal{E}(b|\boldsymbol{h}^{t+1,IDPS})$, where random variables $\boldsymbol{h}^{t,IDPS}$ and $\boldsymbol{h}^{t+1,IDPS}$ represent the (random) histories generated by IDPS up to episodes t and $t+1$, respectively.*

Proof. Similarly to Lemma 1.4.4 of Section 1.4.2, we have

$$
\begin{aligned}
&f^*(h^t) \bullet g(h^t) = \sum_{m=1}^{M}\sum_{k=1}^{K} f^*(m,k|h^t) g(m,k|h^t) \\
&= \sum_{m=1}^{M}\sum_{k=1}^{K} \overset{\text{IDPS}}{\mathbb{P}}(\boldsymbol{\mu}^t = \mu^{*mk}|\boldsymbol{h}^t = h^t) I(b|\boldsymbol{h}^t = h^t; \boldsymbol{o}^t|\boldsymbol{\mu}^t = \mu^{*mk}) \\
&= \sum_{m=1}^{M}\sum_{k=1}^{K} \overset{\text{IDPS}}{\mathbb{P}}(\boldsymbol{\mu}^t = \mu^{*mk}|\boldsymbol{h}^t = h^t) \left[\mathcal{E}(b|\boldsymbol{h}^t = h^t) - \mathcal{E}(b|\boldsymbol{h}^t = h^t, \boldsymbol{o}^t, \boldsymbol{\mu}^t = \mu^{*mk})\right] \\
&= \sum_{m=1}^{M}\sum_{k=1}^{K} \overset{\text{IDPS}}{\mathbb{P}}(\boldsymbol{\mu}^t = \mu^{*mk}|\boldsymbol{h}^t = h^t) \left[\, \mathcal{E}(b|\boldsymbol{h}^t = h^t)\right. \\
&\left. - \sum_{o^t} \mathbb{P}(\boldsymbol{o}^t = o^t|\boldsymbol{h}^t = h^t, \boldsymbol{\mu}^t = \mu^{*mk}) \mathcal{E}(b|\boldsymbol{h}^t = h^t, \boldsymbol{\mu}^t = \mu^{*mk}, \boldsymbol{o}^t = o^t) \,\right] \\
&= \mathcal{E}(b|\boldsymbol{h}^t = h^t) \\
&- \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{o^t} \overset{\text{IDPS}}{\mathbb{P}}(\boldsymbol{\mu}^t = \mu^{*mk}, \boldsymbol{o}^t = o^t|\boldsymbol{h}^t = h^t) \mathcal{E}(b|\boldsymbol{h}^t = h^t, \boldsymbol{\mu}^t = \mu^{*mk}, \boldsymbol{o}^t = o^t).
\end{aligned}
$$

Thus, taking an outer expectation of $f^*(h^t) \bullet g(h^t)$ yields

$$
\begin{aligned}
&\overset{\text{IDPS}}{\underset{\boldsymbol{h}^t}{\mathbb{E}}}(f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)) \\
&= \sum_{h^t} \overset{\text{IDPS}}{\mathbb{P}}(\boldsymbol{h}^t = h^t) \Bigg[\mathcal{E}(b|\boldsymbol{h}^t = h^t) \\
&- \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{o^t} \overset{\text{IDPS}}{\mathbb{P}}(\boldsymbol{\mu}^t = \mu^{*mk}, \boldsymbol{o}^t = o^t|\boldsymbol{h}^t = h^t) \mathcal{E}(b|\boldsymbol{h}^t = h^t, \boldsymbol{\mu}^t = \mu^{*mk}, \boldsymbol{o}^t = o^t) \Bigg] \\
&= \mathcal{E}(b|\boldsymbol{h}^{t,\text{IDPS}}) - \sum_{h^t} \overset{\text{IDPS}}{\mathbb{P}}(\boldsymbol{h}^t = h^t) \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{o^t} \overset{\text{IDPS}}{\mathbb{P}}(\boldsymbol{\mu}^t = \mu^{*mk}, \boldsymbol{o}^t = o^t|\boldsymbol{h}^t = h^t) \\
&\times \mathcal{E}(b|\boldsymbol{h}^t = h^t, \boldsymbol{\mu}^t = \mu^{*mk}, \boldsymbol{o}^t = o^t) \\
&= \mathcal{E}(b|\boldsymbol{h}^{t,\text{IDPS}}) - \sum_{h^t} \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{o^t} \mathbb{P}(\boldsymbol{h}^{t,\text{IDPS}} = h^t, \boldsymbol{\mu}^{\boldsymbol{t},\text{IDPS}} = \mu^{*mk}, \boldsymbol{o}^t = o^t) \\
&\times \mathcal{E}(b|\boldsymbol{h}^{t,\text{IDPS}} = h^t, \boldsymbol{\mu}^{t,\text{IDPS}} = \mu^{*mk}, \boldsymbol{o}^t = o^t)
\end{aligned}
$$

$$= \mathcal{E}(b|\boldsymbol{h}^{t,\text{IDPS}}) - \mathcal{E}(b|\boldsymbol{h}^{t+1,\text{IDPS}}).$$

Hence, the lemma holds.

□

The following proposition proves inequality "(c)" in (2.20).

Proposition 2.2.5. *The cumulative expected information gain due to sampling policies according to pmfs indicated by IDPS over all episodes $t = 1 : T$ is bounded above by the Shannon entropy of the initial belief pmf. That is, $\sum_{t=1}^{T} \underset{\boldsymbol{h}^t}{\overset{IDPS}{\mathbb{E}}} (f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)) \leq \mathcal{E}(b)$.*

Proof. Similarly to Proposition 1.4.5, we have

$$\begin{aligned}\sum_{t=1}^{T} \underset{\boldsymbol{h}^t}{\overset{\text{IDPS}}{\mathbb{E}}} (f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)) &= \sum_{t=1}^{T} \mathcal{E}(b|\boldsymbol{h}^{t,\text{IDPS}}) - \mathcal{E}(b|\boldsymbol{h}^{t+1,\text{IDPS}}) \\ &= \mathcal{E}(b|\boldsymbol{h}^{1,\text{IDPS}}) - \mathcal{E}(b|\boldsymbol{h}^{T+1,\text{IDPS}}) \\ &\leq \mathcal{E}(b|\boldsymbol{h}^{1,\text{IDPS}}) = \mathcal{E}(b).\end{aligned}$$

□

Proposition 2.2.6 below proves inequality "(d)" in (2.20).

Proposition 2.2.6. *The expected value of the information ratio of an optimal solution $f^*(h^t)$ to (2.18) is bounded above by N^2MK. That is, $\underset{\boldsymbol{h}^t}{\overset{IDPS}{\mathbb{E}}} (\phi(f^*|\boldsymbol{h}^t)) \leq N^2MK$.*

Proof. Similarly to as in Section 1.4.2, we begin by showing inequality

$$N^2 g(i,j|h^t) \geq \sum_{m=1}^{M}\sum_{\ell=1}^{K} b(m,\ell|\boldsymbol{h}^t = h^t) \left[V^{ijm\ell} - \sum_{m'=1}^{M}\sum_{\ell'=1}^{K} b(m',\ell'|\boldsymbol{h}^t = h^t) V^{ijm'\ell'}\right]^2$$

holds as

$$g(i,j|h^t) = I(b|\boldsymbol{h}^t = h^t; \boldsymbol{o}^t|\boldsymbol{\mu}^t = \mu^{*ij}) \tag{2.23}$$

$$= \sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k|\boldsymbol{h}^t = h) D_{\mathrm{KL}}\left(\mathbb{P}\left(\boldsymbol{o}^t|\boldsymbol{\mu}^t = \mu^{*ij}, (m,k) \text{ is the true POMDP}\right) || \, \mathbb{P}\left(\boldsymbol{o}^t|\boldsymbol{\mu}^t = \mu^{*ij}\right)\right)$$

$$\geq \frac{2}{N^2}\sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k|\boldsymbol{h}^t = h^t)\times$$

$$\left[\mathbb{E}\left[U(\boldsymbol{o}^t)|\boldsymbol{\mu}^t = \mu^{*ij}, (m,k)\text{is the true POMDP}\right] - \mathbb{E}\left[U(\boldsymbol{o}^t)|\boldsymbol{\mu}^t = \mu^{*ij}\right]\right]^2$$

$$\geq \frac{1}{N^2}\sum_{m=1}^{M}\sum_{k=1}^{K} b(m,k|\boldsymbol{h}^t = h^t)\left[V^{ijmk} - \sum_{m'=1}^{M}\sum_{k'=1}^{K} b(m',k'|\boldsymbol{h}^t = h^t) V^{ijm'k'}\right]^2. \tag{2.24}$$

An argument similar to that ending with (1.4) can be made to show Assumption 1.4.1 implies the difference in the total rewards earned in one episode by any two policies is at most N. Then, Lemma 1.4.6 can be applied with $\psi \equiv U$, where $U(\mathbf{o}^t)$ is the total reward earned in the t^{th} episode given signal measurements $\mathbf{o}^t = (\mathbf{o}_2^t, \ldots, \mathbf{o}_{N+1}^2)$. Justifications for the remaining inequalities in (2.23-2.24) are similar to those of (1.82-1.85).

Next, we show the expected value of the information ratio due to sampling a policy according to PS is at most N^2MK. Given history $\boldsymbol{h}^t = h^t$, define probability vector $f(h^t) \in \mathcal{P}(MK)$ such that $f(m,k|h^t) \stackrel{\text{def}}{=} b(m,k|\boldsymbol{h}^t = h^t)$ for $m = 1, \cdots, M$ and $k = 1, \cdots, K$. Meaning, policy μ^{*mk} will be chosen with the same probability as the decision-maker's belief that the true MDP is the one with transition probability matrix P^m and measurement matrix Ξ^k. Observe that $(f(h^t) \bullet \Delta(h^t))^2 \leq N^2MK(f(h^t) \bullet g(h^t))$ as

$$[f(h^t) \bullet \Delta(h^t)]^2 = \left[\sum_{m=1}^{M}\sum_{k=1}^{K} f(m,k|h^t)\Delta(m,k|h^t)\right]^2 \tag{2.25}$$

$$= \left[\sum_{m=1}^{M}\sum_{k=1}^{K} f(m,k|h^t)\sum_{i=1}^{M}\sum_{j=1}^{K} f(i,j|h^t)(V^{*ij} - V^{mkij})\right]^2$$

$$= \left[\sum_{m=1}^{M}\sum_{k=1}^{K} f(m,k|h^t) \cdot 1 \sum_{i=1}^{M}\sum_{j=1}^{K} f(i,j|h^t)V^{*ij} - \sum_{m=1}^{M}\sum_{k=1}^{K} f(m,k|h^t)\sum_{i=1}^{M}\sum_{j=1}^{K} f(i,j|h^t)V^{mkij}\right]^2$$

$$
\begin{aligned}
&= \left[\sum_{i=1}^{M}\sum_{j=1}^{K} f(i,j|h^t)V^{*ij} - \sum_{m=1}^{M}\sum_{k=1}^{K} f(m,k|h^t)\sum_{i=1}^{M}\sum_{j=1}^{K} f(i,j|h^t)V^{mkij}\right]^2 \\
&= \left[\sum_{m=1}^{M}\sum_{k=1}^{K} f(m,k|h^t)V^{*mk} - \sum_{m=1}^{M}\sum_{k=1}^{K} f(m,k|h^t)\sum_{i=1}^{M}\sum_{j=1}^{K} f(i,j|h^t)V^{mkij}\right]^2 \\
&= \left[\sum_{m=1}^{M}\sum_{k=1}^{K} f(m,k|h^t)\left(V^{*mk} - \sum_{i=1}^{M}\sum_{j=1}^{K} f(i,j|h^t)V^{mkij}\right)\right]^2 \\
&\leq \sum_{m=1}^{M}\sum_{k=1}^{K}(1)^2\sum_{m=1}^{M}\sum_{k=1}^{K}\left(f(m,k|h^t)\left(V^{*mk} - \sum_{i=1}^{M}\sum_{j=1}^{K} f(i,j|h^t)V^{mkij}\right)\right)^2 \\
&= MK\sum_{m=1}^{M}\sum_{k=1}^{K}\left(f(m,k|h^t)\right)^2\left(V^{*mk} - \sum_{i=1}^{M}\sum_{j=1}^{K} f(i,j|h^t)V^{mkij}\right)^2 \\
&\leq MK\sum_{m=1}^{M}\sum_{k=1}^{K} f(m,k|h^t)\sum_{m'=1}^{M}\sum_{k'=1}^{K} f(m',k'|h^t)\left(V^{mkm'k'} - \sum_{i=1}^{M}\sum_{j=1}^{K} f(i,j|h^t)V^{mkij}\right)^2 \\
&\leq N^2MK\sum_{m=1}^{M}\sum_{k=1}^{K} f(m,k|h^t)g(m,k|h^t) = N^2MK(f(h^t)\bullet g(h^t)). \qquad (2.26)
\end{aligned}
$$

Collectively, (2.25-2.26) show $(f(h^t)\bullet\Delta(h^t))^2 \leq MK(f(h^t)\bullet g(h^t))$. Rearranging quantities and recalling the definition of ϕ yields $\phi(f(h^t)|h^t) = (f(h^t)\bullet\Delta(h^t))^2/(f(h^t)\bullet g(h^t)) \leq N^2MK$. Since $f^*(h^t)$ minimizes the information ratio, we have that $\phi(f^*|h^t) \leq \phi(f(h^t)|h^t) \leq N^2MK$, which explains why $\underset{\boldsymbol{h}^t}{\mathbb{E}}^{\text{IDPS}}(\phi(f^*|\boldsymbol{h}^t)) \leq \underset{\boldsymbol{h}^t}{\mathbb{E}}^{\text{IDPS}}(N^2MK) = N^2MK$. □

Finally, Theorem 2.2.7 below combines the results of the previous propositions to prove the regret of IDPS is at most $N\sqrt{TMK\mathcal{E}(b)}$.

Theorem 2.2.7. *The regret of IDPS is bounded by $N\sqrt{TMK\mathcal{E}(b)}$, where N is the number of time-stages, T is the number of episodes, M is the number of possible transition probability matrices, K is the number of possible measurement matrices, and $\mathcal{E}(b)$ is the Shannon entropy of the initial belief pmf b. That is, Regret$(IDPS, T) \leq N\sqrt{TMK\mathcal{E}(b)}$.*

Proof. Following the outlining chain of inequalities provided by (2.19-2.20) and the justifi-

cation in Theorem 1.4.2, we have

$$\begin{aligned}
\text{Regret}(\text{IDPS}, T) &= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{IDPS}} (f^*(\boldsymbol{h}^t) \bullet \Delta(\boldsymbol{h}^t)) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{IDPS}} \left(\sqrt{\phi(f^*|\boldsymbol{h}^t) \, [f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)]} \right) \\
&\leq \sum_{t=1}^{T} \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{IDPS}} (\phi(f^*|\boldsymbol{h}^t))} \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{IDPS}} (f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t))} \\
&\leq \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{IDPS}} (\phi(f^*|\boldsymbol{h}^t))} \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{IDPS}} (f^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t))} \\
&\leq \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{IDPS}} (\phi(f^*|\boldsymbol{h}^t))} \sqrt{\mathcal{E}(b)} \leq \sqrt{TN^2MK} \sqrt{\mathcal{E}(b)} \\
&= N\sqrt{TMK\mathcal{E}(b)}.
\end{aligned}$$

□

Corollary 2.2.8 and 2.2.9 below follow from Theorem 2.2.7.

Corollary 2.2.8. *The regret of IDPS is bounded by* $N\sqrt{TMK\log(MK)}$*, where* N *is the number of time-stages,* T *is the number of episodes,* M *is the number of possible transition probability matrices, and* K *is the number of possible measurement matrices. That is,* $\textit{Regret}(\textit{IDPS}, T) \leq N\sqrt{TMK\log(MK)}$.

Proof. Similarly to Corollary 1.4.9, we have

$$\text{Regret}(\text{IDPS}, T) \leq N\sqrt{TMK\mathcal{E}(b)} \leq \sqrt{TMK\log(MK)}.$$

□

Corollary 2.2.9. *The following two statements are true.*

(a) Suppose the measurement matrix Ξ *is known, and* b *is a belief pmf over the set of* M *possible transition probability matrices* $\{P^1, \cdots, P^M\}$ *so that* $b(m)$ *is the decision-maker's*

belief that transition probability matrix P^m is the true one. Then, the regret of IDPS is at most $N\sqrt{TM\mathcal{E}(b)} \leq N\sqrt{TM\log(M)}$.

(b) Suppose the transition probability matrix P is known, and b is a belief pmf over the set of K possible measurement matrices $\{\Xi^1, \cdots, \Xi^K\}$ so that $b(k)$ is the decision-maker's belief that measurement matrix Ξ^k is the true one. Then, the regret of IDPS is at most $N\sqrt{TK\mathcal{E}(b)} \leq N\sqrt{TK\log(K)}$.

Finally, consider the following remark. Briefly, it states that a similar regret bound will hold if the reward matrix R is unknown along with the transition probability matrix P and observation matrix O.

Remark 2.2.10. *Consider the problem where in addition to incomplete information about P and Ξ, the decision-maker only knows that the reward matrix R is one of L possibilities $\{R^1, \cdots, R^L\}$. In this scenario, the decision-maker has an initial joint belief pmf $b_1(m, k, \ell)$ on the true P, Ξ, R combination. Signal and reward trajectories observed during episodes are used to maintain and update the the joint belief pmf $b_t(m, k, \ell)$ on the true P, Ξ, R combination. The episodic regret and information gain of a policy μ can be defined similarly as before for all $\mu^{*mk\ell}$, where $\mu^{*mk\ell}$ is an optimal policy to the POMDP with transition probability matrix P^m, observation matrix Ξ^k, and reward matrix R^ℓ. The decision variable in the resulting convex information ratio minimization problem will be a pmf $f \in \mathcal{P}(MKL)$ and the problem will be convex in such an f.*

The T-episode regret of IDPS is defined as

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{\ell=1}^{L} b_1(m,k,\ell)\left(TV^{*mk\ell} - \overset{IDPS}{\mathbb{E}}\left[\sum_{t=1}^{T} V^{\boldsymbol{\mu}^t mk\ell}\right]\right). \tag{2.27}$$

A chain of inequalities similar to (2.19-2.20) can be used to derive a regret bound of $N\sqrt{TMKL\mathcal{E}(b_1)} \leq N\sqrt{TMLK\log(MKL)}$. Finally, corollaries similar to Corollary 2.2.7 and 2.2.9 can be stated.

2.3 Conclusions

We formulated the problem of episodic Bayesian POMDPs when there is incomplete information about the transition probability and observation matrices. We then proposed a version of IDPS for this setting. The notions of episodic regret and information gain were extended so that a proof technique similar to that of Chapter 1 could be applied to derive regret bounds. An extension of this framework to the case of incomplete information about the reward matrices was described.

Chapter 3

HIERARCHICAL MARKOV DECISION PROCESSES WITH INCOMPLETE INFORMATION

3.1 Introduction

Complex problems in the broader statistical literature are often modeled via a hierarchical structure that combines simpler models. Such integration provides enhanced representation power and can facilitate model-fitting with less data than a flat model. Some problems are especially amenable to this approach owing to their inherent structure. Bayesian methods can exploit such structure to improve model selection and estimation. See, for instance, [5, 13, 56].

We consider an episodic Bayesian framework similar to Chapters 1 and 2 for addressing the problem of decision-making in MDPs with incomplete hierarchical information. For an example of such a setup, consider the following variation of the classical finite-stage stochastic inventory control problem. Here, the demand can be either Poisson or Binomial. If the demand is Poisson distributed, then its mean is one of, say, five possible values. In contrast, if the demand follows a Binomial distribution, then the parameter is one of, say, three possible values. Since the parameters of the model characterize the transition probability matrices of the associated MDP, the decision faces five MDPs if the demand model is Poisson and three MDPs if the demand model is Binomial. However, in this case, the decision-maker does not know which model (higher-level incomplete information) and which transition probability matrix (lower-level incomplete information) for the true model characterize the MDP.

In the episodic Bayesian setting, the decision-maker starts with an initial belief as to which model and which transition probability matrix for the model are the true ones. The decision-maker then interacts with the unknown true MDP in an online fashion over a series

of episodes. In the beginning of each episode, the decision-maker chooses a policy and implements it throughout all stages of the MDP. The states visited due to implementing the chosen policy enable the decision-maker to update its belief before the start of the next episode.

3.2 Problem formulation

Consider a hierarchical extension of the N-stage MDP from Section 1.3. In this case, the decision-maker knows that the true system-dynamics are characterized by one of I possible models indexed by $i = 1 : I$. For model i, there are M possible transition matrices $\{P^{i1}, \cdots, P^{iM}\}$. The decision-maker does not know which model (higher-level incomplete information) and which transition probability matrix for the model (lower-level incomplete information) characterize the true MDP.

In the Bayesian episodic setting, the decision-maker begins with an initial belief pmf $b_1 \in \mathcal{P}(I)$ on the true model and conditional belief pmfs $\beta_1^i \in \mathcal{P}(M)$ on the true transition probability matrix for model i, $i = 1 : I$. The decision-maker interacts with the same true but unknown MDP for a total of T episodes. At the beginning of episode t, the decision-maker chooses policy π^t to implement for all N stages. An initial state s_1^t is sampled from the pmf ρ_1 and the previously chosen policy is implemented. The states $s^t = (s_1^t : s_{N+1}^t)$ visited during the N stages are then used to update the decision-maker's belief pmf b_t over the model and the conditional belief pmfs $\beta_t^1 : \beta_t^I$ over the transition probability matrix for each model. The decision-maker wishes to maximize the expected total reward earned over all T episodes.

If we ignore the hierarchical structure of this problem formulation, then we can use a proof similar to that of Chapter 1 to obtain a regret bound of $N\sqrt{TIM\log(IM)}$. However, if we do this, then the minimization problem will have a IM-dimensional decision variable. We will show that exploiting the hierarchical structure of this problem leads to a more efficient computation.

In Section 3.3, we introduce a family of algorithms tailored for this setting as well as four

specific algorithms of interest. A general regret bound for any algorithm in this family is derived in Section 3.4.1. Two algorithm-specific bounds which scale nicely with the number of stages N, episodes T, models I, and transition probability matrices per model M are derived in Section 3.4.2. Finally, the four algorithms of interest introduced in Section 3.3 are compared computationally in Section 3.5.

3.3 Hierarchical algorithms

Let π^{*ij} denote the optimal policy if the true MDP is characterized by model i and transition probability matrix P^{ij} (henceforth referred to as the event $(i \downarrow j)$).

We call an algorithm *hierarchical* if it takes advantage of the hierarchical structure of the problem formulation. Specifically, at the start of the tth episode, the hierarchical algorithm ALGO = (U $\downarrow$ L) first selects model i according to pmf $\gamma_t \in \mathcal{P}(I)$ determined by rule U and then selects the jth transition probability for model i with probability $\alpha_t^i(j)$ where pmf $\alpha_t^i \in \mathcal{P}(M)$ is determined by rule L. Thus, the probability that the hierarchical algorithm ALGO = (U $\downarrow$ L) chooses policy π^{*ij} is $\mathbb{P}^{\text{ALGO}}(\pi^{*ij}) = \gamma_t(i)\alpha_t^i(j)$. We consider hierarchical algorithms.

The posterior sampling hierarchical algorithm (PS $\downarrow$ PS) outlined in Algorithm 3 below first samples model i from the upper-level posterior belief b_t and then chooses policy π^{*ij} by sampling j from the lower-level posterior belief β_t^i for model i; here, $\gamma_t = b_t$ and $\alpha_t^i = \beta_t^i$. A theoretical analysis of this algorithm is provided in Section 3.4.2.

To apply the information-directed policy sampling to this framework, we first define a lower-level information ratio and an upper-level information ratio. The former will be used to select a policy for each model and the latter will be used to select the model corresponding to the policy.

Let $q_t(i,j)$ be the decision-maker's belief that the true model is characterized by model k and the transition probability matrix P^{ij} at the beginning of episode t. The regret of policy π^{*ij} given that the true MDP is characterized by model i is defined as $\tilde{\Delta}_t^i(j) \stackrel{\text{def}}{=} \sum_{k=1}^{I}\sum_{\ell=1}^{M} q_t(k,\ell)(V^{*k\ell} - V^{ijk\ell})$. Thus, the regret of choosing a policy at the lower level

Algorithm 3 (PS $\downarrow$ PS) for episodic Bayesian MDPs with hierarchical incomplete information

1: Start with higher-level belief pmf $b_1 \in \mathcal{P}(I)$, and lower-level belief pmfs $\beta_1^i \in \mathcal{P}(M), i = 1:I$.
2: **for** episodes $t = 1:T$ **do**
3: Observe initial state s_1^t sampled from $\rho_1 \in \mathcal{P}(|S|)$.
4: Sample a model index i from $\{1, \cdots, I\}$ according to $b_t(i)$.
5: Execute π^t sampled from $\{\pi^{*i1}, \cdots, \pi^{*iM}\}$ according to β_t^i; observe states $s_2^t, \cdots, s_{N+1}^t$.
6: Update lower-level belief pmfs to β_{t+1}^i using

$$\beta_{t+1}^i(j) \propto \rho_1(s_1^t) p^{ij}(s_2^t|s_1^t, \pi_1^t(s_1^t)) \cdots p^{ij}(s_{N+1}^t|s_N^t, \pi_N^t(s_N^t)) \beta_t^i(j), \text{ for } i = 1:I. \tag{3.1}$$

7: Update higher-level belief pmf to b_{t+1} using

$$b_{t+1}(i) \propto b_t(i) \sum_{j=1}^{M} \rho_1(s_1^t) p^{ij}(s_2^t|s_1^t, \pi_1^t(s_1^t)) \cdots p^{ij}(s_{N+1}^t|s_N^t, \pi_N^t(s_N^t)) \beta_t^i(j). \tag{3.2}$$

8: **end for**

according to pmf α_t^i determined by rule L given that the true MDP is characterized by model i is $\Delta_t^L(i) \stackrel{\text{def}}{=} \sum_{j=1}^M \alpha_t^i(j) \tilde{\Delta}_t^i(j)$; $\Delta_t^L(i)$ is referred to as the lower-level regret for model i.

Analogously, as $\tilde{g}_t^i(j) \stackrel{\text{def}}{=} I(q_t; \boldsymbol{s}^t | \boldsymbol{\pi}^t = \pi^{*ij})$ is the information gain of policy π^{*ij} given that the true MDP is characterized by model i, we define the lower-level information gain as $g_t^L(i) = \sum_{j=1}^M \alpha_t^i(j) \tilde{g}_t^i(j)$; this is the information gain of choosing a policy at the lower level according to pmf α_t^i determined by rule L given that the true MDP is characterized by model i. Combining the lower-level episodic regret and information gain yields the following lower-level information ratio when the true MDP is characterized by model i:

$$\phi_t^i(w) = \frac{\left(w \bullet \tilde{\Delta}_t^i\right)^2}{w \bullet \tilde{g}_t^i} = \frac{\left(\sum_{j=1}^M w(j) \tilde{\Delta}_t^i(j)\right)^2}{\sum_{j=1}^M w(j) \tilde{g}_t^i(j)} = \frac{\left(\sum_{j=1}^M w(j) \sum_{k=1}^I \sum_{\ell=1}^M q_t(k,\ell)(V^{*k\ell} - V^{ijk\ell})\right)^2}{\sum_{j=1}^M w(j) I(q_t; \boldsymbol{s}^t | \boldsymbol{\pi}^t = \pi^{*ij})}.$$

At the upper level, the (upper-level) episodic regret of choosing a policy according to $\gamma_t \in \mathcal{P}(I)$ as determined by rule U in episode t is

$$\begin{aligned}\gamma_t \bullet \Delta_t^L &\stackrel{\text{def}}{=} \sum_{i=1}^{I} \gamma_t(i) \Delta_t^L(i) = \sum_{i=1}^{I} \gamma_t(i) \left[\sum_{j=1}^{M} \alpha_t^i(j) \tilde{\Delta}_t^i(j) \right] \\ &= \sum_{i=1}^{I} \gamma_t(i) \left[\sum_{j=1}^{M} \alpha_t^i(j) \sum_{k=1}^{I} \sum_{\ell=1}^{M} q_t(k,\ell)(V^{*k\ell} - V^{ijk\ell}) \right],\end{aligned}$$

while the (upper-level) information gain is

$$\gamma_t \bullet g_t^L \stackrel{\text{def}}{=} \sum_{i=1}^{I} \gamma_t(i) g_t^L(i) = \sum_{i=1}^{I} \gamma_t(i) \left[\sum_{j=1}^{M} \alpha_t^i(j) \tilde{g}_t^i(j) \right] = \sum_{i=1}^{I} \gamma_t(i) \left[\sum_{j=1}^{M} \alpha_t^i(j) I(q_t; \boldsymbol{s}^t | \boldsymbol{\pi}^t = \pi^{*ij}) \right].$$

Hence, the (upper-level) information ratio is

$$\psi_t^L(u) = \frac{\left(u \bullet \Delta_t^L\right)^2}{u \bullet g_t^L} = \frac{\left(\sum_{i=1}^{I} u(i) \Delta_t^L(i) \right)^2}{\sum_{i=1}^{I} u(i) g_t^L(i)} = \frac{\left(\sum_{i=1}^{I} u(i) \left[\sum_{j=1}^{M} \alpha_t^i(j) \tilde{\Delta}_t^i(j) \right] \right)^2}{\sum_{i=1}^{I} u(i) \left[\sum_{j=1}^{M} \alpha_t^i(j) \tilde{g}_t^i(j) \right]}.$$

Note that the quantities defined for the upper level have superscripts indicating that rule L was used at the lower level. This will become important in Section 3.4.2.

Algorithms within the information-directed framework make a decision by sampling from a pmf minimizing the information ratio. Thus, the hierarchical information-directed algorithm (IDPS $\downarrow$ IDPS) chooses an index $i = 1 : I$ from a pmf $u_t^* \in \mathcal{P}(I)$ minimizing the upper-level information ratio $\psi_t^{\text{IDPS}}(u)$ and index $j = 1 : M$ from a pmf $f_t^{*i} \in \mathcal{P}(\mathcal{M})$ minimizing the lower-level information ratio $\phi_t^i(w)$ as summarized in Algorithm 4 below. The superscript in the upper-level information ratio ψ_t^{IDPS} is defined using the lower-level information ratio for all models $i = 1 : L$. Hence, the solutions to these lower-level information ratios must be determined in order to solve the upper-level information ratio minimization problem; they are determined first in Algorithm 4.

Algorithm 4 (IDPS $\downarrow$ IDPS) for episodic Bayesian MDPs with hierarchical incomplete information

1: Start with higher-level belief pmf $b_1 \in \mathcal{P}(I)$, and lower-level belief pmfs $\beta_1^i \in \mathcal{P}(M), i = 1 : I$.
2: **for** episodes $t = 1 : T$ **do**
3: Let f_t^{*i} be solutions of convex lower-level information ratio minimization problems

$$f_t^{*i} \in \underset{w \in \mathcal{P}(M)}{\operatorname{argmin}} \frac{(w \bullet \tilde{\Delta}_t^i)^2}{w \bullet \tilde{g}_t^i}, i = 1 : I.$$

4: Let u_t^* be a solution of the convex ratio minimization problem

$$u_t^* \in \underset{u \in \mathcal{P}(I)}{\operatorname{argmin}} \frac{(u \bullet \Delta_t^{\text{IDPS}})^2}{u \bullet g_t^{\text{IDPS}}}.$$

5: Observe initial state s_1^t sampled from $\rho_1 \in \mathcal{P}(|S|)$.
6: Sample a model index i from $\{1, \cdots, I\}$ according to u_t^*.
7: Execute π^t sampled from $\{\pi^{*i1}, \cdots, \pi^{*iM}\}$ according to f_t^{*i}; observe states $s_2^t, \cdots, s_{N+1}^t$.
8: Update lower-level belief pmfs to β_{t+1}^i using formula (3.1), for $i = 1 : I$.
9: Update higher-level belief pmf to b_{t+1} using formula (3.2).
10: **end for**

Combinations of posterior sampling and information-directed sampling are also possible. Algorithm 5 outlines (IDPS ↓ PS) where decisions in the lower level are determined by PS and those in the upper level are determined by IDPS. Here, only the upper-level minimization problem needs to be solved.

Algorithm 5 (IDPS ↓ PS) for episodic Bayesian MDPs with hierarchical incomplete information

1: Start with higher-level belief pmf $b_1 \in \mathcal{P}(I)$, and lower-level belief pmfs $\beta_1^i \in \mathcal{P}(M)$, $i = 1 : I$.
2: **for** episodes $t = 1 : T$ **do**
3: Let u_t^* be a solution of the convex ratio minimization problem

$$u_t^* \in \underset{u \in \mathcal{P}(I)}{\operatorname{argmin}} \frac{(u \bullet \Delta_t^{\mathrm{PS}})^2}{u \bullet g_t^{\mathrm{PS}}}.$$

4: Observe initial state s_1^t sampled from $\rho_1 \in \mathcal{P}(|S|)$.
5: Sample a model index i from $\{1, \cdots, I\}$ according to u_t^*.
6: Execute π^t sampled from $\{\pi^{*i1}, \cdots, \pi^{*iM}\}$ according to β_t^i; observe states $s_2^t, \cdots, s_{N+1}^t$.
7: Update lower-level belief pmfs to β_{t+1}^i using formula (3.1), for $i = 1 : I$.
8: Update higher-level belief pmf to b_{t+1} using formula (3.2).
9: **end for**

Distinctly, (PS ↓ IDPS) in Algorithm 6 chooses model i based on the upper-level belief b_t and policy π^t by sampling from pmf f_t^{*i}. Since the upper-level decision does *not* depend on the lower-level information ratio, model i can be chosen first and then the lower-level information ratio ϕ_t^i can be minimized. This is different from (IDPS ↓ IDPS) in that only one lower-level information ratio has to be determined even though the information-directed framework is being used at the lower-level for both Algorithm 4 and 6.

Table 3.1 provides a clear comparison of the computation required by Algorithms 3-6. Theoretical analysis of Algorithms 3-6 is provided, to varying extent, in Section 3.4.

Algorithm 6 (PS ↓ IDPS) for episodic Bayesian MDPs with hierarchical incomplete information

Start with higher-level belief pmf $b_1 \in \mathcal{P}(I)$, and lower-level belief pmfs $\beta_1^i \in \mathcal{P}(M), i = 1 : I$.
for episodes $t = 1 : T$ **do**
    Observe initial state s_1^t sampled from $\rho_1 \in \mathcal{P}(|S|)$.
    Sample a model index i from $\{1, \cdots, I\}$ according to $b_t(i)$.
    Let f_t^{*i} be a solution of convex lower-level information ratio minimization problem

$$f_t^{*i} \in \underset{w \in \mathcal{P}(M)}{\operatorname{argmin}} \frac{(w \bullet \tilde{\Delta}_t^i)^2}{w \bullet \tilde{g}_t^i}.$$

    Execute π^t sampled from $\{\pi^{*i1}, \cdots, \pi^{*iM}\}$ according to f_t^{*i}; observe states $s_2^t, \cdots, s_{N+1}^t$.
    Update lower-level belief pmfs to β_{t+1}^i using formula (3.1), for $i = 1 : I$.
    Update higher-level belief pmf to b_{t+1} using formula (3.2).
end for

Algorithm	**Upper-level computation**	**Lower-level computation**
(PS ↓ PS)	Use the upper-level belief pmf.	Use the lower-level belief pmf.
(IDPS ↓ PS)	Use the solution to an I-dimensional min problem.	Use the lower-level belief pmfs.
(PS ↓ IDPS)	Use the upper-level belief pmf.	Use the solution to an M-dimensional min problem.
(IDPS ↓ IDPS)	Use the solution to an I-dimensional min problem.	Use the solutions of I-many M-dimensional min problems.

Table 3.1: A comparison of the computational requirements for several algorithms.

3.4 Theoretical results

Let π^t denote the policy executed by the hierarchical algorithm ALGO = (U ↓ L) in episode t, and $V^{\pi^t ij}$ denote the expected total reward earned by this policy over stages $n = 1 :$

$N+1$ of episode t given that the initial state was sampled according to pmf ρ_1 given event $(i \downarrow j)$. Thus, the expected total reward earned by ALGO = (U $\downarrow$ L) in T episodes equals $\overset{\text{ALGO}}{\mathbb{E}}\left[\sum_{t=1}^{T} V^{\boldsymbol{\pi}^t ij}\right]$. The randomness in π^t arises from several sources. Pmfs α_t^i and γ_t depend (through the posteriors b_t and β_t^i, for $i = 1 : I$) on the history $h^t = (\pi^1, s^1, \cdots \pi^{t-1}, s^{t-1})$ consisting of previously chosen policies and resulting state trajectories in episodes $1 : t-1$. Model index i is sampled from pmf γ_t and then the policy for episode t is sampled according to pmf α_t^i. Recall that V^{*ij} is the optimal expected total reward earned by the (optimal) policy π^{*ij} in stages $n = 1 : N+1$ given event $(i \downarrow j)$. Thus, if the decision-maker knew that the event was $(i \downarrow j)$, it would execute policy π^{*ij} in each episode and earn the optimal reward of TV^{*ij}. That is, the regret of ALGO would be $TV^{*ij} - \overset{\text{ALGO}}{\mathbb{E}}\left[\sum_{t=1}^{T} V^{\boldsymbol{\pi}^t ij}\right]$. Since $b_1(i)\beta_1^i(j)$ denotes the probability of event $(i \downarrow j)$, the regret of ALGO is given by

$$\text{Regret}(\text{ALGO}, T) \overset{\text{def}}{=} \sum_{i=1}^{I}\sum_{j=1}^{M} b_1(i)\beta_1^i(j)\left[TV^{*ij} - \overset{\text{ALGO}}{\mathbb{E}}\left(\sum_{t=1}^{T} V^{\boldsymbol{\pi}^t ij}\right)\right]. \tag{3.3}$$

To bound the regret, we need to be careful with the expectation. To do so, we introduce notation where the history is explicit. Specifically, we alter the notation so that quantities with a subscript t have a history h^t. For instance, we change the decision-maker's belief $\beta_t(i)$ that the true model is model i at the beginning of episode t with $\beta(i|h^t)$. Here, $\beta(i|h^t)$ is the decision-maker's belief that i is the true model given history h^t. This level of detail is not needed to implement the algorithm but is is necessary for the regret analysis. We change $\beta_t(\cdot), \beta_t^i(\cdot), \gamma_t(\cdot), \alpha_t^i(\cdot), \tilde{\Delta}_t^i(\cdot), \Delta_t^L(\cdot), \tilde{g}_t^i(\cdot), g_t^L(\cdot), \phi_t^i(\cdot), \psi_t^L(\cdot)$ to $\beta(\cdot|h^t), \beta^i(\cdot|h^t), \gamma(\cdot|h^t), \alpha^i(\cdot|h^t)$, $\tilde{\Delta}^i(\cdot|h^t), \Delta^L(\cdot|h^t), \tilde{g}^i(\cdot|h^t), g^L(\cdot|h^t), \phi^i(\cdot|h^t), \psi^L(\cdot|h^t)$, respectively.

Table 3.2 summarizes the theoretical results.

3.4.1 Regret analysis for a general hierarchical algorithm

A general regret bound in terms of the cumulative expected information ratio can be derived for *any* hierarchical algorithm as demonstrated by Theorem 3.4.1.

Algorithm	Upper bound
(U $\downarrow$ L)	$\sqrt{\mathcal{E}(q_1)\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{(\mathrm{U}\downarrow\mathrm{L})}(\psi^{\mathrm{L}}(\gamma\|\boldsymbol{h}^t))}$
(PS $\downarrow$ PS)	$N\sqrt{TIM\mathcal{E}(q_1)} \leq N\sqrt{TIM\log(IM)}$
(IDPS $\downarrow$ PS)	$N\sqrt{TIM\mathcal{E}(q_1)} \leq N\sqrt{TIM\log(IM)}$

Table 3.2: Summary of theoretical results.

Theorem 3.4.1. *An upper bound on the regret of hierarchical algorithm ALGO = (U $\downarrow$ L) is*

$$\mathit{Regret}(ALGO,T) \leq \sqrt{\mathcal{E}(q_1)\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{ALGO}(\psi^{L}(\gamma|\boldsymbol{h}^t))} \leq \sqrt{\log(IM)\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{ALGO}(\psi^{L}(\gamma|\boldsymbol{h}^t))}.$$

Proof. We have

$$\mathrm{Regret}(\mathrm{ALGO},T) = \sum_{t=1}^{T}\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\mathrm{ALGO}}(\gamma(\boldsymbol{h}^t)\bullet\Delta(\boldsymbol{h}^t)) \tag{3.4}$$

$$= \sum_{t=1}^{T}\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\mathrm{ALGO}}\left(\sqrt{\psi^{L}(\gamma|\boldsymbol{h}^t)\,[\gamma(\boldsymbol{h}^t)\bullet g(\boldsymbol{h}^t)]}\right) \tag{3.5}$$

$$\leq \sum_{t=1}^{T}\sqrt{\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\mathrm{ALGO}}(\psi^{L}(\gamma|\boldsymbol{h}^t))}\sqrt{\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\mathrm{ALGO}}(\gamma(\boldsymbol{h}^t)\bullet g(\boldsymbol{h}^t))} \tag{3.6}$$

$$\leq \sqrt{\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\mathrm{ALGO}}(\psi^{L}(\gamma|\boldsymbol{h}^t))}\sqrt{\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\mathrm{ALGO}}(\gamma(\boldsymbol{h}^t)\bullet g(\boldsymbol{h}^t))} \tag{3.7}$$

$$\leq \sqrt{\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\mathrm{ALGO}}(\psi^{L}(\gamma|\boldsymbol{h}^t))\mathcal{E}(q_1)}. \tag{3.8}$$

The equality in (3.4) holds by Proposition 3.4.2, and the equality in (3.5) holds by the

definition of $\psi^L(h^t)$. Hölder's inequality justifies the equality in (3.6), and the Cauchy-Schwarz inequality justifies the one in (3.7). Finally, (3.8) is justified by Proposition 3.4.4.

Since the cardinality of q_1's sample space is IM, a basic property of Shannon entropy implies inequality "(a)" in

$$\text{Regret}(\text{ALGO}, T) \leq \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{ALGO}} (\psi^L(\gamma|\boldsymbol{h}^t))\mathcal{E}(q_1)} \overset{(a)}{\leq} \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{ALGO}} (\psi^L(\gamma|\boldsymbol{h}^t)) \log(IM)}. \ \square$$

Theorem 3.4.1 depends on Proposition 3.4.2 and Proposition 3.4.7, which we prove now with the help of Lemma 3.4.3.

Proposition 3.4.2. *The regret of ALGO = (U ↓ L) can be split additively over episodes. That is, Regret(ALGO, T) = $\sum_{t=1}^{T} \mathbb{E}^{ALGO}(\gamma(\boldsymbol{h}^t) \bullet \Delta^L(\boldsymbol{h}^t))$.*

Proof. By induction on the total number of episodes t, we assume the claim holds for t and show it holds for $t+1$. The regret can be expanded as

$$\begin{aligned}
\text{Regret}(\text{ALGO}, t+1) &= \sum_{i=1}^{I}\sum_{j=1}^{M} b(i)\beta^i(j) \left[(t+1)V^{*ij} - \mathop{\mathbb{E}}^{\text{ALGO}} \left(\sum_{k=1}^{t+1} V^{\boldsymbol{\pi}^k ij}\right)\right] \\
&= \sum_{i=1}^{I}\sum_{j=1}^{M} b(i)\beta^i(j) \left[(t+1)V^{*ij} - \sum_{\pi^1,\ldots,\pi^{t+1}} \mathop{\mathbb{P}}^{\text{ALGO}}(\pi^1,\ldots,\pi^{t+1}|(i \downarrow j)) \sum_{k=1}^{t+1} V^{\pi^k ij}\right] \\
&= \sum_{i=1}^{I}\sum_{j=1}^{M} b(i)\beta^i(j) \left[tV^{*ij} - \sum_{\pi^1,\ldots,\pi^{t+1}} \mathop{\mathbb{P}}^{\text{ALGO}}(\pi^1,\ldots,\pi^{t+1}|(i \downarrow j)) \sum_{k=1}^{t} V^{\pi^k ij}\right] \qquad (3.9) \\
&\quad + \sum_{i=1}^{I}\sum_{j=1}^{M} b(i)\beta^i(j) \left[V^{*ij} - \sum_{\pi^1,\ldots,\pi^{t+1}} \mathop{\mathbb{P}}^{\text{ALGO}}(\pi^1,\ldots,\pi^{t+1}|(i \downarrow j)) V^{\pi^{t+1} ij}\right].
\end{aligned}$$

The quantity in (3.9) can be rewritten using the inductive hypothesis as

$$\sum_{i=1}^{I}\sum_{j=1}^{M} b(i)\beta^i(j) \left[tV^{*ij} - \sum_{\pi^1,\ldots,\pi^{t+1}} \mathop{\mathbb{P}}^{\text{ALGO}}(\pi^1,\ldots,\pi^{t+1}|(i \downarrow j)) \sum_{k=1}^{t} V^{\pi^k ij}\right]$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{M} b(i)\beta^i(j)\,\Big[tV^{*ij}$$

$$- \sum_{\pi^1,\ldots,\pi^t} {}^{\text{ALGO}}\mathbb{P}(\pi^1,\ldots,\pi^t|(i\downarrow j)) \sum_{\pi^{t+1}} {}^{\text{ALGO}}\mathbb{P}(\pi^{t+1}|\pi^1,\ldots,\pi^t,(i\downarrow j)) \sum_{k=1}^{t} V^{\pi^k ij}\Big]$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{M} b(i)\beta^i(j)\left[tV^{*ij} - \sum_{\pi^1,\ldots,\pi^t} {}^{\text{ALGO}}\mathbb{P}(\pi^1,\ldots,\pi^t|(i\downarrow j)) \sum_{k=1}^{t} V^{\pi^k ij}\right]$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{M} b(i)\beta^i(j)\left[tV^{*ij} - {}^{\text{ALGO}}\mathbb{E}\left(\sum_{k=1}^{t} V^{\boldsymbol{\pi}^k ij}\right)\right]$$

$$= \sum_{k=1}^{t} {}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^k)\bullet\Delta^L(\boldsymbol{h}^k)). \tag{3.10}$$

Substituting (3.10) for (3.9) yields

$$\text{Regret}(\text{ALGO}, t+1) = \sum_{k=1}^{t} {}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^k)\bullet\Delta^L(\boldsymbol{h}^k))$$

$$+ \sum_{i=1}^{I}\sum_{j=1}^{M} b(i)\beta^i(j)\left[V^{*ij} - \sum_{\pi^1,\ldots,\pi^{t+1}} {}^{\text{ALGO}}\mathbb{P}(\pi^1,\ldots,\pi^{t+1}|(i\downarrow j))V^{\pi^{t+1}ij}\right]$$

$$= \sum_{k=1}^{t} {}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^k)\bullet\Delta^L(\boldsymbol{h}^k))$$

$$+ \sum_{i=1}^{I}\sum_{j=1}^{M} b(i)\beta^i(j)\left[V^{*ij} - \sum_{\pi^1,s^1,\ldots,s^t,\pi^{t+1}} {}^{\text{ALGO}}\mathbb{P}(\pi^1,s^1,\ldots,s^t,\pi^{t+1}|(i\downarrow j))V^{\pi^{t+1}ij}\right]$$

$$= \sum_{k=1}^{t} {}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^k)\bullet\Delta^L(\boldsymbol{h}^k))$$

$$+ \sum_{i=1}^{I}\sum_{j=1}^{M} b(i)\beta^i(j)\left[V^{*ij} - \sum_{h^{t+1},\pi^{t+1}} {}^{\text{ALGO}}\mathbb{P}(h^{t+1},\pi^{t+1}|(i\downarrow j))V^{\pi^{t+1}ij}\right]$$

$$= \sum_{k=1}^{t} {}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^k)\bullet\Delta^L(\boldsymbol{h}^k))$$

$$
\begin{aligned}
&+\sum_{i=1}^{I}\sum_{j=1}^{M} b(i)\beta^i(j)\left[V^{*ij}-\sum_{h^{t+1}}{}^{\text{ALGO}}\mathbb{P}(h^{t+1}|(i\downarrow j))\sum_{\pi^{t+1}}{}^{\text{ALGO}}\mathbb{P}(\pi^{t+1}|h^{t+1},\cancel{(i\downarrow j)})V^{\pi^{t+1}ij}\right]\\
&=\sum_{k=1}^{t}{}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^k)\bullet\Delta^L(\boldsymbol{h}^k))\\
&\quad+\sum_{i=1}^{I}\sum_{j=1}^{M} b(i)\beta^i(j)\sum_{h^{t+1}}{}^{\text{ALGO}}\mathbb{P}(h^{t+1}|(i\downarrow j))\sum_{\pi^{t+1}}{}^{\text{ALGO}}\mathbb{P}(\pi^{t+1}|h^{t+1})\left[V^{*ij}-V^{\pi^{t+1}ij}\right]\\
&=\sum_{k=1}^{t}{}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^k)\bullet\Delta^L(\boldsymbol{h}^k))\\
&\quad+\sum_{h^{t+1}}{}^{\text{ALGO}}\mathbb{P}(h^{t+1})\sum_{i=1}^{I}\sum_{j=1}^{M} b(i|h^{t+1})\beta^i(j|h^{t+1})\sum_{\pi^{t+1}}{}^{\text{ALGO}}\mathbb{P}(\pi^{t+1}|h^{t+1})\left[V^{*ij}-V^{\pi^{t+1}ij}\right]\\
&=\sum_{k=1}^{t}{}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^k)\bullet\Delta^L(\boldsymbol{h}^k))\\
&\quad+\sum_{h^{t+1}}{}^{\text{ALGO}}\mathbb{P}(h^{t+1})\sum_{\pi^{t+1}}{}^{\text{ALGO}}\mathbb{P}(\pi^{t+1}|h^{t+1})\sum_{i=1}^{I}\sum_{j=1}^{M} b(i|h^{t+1})\beta^i(j|h^{t+1})\left[V^{*ij}-V^{\pi^{t+1}ij}\right]\\
&=\sum_{k=1}^{t}{}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^k)\bullet\Delta^L(\boldsymbol{h}^k))\\
&\quad+\sum_{h^{t+1}}{}^{\text{ALGO}}\mathbb{P}(h^{t+1})\sum_{k=1}^{I}\sum_{\ell=1}^{M}{}^{\text{ALGO}}\mathbb{P}(\pi^{*k\ell}|h^{t+1})\sum_{i=1}^{I}\sum_{j=1}^{M} b(i|h^{t+1})\beta^i(j|h^{t+1})\left[V^{*ij}-V^{k\ell ij}\right]\\
&=\sum_{k=1}^{t}{}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^k)\bullet\Delta^L(\boldsymbol{h}^k))\\
&\quad+\sum_{h^{t+1}}{}^{\text{ALGO}}\mathbb{P}(h^{t+1})\sum_{k=1}^{I}\sum_{\ell=1}^{M}\gamma(k|h^{t+1})\alpha^k(\ell|h^{t+1})\sum_{i=1}^{I}\sum_{j=1}^{M} b(i|h^{t+1})\beta^i(j|h^{t+1})\left[V^{*ij}-V^{k\ell ij}\right]
\end{aligned}
\tag{3.11}
$$

$$
\begin{aligned}
&=\sum_{k=1}^{t}{}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^k)\bullet\Delta^L(\boldsymbol{h}^k))+\sum_{h^{t+1}}{}^{\text{ALGO}}\mathbb{P}(h^{t+1})\sum_{k=1}^{I}\gamma(k|h^{t+1})\Delta^L(k|h^{t+1})\\
&=\sum_{k=1}^{t}{}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^k)\bullet\Delta^L(\boldsymbol{h}^k))+\sum_{h^{t+1}}{}^{\text{ALGO}}\mathbb{P}(h^{t+1})(\gamma(h^{t+1})\bullet\Delta^L(h^{t+1}))\\
&=\sum_{k=1}^{t}{}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^k)\bullet\Delta^L(\boldsymbol{h}^k))+{}^{\text{ALGO}}\mathbb{E}(\gamma(\boldsymbol{h}^{t+1})\bullet\Delta^L(\boldsymbol{h}^{t+1}))
\end{aligned}
$$

$$= \sum_{k=1}^{t+1} \overset{\text{ALGO}}{\mathbb{E}} (\gamma(\boldsymbol{h}^k) \bullet \Delta^L(\boldsymbol{h}^k))$$

where (3.11) uses $\mathbb{P}^{\text{ALGO}}(\pi^{*k\ell}|h^{t+1}) = \gamma(k|h^{t+1})\alpha^k(\ell|h^{t+1})$. □

Lemma 3.4.3 will be used by Proposition 3.4.4.

Lemma 3.4.3. *The expected information gain due to sampling according to ALGO can be expressed as the difference in entropies of q: $\mathbb{E}^{ALGO}[\gamma(h^t) \bullet g(h^t)] = \mathcal{E}(q|\boldsymbol{h}^t) - \mathcal{E}(q|\boldsymbol{h}^{t+1})$.*

Proof. Expanding $\gamma(h^t) \bullet g(h^t)$ yields

$$\gamma(h^t) \bullet g(h^t) = \sum_{i=1}^{I} \gamma(i|h^t) g(i|h^t) \tag{3.12}$$

$$= \sum_{i=1}^{I} \gamma(i|h^t) \sum_{j=1}^{M} \alpha^i(j) I(q|h^t; \boldsymbol{s}^t|\boldsymbol{\pi}^t = \pi^{*ij}) \tag{3.13}$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{M} \gamma(i|h^t) \alpha^i(j) I(q|h^t; \boldsymbol{s}^t|\boldsymbol{\pi}^t = \pi^{*ij}) \tag{3.14}$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{M} \overset{\text{ALGO}}{\mathbb{P}} (\boldsymbol{\pi}^t = \pi^{*ij}|h^t) I(q|h^t; \boldsymbol{s}^t|\boldsymbol{\pi}^t = \pi^{*ij}) \tag{3.15}$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{M} \overset{\text{ALGO}}{\mathbb{P}} (\boldsymbol{\pi}^t = \pi^{*ij}|h^t) \left[\mathcal{E}(q|h^t) - \mathcal{E}(q|h^t, \boldsymbol{\pi}^t = \pi^{*ij}, \boldsymbol{s}^t)\right] \tag{3.16}$$

$$= \left[\mathcal{E}(q|h^t) - \mathcal{E}(q|h^t, \boldsymbol{\pi}^t, \boldsymbol{s}^t)\right]. \tag{3.17}$$

The definition of $g(i|h^t)$ is used in (3.13). Line (3.15) uses the fact that $\mathbb{P}^{\text{ALGO}}(\boldsymbol{\pi}^t = \pi^{*ij}|h^t) = \gamma(i|h^t)\alpha^i(j)$; this is due to the way ALGO is defined. A well-known property of mutual information is used in (3.16).

Below, taking an outer expectation with respect to the history up to episode t yields the result as

$$\overset{\text{ALGO}}{\mathbb{E}} \left[\gamma(h^t) \bullet g(h^t)\right] = \sum_{h^t} \overset{\text{ALGO}}{\mathbb{P}} (h^t) \left[\mathcal{E}(q|h^t) - \mathcal{E}(q|h^t, \boldsymbol{\pi}^t, \boldsymbol{s}^t)\right]$$

$$= \mathcal{E}(q|\boldsymbol{h}^t) - \mathcal{E}(q|\boldsymbol{h}^t, \boldsymbol{\pi}^t, \boldsymbol{s}^t) = \mathcal{E}(q|\boldsymbol{h}^t) - \mathcal{E}(q|\boldsymbol{h}^{t+1}).$$

□

Finally, we prove Proposition 3.4.4 as the last step of proving Theorem 3.4.1.

Proposition 3.4.4. *The cumulative expected information gain of ALGO = (U ↓ L) is at most $\mathcal{E}(q)$. Meaning, $\sum_{t=1}^{T} \mathbb{E}^{ALGO}[\gamma_t^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)] \leq \mathcal{E}(q)$.*

Proof. Summing the result from Lemma 3.4.3 over all episodes yields

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}^{\text{ALGO}} \left[\gamma_t^*(\boldsymbol{h}^t) \bullet g(\boldsymbol{h}^t)\right] = \sum_{t=1}^{T} \mathcal{E}(q|\boldsymbol{h}^t) - \mathcal{E}(q|\boldsymbol{h}^{t+1}) = \mathcal{E}(q|\boldsymbol{h}^1) - \mathcal{E}(q|\boldsymbol{h}^{T+1}) \leq \mathcal{E}(q|\boldsymbol{h}^1) = \mathcal{E}(q).$$

□

3.4.2 Regret analysis for (PS ↓ PS) and (IDPS ↓ PS)

Section 3.4.1 provided a regret bound for a general hierarchical algorithm. Under Assumption 1.4.1, we can provide an upper bound of N^2IM on the expected information gain of (PS ↓ PS) and (IDPS ↓ PS), yielding an overall regret bound of $N\sqrt{TIM\log(IM)}$ for these two algorithms as noted in Theorem 3.4.5.

Theorem 3.4.5. *For ALGO = (PS ↓ PS) and ALGO = (IDPS ↓ PS), we have*

$$\textit{Regret}(ALGO, T) \leq \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{ALGO} (\psi^{PS}(\gamma|\boldsymbol{h}^t))\mathcal{E}(q_1)} \leq N\sqrt{TIM\log(IM)}.$$

Proof. For ALGO = (PS ↓ PS) or ALGO = (IDPS ↓ PS), we have

$$\text{Regret}(ALGO, T) \overset{(a)}{\leq} \sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{\text{ALGO}} (\psi^{\text{PS}}(\gamma|\boldsymbol{h}^t))\mathcal{E}(q_1)} \overset{(b)}{\leq} N\sqrt{TIM\mathcal{E}(q_1)} \overset{(c)}{\leq} N\sqrt{TIM\log(IM)}.$$

Above, inequality "(a)" follows from Theorem 3.4.5, inequality "(b)" follows from Proposition 3.4.7, and inequality "(c)" follows from a basic property of Shannon entropy since the sample space of q_1 has cardinality IM. □ □

We first prove Lemma 3.4.6 to assist in the proof of Proposition 3.4.7.

Lemma 3.4.6. *The following inequality is true:*

$$g_t(k) \geq \frac{1}{N^2} \sum_{\ell=1}^{M} \beta_t^k(\ell) \sum_{x=1}^{I} \sum_{y=1}^{M} b_t(x) \beta_t^x(y) \left(V^{k\ell xy} - \sum_{i=1}^{I} \sum_{j=1}^{M} V^{k\ell ij} \right)^2. \tag{3.18}$$

Proof. This claim uses Lemma 1.4.6. Namely, we have

$$g_t(k) = \sum_{\ell=1}^{M} \beta_t^k(\ell) I(q_t; \boldsymbol{s}^t | \boldsymbol{\pi}^t = \pi^{*k\ell}) \tag{3.19}$$

$$= \sum_{\ell=1}^{M} \beta_t^k(\ell) \sum_{x=1}^{I} \sum_{y=1}^{M} b_t(x) \beta_t^x(y) D_{KL} \left(\mathbb{P}(\boldsymbol{s}^t | \boldsymbol{\pi}^t = \pi^{*k\ell}, (x \downarrow y)) || \, \mathbb{P}(\boldsymbol{s}^t | \boldsymbol{\pi}^t = \pi^{*k\ell}) \right) \tag{3.20}$$

$$\geq \frac{1}{N^2} \sum_{\ell=1}^{M} \beta_t^k(\ell) \sum_{x=1}^{I} \sum_{y=1}^{M} b_t(x) \beta_t^x(y) \left(\mathbb{E}(U(\boldsymbol{s}^t) | \boldsymbol{\pi}^t = \pi^{*k\ell}, (x \downarrow y)) - \mathbb{E}(U(\boldsymbol{s}^t) | \boldsymbol{\pi}^t = \pi^{*k\ell}) \right)^2 \tag{3.21}$$

$$\geq \frac{1}{N^2} \sum_{\ell=1}^{M} \beta_t^k(\ell) \sum_{x=1}^{I} \sum_{y=1}^{M} b_t(x) \beta_t^x(y) \left(V^{k\ell xy} - \sum_{i=1}^{I} \sum_{j=1}^{M} V^{k\ell ij} \right)^2. \tag{3.22}$$

□

Finally, Proposition 3.4.7 is the last step in proving the regret bound that is specific to (PS ↓ PS) and (IDPS ↓ PS). Line (3.27) in this proposition allows for the proof technique to hold when the lower-level algorithm is PS. This equality does not hold when IDPS is used at the lower level. Moreover, this equality cannot be exchanged with an inequality when IDPS is used at the lower level. Meaning, the inequality in (3.27) is the reason why this proof technique does not work for (PS ↓ IDPS) or (IDPS ↓ IDPS).

Proposition 3.4.7. *For ALGO = (PS ↓ PS) and ALGO = (IDPS ↓ PS), we have* $\mathop{\mathbb{E}}_{\boldsymbol{h}^t}^{ALGO}(\psi^{PS}(u^*|\boldsymbol{h}^t)) \leq N^2IM$.

Proof. We begin by showing $(b_t \bullet \Delta_t^{\text{PS}})^2 \leq N^2IM(b_t \bullet g_t^{\text{PS}})$. Specifically, we have that

$$(b_t \bullet \Delta_t^{\text{PS}})^2 = \left(\sum_{i=1}^{I} b_t(i)\Delta_t^{\text{PS}}(i)\right)^2 \tag{3.23}$$

$$= \left(\sum_{i=1}^{I} b_t(i) \sum_{j=1}^{M} \beta_t^i(j) \sum_{k=1}^{I}\sum_{\ell=1}^{M} b_t(k)\beta_t^k(\ell)(V^{*k\ell} - V^{ijk\ell})\right)^2 \tag{3.24}$$

$$= \left(\sum_{i=1}^{I}\sum_{j=1}^{M} b_t(i)\beta_t^i(j) \sum_{k=1}^{I}\sum_{\ell=1}^{M} b_t(k)\beta_t^k(\ell)(V^{*k\ell} - V^{ijk\ell})\right)^2 \tag{3.25}$$

$$= \left(\left[\sum_{k=1}^{I}\sum_{\ell=1}^{M} b_t(k)\beta_t^k(\ell)V^{*k\ell}\right] - \left[\sum_{i=1}^{I}\sum_{j=1}^{M} b_t(i)\beta_t^i(j) \sum_{k=1}^{I}\sum_{\ell=1}^{M} b_t(k)\beta_t^k(\ell)V^{ijk\ell}\right]\right)^2 \tag{3.26}$$

$$= \left(\left[\sum_{k=1}^{I}\sum_{\ell=1}^{M} b_t(k)\beta_t^k(\ell)V^{*k\ell}\right] - \left[\sum_{i=1}^{I}\sum_{j=1}^{M} b_t(i)\beta_t^i(j) \sum_{k=1}^{I}\sum_{\ell=1}^{M} b_t(k)\beta_t^k(\ell)V^{k\ell ij}\right]\right)^2 \tag{3.27}$$

$$= \left(\sum_{k=1}^{I}\sum_{\ell=1}^{M} b_t(k)\beta_t^k(\ell)\left[V^{*k\ell} - \sum_{i=1}^{I}\sum_{j=1}^{M} b_t(i)\beta_t^i(j)V^{k\ell ij}\right]\right)^2 \tag{3.28}$$

$$\leq \sum_{m=1}^{I}\sum_{m'=1}^{M_m}(1)^2 \sum_{k=1}^{I}\sum_{\ell=1}^{M}\left(b_t(k)\beta_t^k(\ell)\left[V^{*k\ell} - \sum_{i=1}^{I}\sum_{j=1}^{M} b_t(i)\beta_t^i(j)V^{k\ell ij}\right]\right)^2 \tag{3.29}$$

$$= IM\sum_{k=1}^{I}\sum_{\ell=1}^{M}\left(b_t(k)\beta_t^k(\ell)\right)^2\left(\left[V^{*k\ell} - \sum_{i=1}^{I}\sum_{j=1}^{M} b_t(i)\beta_t^i(j)V^{k\ell ij}\right]\right)^2 \tag{3.30}$$

$$\leq IM\sum_{k=1}^{I}\sum_{\ell=1}^{M}\sum_{x=1}^{I}\sum_{y=1}^{M} b_t(k)\beta_t^k(\ell)b_t(x)\beta_t^x(y)\left(\left[V^{k\ell xy} - \sum_{i=1}^{I}\sum_{j=1}^{M} b_t(i)\beta_t^i(j)V^{k\ell ij}\right]\right)^2 \tag{3.31}$$

$$= IM\sum_{k=1}^{I} b_t(k)\sum_{\ell=1}^{M}\beta_t^\ell(k)\sum_{x=1}^{I}\sum_{y=1}^{M} b_t(x)\beta_t^x(y)\left(\left[V^{k\ell xy} - \sum_{i=1}^{I}\sum_{j=1}^{M} b_t(i)\beta_t^i(j)V^{k\ell ij}\right]\right)^2 \tag{3.32}$$

$$= N^2IM\sum_{k=1}^{I} b_t(k)g_t^{\text{PS}}(k) \tag{3.33}$$

$$= N^2IM(b_t \bullet g_t^{\text{PS}}). \tag{3.34}$$

The definition of $\Delta_t^{\text{PS}}(i)$ was used in (3.25). The important equality in (3.27) uses the fact that posterior sampling is used. The inequality in (3.29) applies the Cauchy-Schwarz inequality. More terms are included in (3.31) than in the previous line, justifying the inequality in (3.31). The inequality in (3.33) uses the result of Proposition 3.4.6. Thus, $(b_t \bullet \Delta_t^{\text{PS}})^2/(b_t \bullet g_t^{\text{PS}}) \leq N^2IM$. Equivalently, for particular history h^t, $\psi^{\text{PS}}(u^*|h^t) = (u^*(h^t) \bullet \Delta^{\text{PS}}(h^t))^2/(u^*(h^t) \bullet g^{\text{PS}}(h^t)) \leq N^2IM$. Taking an expectation with respect to the history yields the result: $\underset{\boldsymbol{h}^t}{\overset{(\text{PS} \downarrow \text{PS})}{\mathbb{E}}}(\psi^{\text{PS}}(u^*|\boldsymbol{h}^t)) \leq N^2IM$.

Since u_t^* minimizes information ratio ψ_t^{PS}, we also have that $(u_t^* \bullet \Delta_t^{\text{PS}})^2/(u_t^* \bullet g_t^{\text{PS}}) \leq N^2IM$. Similarly as above, this yields $\underset{\boldsymbol{h}^t}{\overset{(\text{IDPS} \downarrow \text{PS})}{\mathbb{E}}}(\psi^{\text{PS}}(u^*|\boldsymbol{h}^t)) \leq N^2IM$. □

3.4.3 Extension to the case of unknown rewards

Similarly to Chapter 1, the results of this chapter can be extended to the case where the reward matrices are uncertain.

Remark 3.4.8. *When the problem formulation is extended to include the case where model i not only has M possible transition probability matrices but also has rule L possible reward matrices $\{R^{i1}, \cdots, R^{iL}\}$, then the regret will be at most $N\sqrt{TIML\log(IML)}$.*

3.5 Computational results

Algorithms (IDPS ↓ PS), (PS ↓ PS), (PS ↓ IDPS), and (IDPS ↓ IDPS) are compared on an extension of the dynamic pricing application outlined in Section 1.5. Namely, the decision-maker does not know whether the demand is Poisson distributed or Binomial distributed. In both cases, the decision-maker has incomplete information about the price-demand parameters. Specifically, if the demand is Poisson distributed, then the decision-maker has incomplete information about the price-demand parameter (λ, α). If it is Binomial distributed, then the decision-maker has incomplete information about the price-demand parameter (B, α). For simplicity, we let $\lambda = B$ and assume there are three possible values of (λ, α). We sample the value of λ uniformly from the set $\{2, \ldots, 5\}$ and the value of α uniformly from the inter-

val $(0, 1)$. The remaining parameter values are chosen to be the same as those from Section 1.5. Specifically, throughout this section, we assume there is an initial inventory of $S = 4$, a set of possible prices $A = \{5, 10, 15, 20\}$, a holding cost of $h = 0.5$, and a penalty of $c = 1$ for the loss of excess demand. Further, we assume there are $N = 5$ time-stages and $T = 100$ episodes.

We display the results from 20 independent experiments. For each experiment, we generate three random parameter values for $(\lambda, \alpha) = (B, \alpha)$. For the first 10 experiments, we assume the true model is Poisson. We assume the true model is Binomial for the last 10 experiments. In all experiments, we assume the true parameter values at the lower level are given by the first sample of $(\lambda, \alpha) = (B, \alpha)$. We generate 50 independent replications for each experiment and consider the averaged results. The average cumulative regret at the end of episode $T = 100$ for each experiment and algorithm is displayed in Table 3.3 and Table 3.4. Table 3.3 (Table 3.4) pertains to the experiments when the true model is Poisson (Binomial).

Furthermore, for each experiment, we test whether the cumulative regret of (IDPS $\downarrow$ PS) is statistically lower than that of (PS $\downarrow$ PS). If the statistical experiment proved to be *in*significant, meaning that there was not enough evidence to conclude that (IDPS $\downarrow$ PS) outperformed (PS $\downarrow$ PS), then we included an asterisk in the corresponding row of Table 3.3 and Table 3.4.

Table 3.3 shows the performance of (IDPS $\downarrow$ PS) was statistically lower than that of (PS $\downarrow$ PS) for one experiment; this was not the case for the remaining nine experiments. Note that (PS $\downarrow$ PS) obtained the smallest average cumulative regret in six of the experiments, and (PS $\downarrow$ IDPS) obtained the smallest average cumulative regret in the remaining experiments. In all experiments when the true model was Poisson, it was best to use PS in the upper level of the hierarchy.

Table 3.4 shows the performance of (IDPS $\downarrow$ PS) was statistically lower than that of (PS $\downarrow$ PS) in ten out of ten experiments. Moreover, (IDPS $\downarrow$ IDPS) obtained the smallest average cumulative regret in six experiments, and (IDPS $\downarrow$ PS) obtained the smallest average

Table 3.3: Average cumulative regret when the true model is Poisson. The bolded price-demand parameters are the true parameters. Rows containing an asterisk indicate a lack of statistical evidence to show that (IDPS ↓ PS) outperformed (PS ↓ PS).

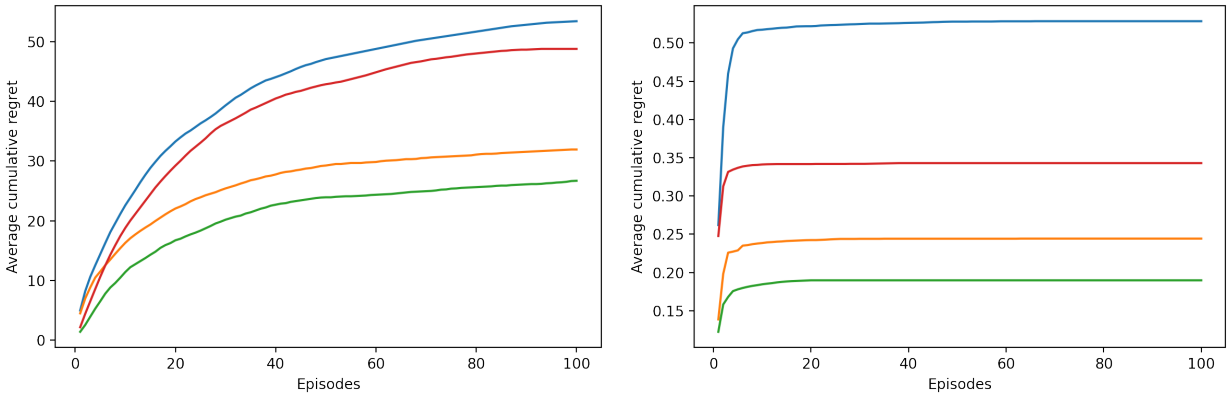
(λ, α)	(IDPS ↓ PS)	(PS ↓ PS)	(PS ↓ IDPS)	(IDPS ↓ IDPS)
(4, 0.07), (5, 0.29), (2, 0.57)*	53.467	31.946	26.701	48.822
(4, 0.9), (3, 0.97), (5, 0.61)*	0.529	0.244	0.19	0.343
(3, 0.06), (2, 0.65), (2, 0.41)	6.361	8.537	1.65	2.466
(2, 0.79), (5, 0.62), (4, 0.7)*	0.199	0.068	0.239	0.394
(3, 0.81), (2, 0.07), (2, 0.97)*	0.324	0.208	0.497	0.641
(3, 0.64), (5, 0.78), (2, 0.31)*	2.23	0.91	1.24	2.0
(5, 0.93), (5, 0.52), (4, 0.62)*	1.096	0.37	0.903	1.447
(4, 0.95), (2, 0.82), (5, 0.37)*	0.226	0.091	0.714	0.172
(3, 0.92), (4, 0.67), (5, 0.7)*	0.845	0.244	0.219	0.809
(2, 0.71), (2, 0.41), (4, 0.21)*	2.02	0.966	1.797	2.381

cumulative regret in the remaining experiments. In contrast to when the true model is Poisson, empirically, it was always best to utilize IDPS in the upper level of the hierarchy when the true model was Binomial.

For the first four experiments, Figure 3.1 plots the average cumulative regret for all four algorithms over all $T = 100$ episodes, when the true model is Poisson. Figure 3.2 shows the analogous plots when the true model is Binomial.

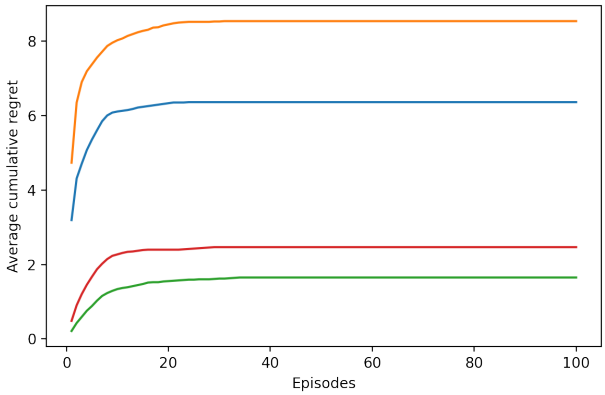
3.6 Conclusion

This chapter introduced a family of algorithms which exploit the structure of the problem. A general regret bound for all algorithms in this family was provided. Moreover, we introduced four hierarchical algorithms of interest in Section 3.3 and provided regret bounds specific to two of these algorithms in Section 3.4.2. All four algorithms were compared computationally in Section 3.5 and statistical tests were performed to compare the results of (IDPS ↓ PS) versus (PS ↓ PS). It was shown that (IDPS ↓ PS) outperformed (PS ↓ PS) when the price-demand function is Binomially distributed.

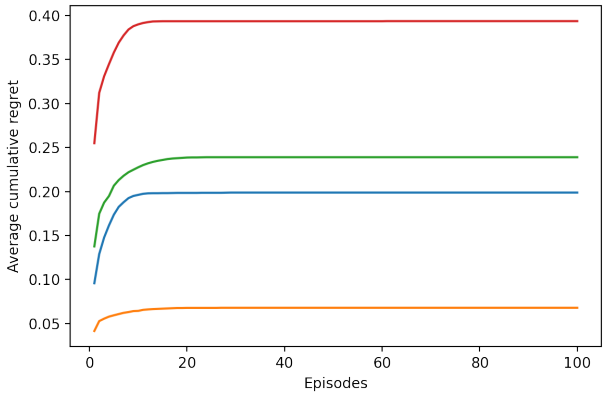


(a) $(\lambda, \alpha) = (4,\ 0.07)$

(b) $(\lambda, \alpha) = (4,\ 0.9)$

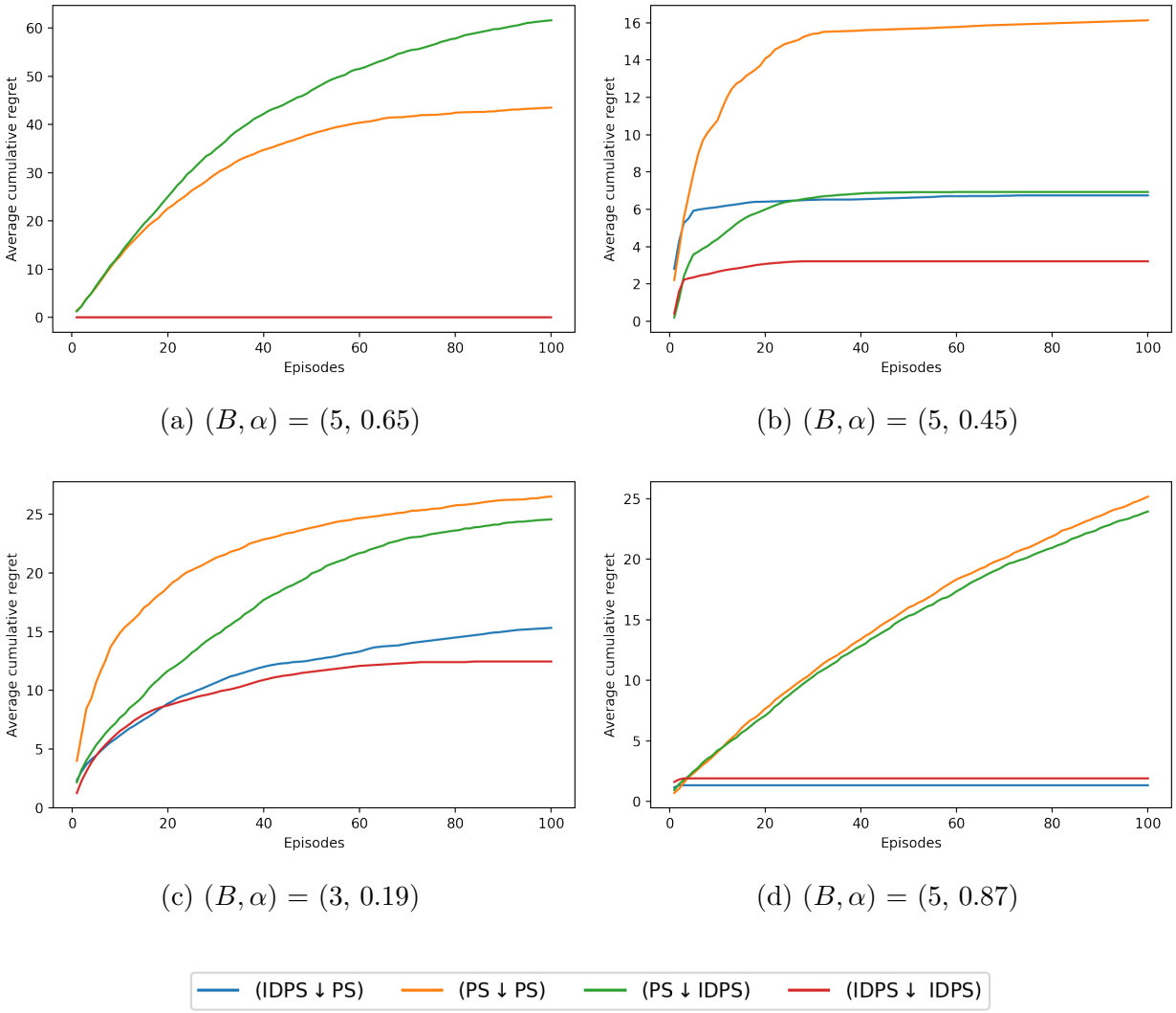


(c) $(\lambda, \alpha) = (3,\ 0.06)$



(d) $(\lambda, \alpha) = (2,\ 0.79)$

Figure 3.1: Average cumulative regret when the true model is Poisson. The true parameter values for (λ, α) are provided under each plot.



(a) $(B, \alpha) = (5,\ 0.65)$

(b) $(B, \alpha) = (5,\ 0.45)$

(c) $(B, \alpha) = (3,\ 0.19)$

(d) $(B, \alpha) = (5,\ 0.87)$

Figure 3.2: Average cumulative regret when the true model is Binomial. The true parameter values for (B, α) are provided under each plot.

Table 3.4: Average cumulative regret when the true model is Binomial. The bolded price-demand parameters are the true parameters. Note that no rows contain an asterisk, meaning that there was enough statistical evidence to reject the null hypothesis that (IDPS ↓ PS) underperformed against (PS ↓ PS).

(B, α)	(IDPS ↓ PS)	(PS ↓ PS)	(PS ↓ IDPS)	(IDPS ↓ IDPS)
(5, 0.9), (3, 0.04), (3, 0.05)	0.0	43.512	61.625	0.0
(3, 0.58), (3, 0.94), (4, 0.37)	6.744	16.136	6.928	3.211
(4, 0.71), (4, 0.03), (4, 0.07)	15.323	26.515	24.572	12.456
(5, 0.65), (5, 0.32), (2, 0.86)	1.323	25.181	23.948	1.895
(3, 0.22), (5, 0.22), (5, 0.25)	5.718	13.627	5.577	4.501
(3, 0.36), (3, 0.42), (4, 0.07)	1.284	14.382	8.289	1.811
(4, 0.44), (2, 0.32), (4, 0.85)	2.892	45.746	33.21	3.176
(3, 0.71), (4, 0.45), (3, 0.59)	6.391	13.76	6.448	3.965
(2, 0.61), (5, 0.35), (2, 0.3)	4.337	47.418	49.816	7.065
(2, 0.4), (5, 0.05), (4, 0.47)	4.48	15.239	4.344	4.145

BIBLIOGRAPHY

[1] P Afeche and B Atta. Bayesian dynamic pricing in queueing systems with unknown delay cost characteristics. *Manufacturing & Service Operations Management*, 15(2):292–304, 2013.

[2] S Agrawal and N Goyal. Further optimal regret bounds for Thompson Sampling. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pages 99–107, Scottsdale, AZ, USA, 2013.

[3] V Ahuja and J R Birge. Response-adaptive designs for clinical trials: Simultaneous learning from multiple patients. *European Journal of Operational Research*, 248:619–633, 2016.

[4] V Ahuja and J R Birge. An approximation approach for response-adaptive clinical trial design. *INFORMS Journal on Computing*, 32(4):877–894, 2020.

[5] G M Allenby and P E Rossi. Hierarchical Bayes models: a practitioners guide. In R Grover and M Vriens, editors, *The handbook of marketing research*. Sage, Newbury Park, CA, USA, 2005.

[6] P Auer, N Cesa-Bianchi, and P Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[7] D Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, Nashua, NH, USA, 3rd edition, 2005.

[8] S Boyd and L Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 1st edition, 2004.

[9] R I Brafman and M Tennenholtz. R-max — a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2003.

[10] E Brunskill and L Li. Sample complexity of multi-task reinforcement learning. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 122–131, 2013.

[11] P Buchholz and D Scheftelowitsch. Computation of weighted sums of rewards for concurrent mdps. *Mathematical Methods of Operations Research*, 89(1):1–42, 2019.

[12] O Chapelle and L Li. An empirical evaluation of Thompson Sampling. In *Proceedings of the 25th Conference on Neural Information Processing Systems*, pages 2249–2257, Granada, Spain, 2011.

[13] H Chipman and R E McCulloch. Hierarchical priors for Bayesian CART shrinkage. *Statistics and Computing*, 10(1):17–24, 2000.

[14] R Dearden, N Friedman, and D Andre. Model based Bayesian exploration. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 150–159, Stockholm, Sweden, 1999.

[15] E Delage and S Mannor. Percentile optimization for markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203–213, 2010.

[16] A V den Boer. Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, 20(1):1–18, 2015.

[17] M O Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massachusetts Amherst, 2002.

[18] P Frazier, W Powell, and S Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

[19] A Ghate. Optimal minimum bids and inventory scrapping in sequential, single-unit, vickrey auctions with demand learning. *European Journal of Operational Research*, 245(2):555–570, 2015.

[20] M Ghavamzadeh, S Mannor, J Pineau, and A Tamar. Bayesian Reinforcement Learning: A Survey. *Foundations and Trends in Machine Learning*, 8(5-6):359–492, 2015.

[21] J C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

[22] J C Gittins, K Glazebrook, and R Weber. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, Hoboken, NJ, USA, 2011.

[23] A Gopalan and S Mannor. Thompson Sampling for learning parameterized Markov decision processes. In *Proceedings of Machine Learning Research*, volume 40, pages 1–38, 2015.

[24] R M Gray. *Entropy and Information Theory*. Springer, New York, NY, USA, 2011.

[25] A Guez, D Silver, and P Dayan. Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Advances in Neural Information Processing Systems*, pages 1025–1033, Lake Tahoe, CA, USA, 2012.

[26] A Hallak, D Di Castro, and S Mannor. Contextual Markov decision processes. `https://arxiv.org/pdf/1502.02259.pdf`, 2015.

[27] Jochen Hardt. A new questionnaire for measuring quality of life - the Stark QoL. *Health and quality of life outcomes*, 13(1):174–174, 2015.

[28] J M Harrison, N B Keskin, and A Zeevi. Bayesian Dynamic Pricing Policies: Learning and Earning Under a Binary Prior Distribution. *Management Science*, 58(3):570–586, 2012.

[29] G N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[30] T Jaksch, R Ortner, and P Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

[31] C Jin, Z Allen-Zhu, S Bubeck, and M I Jordan. Is Q-learning provably efficient? In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, Montreal, Canada, 2018.

[32] E Kauffmann, N Korda, and R Munos. Thompson Sampling: an asymptotically optimal finite-time analysis. In N H Bshouty, G Stoltz, N Vayatis, and T Zeugmann, editors, *Lecture Notes in Computer Science*, volume 7568, pages 199–213, Berlin/Heidelberg, Germany, 2012. Springer.

[33] M Kearns and S Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

[34] J Kirschner and A Krause. Information directed sampling and bandits with heteroscedastic noise. In *Proc. International Conference on Learning Theory (COLT)*, July 2018.

[35] J Kirschner, T Lattimore, and A Krause. Information directed sampling for linear partial monitoring. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 2328–2369, 09-12 Jul 2020.

[36] J Z Kolter and A Y Ng. Near-Bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 513–520, Montreal, Canada, 2009.

[37] V Krishnamurthy. *Partially Observed Markov Decision Processes*. Cambridge University Press, Cambridge, UK, 1st edition, 2016.

[38] P Kumar. *Information theoretic learning methods for Markov decision processes with parametric uncertainty*. PhD thesis, University of Washington, Seattle, WA, USA, 2018.

[39] P Kumar and A Ghate. Information directed policy sampling for partially observable markov decision processes with parametric uncertainty. In *Proceedings of the INFORMS International Conference on Service Science*, Phoenix, AZ, USA, 2018.

[40] T Lai and H Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[41] X Lu. *Information-directed sampling for reinforcement learning*. PhD thesis, Stanford University, Stanford, CA, USA, 2020.

[42] X Lu and B Van Roy. Information-theoretic confidence bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2019.

[43] D M Negoescu, K Bimpkis, M L Brandeau, and D A Iancu. Dynamic learning of patient response types: An application to treating chronic diseases. *Management Science*, 64(8):3469–3488, 2017.

[44] N Nikolov, J Kirschner, F Berkenkamp, and A Krause. Information-directed exploration for deep reinforcement learning. https://arxiv.org/abs/1812.07544, 2019.

[45] I Osband, D Russo, and B Van Roy. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, Lake Tahoe, CA, USA, 2013.

[46] I Osband and B Van Roy. Why is Posterior Sampling better than optimism for Reinforcement Learning? In *Proceedings of the 22nd International Conference on Machine Learning*, pages 2701–2710, Sydney, Australia, 2017.

[47] Y Ouyang, M Gagrani, A Nayyar, and R Jain. Learning unknown Markov decision processes: A Thompson Sampling approach. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 1333–1342, Long Beach, CA, USA, 2017.

[48] B Paria, W Neiswanger, R Ghods, J Schneider, and B Póczos. Cost-aware bayesian optimization via information directed sampling. 2020. *ICML 2020 Workshop on Real World Experiment Design and Active Learning*.

[49] W B Powell and I O Ryzhov. *Optimal Learning*. John Wiley & Sons, Hoboken, NJ, USA, 2012.

[50] M L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Hoboken, NJ, USA, 2005.

[51] J Qian and J Zhang. Information-directed Sampling for Reinforcement Learning. https://www.researchgate.net/publication/337335562_Information-Directed_Sampling_for_Reinforcement_Learning, 2017.

[52] S M Ross. *Introduction to Probability Models*. Academic Press, Cambridge, MA, USA, tenth edition, 2010.

[53] D Russo and B Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[54] D Russo and B Van Roy. An information-theoretic analysis of Thompson Sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

[55] D Russo and B Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

[56] C H Schmid and E N Brown. Bayesian hierarchical models. *Methods in Enzymology*, 321:305–330, 2000.

[57] D Silver, A Huang, C J Maddison, A Guez, L Sifre, G van den Driessche, J Schrittwieser, I Antonoglou, V Pannershelvam, M Lanctot, S Dieleman, D Grewe, J Nham, N Kalchbrenner, I Sutskever, T P Lillicrap, M Leach, K Kavukcuoglu, T Graepel, and D Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(484-489), 2016.

[58] R D Smallwood and E J Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

[59] J Sorg, S Singh, and R L Lewis. Variance-based rewards for approximate bayesian reinforcement learning. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, Catalina Island, CA, USA, 2010.

[60] L N Steimle, V S Ahluwalia, C Kamdar, and B T Denton. Decomposition methods for solving Markov decision processes with multiple models of the parameters. *IISE Transactions*, 53(12):1295–1310, 2021.

[61] L N Steimle, D L Kaufman, and B T Denton. Multi-model markov decision processes. *IISE Transactions*, 53(10):1124–1139, 2021.

[62] M A Strens. A Bayesian framework for reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 943–950, San Francisco, CA, USA, 2000.

[63] R S Sutton and A G Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, second edition, 2018.

[64] W Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[65] A Tossou, D Basu, and C Dimitrakakis. Near-optimal optimistic reinforcement learning using empirical Bernstein inequalities. https://arxiv.org/abs/1905.12425, 2019.

[66] C Watkins and P Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

[67] P Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25(A):287–298, 1988.

[68] A Zanette and R Sarkar. Information directed reinforcement learning. http://web.stanford.edu/ rsarkar/materials/CS234-Project-Report.pdf, 2017.