Information theoretic learning methods for Markov decision processes with parametric uncertainty

Peeyush Kumar

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

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Archis Ghate, Chair

Youngjun Choe, Industrial & Systems Engineering

Emo Todorov, Paul Allen School of Computer Science & Engineering

Program Authorized to Offer Degree:
Industrial & Systems Engineering

University of Washington

Abstract

Information theoretic learning methods for Markov decision processes with parametric uncertainty

Peeyush Kumar

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

Markov decision processes (MDPs) model a class of stochastic sequential decision problems with applications in engineering, medicine, and business analytics. There is considerable interest in the literature in MDPs with imperfect information, where the search for well-performing policies faces many challenges. There is no rigorous universally accepted optimality criterion. The search space explodes and the decision-maker suffers from the curse-of-dimensionality. Finding good policies requires careful balancing of the trade-off between exploration to acquire information and exploitation of this information to earn high rewards. This dissertation contributes to this area by building a rigorous framework rooted in information theory for solving MDPs with model uncertainty.

In the first chapter, the value of a parameter that characterizes the transition probabilities is unknown to the decision-maker. The decision-maker updates its Bayesian belief about this parameter using state observations induced by policies it chooses. Information Directed Policy Sampling (IDPS) is proposed to manage the exploration-exploitation trade-off. At each time-stage, the decision-maker solves a convex problem to sample a policy from a distribution that minimizes a particular ratio. The numerator of this ratio equals the square of the expected regret of distributions over policy trajectories (exploitation). The denominator equals the expected mutual information between the resulting system-state trajectory and the parameter's posterior (exploration). A generalization of Hoeffding's inequality is employed to bound regret. The bound grows at a square-root rate with the planning horizon, and a square-root log-linear rate with the parameter-set cardinality. It is insensitive to state- and action-space cardinalities. The regret per stage converges to zero as the planning horizon increases. IDPS is thus asymptotically optimal. Numerical results on a stylized example, an auction-design problem, and a response-guided dosing problem demonstrate its benefits.

Uncertainty in transition probabilities arises from two levels in the second chapter. The top level corresponds to the ambiguity about the system model. Bottom-level uncertainty is rooted in the unknown parameter values for each possible model. Prior-update formulas using a hierarchical Bayesian framework are derived and incorporated into two learning algorithms: Thompson Sampling and a hierarchical extension of IDPS. Analytical performance bounds for these algorithms are developed. Numerical results on the response-guided dosing problem, which is amenable to hierarchical modeling, are presented.

The third chapter extends the above to partially observable Markov decision processes (POMDPs). In POMDPs, the decision-maker cannot observe the actual state of the system. Instead, it can take a measurement that provides probabilistic information about the true state. Such POMDPs are equivalent to Bayesian adaptive MDPs (BAMDPs) from the first two chapters. This connection is exploited to devise algorithms and provide analytical performance guarantees for POMDPs in three separate cases: a) uncertainty in the transition probabilities; b) uncertainty in the measurement outcome probabilities; and c) uncertainty in both. Numerical results on partially observed response-guided dosing are included.

The fourth chapter proposes a formal information theoretic framework inspired by stochastic thermodynamics. It utilizes the idea that information is physical. An explicit link between information entropy and stochastic dynamics of a system coupled to an environment is developed from fundamental principles. Unlike the heuristic method of defining information ratio, this provides an optimization program that is built from system dynamics, problem

objective, and the feedback from observations. To the best of my knowledge, this is the first comprehensive work in MDPs with model uncertainty, which builds a problem formulation entirely grounded in system and informational dynamics without the use of ad-hoc heuristics.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

ACKNOWLEDGMENTS

To my adviser, Prof. Archis Ghate: For the support and advice in various capacities, and for the opportunity to explore various unconventional ideas. I have learned a great deal from you both professionally and personally; and it has been my pleasure to work with you.

To my Reading Committee: Thank you for your time and effort. I really appreciate your guidance.

To the University of Washington community: For the financial support and the opportunity to work on various research problems, and to help make my time at the University an enriching experience.

To NSF: For providing financial support.

To the staff at Industrial and Systems Engineering: Thank you for your support.

To my friends in the lab: Thank you for being there and for lunch time conversations.

Chapter 1

INFORMATION DIRECTED POLICY SAMPLING FOR MARKOV DECISION PROCESSES WITH PARAMETRIC UNCERTAINTY

1.1 Introduction

Markov decision processes (MDPs) are perhaps the most widely studied models of sequential decision problems under uncertainty in Operations Research (Puterman, 1994). In this chapter, an MDP is described by the tuple $\mathcal{M} = (S, A, T, R, N)$. Here, S is a finite set of states; A is a finite set of actions; T denotes a transition probability function of the form $T(s'|s, a)$, for $s, s' \in S$ and $a \in A$; R denotes a reward function of the form $R(s'|s, a)$, for $s, s' \in S$ and $a \in A$; and N denotes a finite planning horizon. This MDP models the following time-invariant, finite-horizon, sequential decision-making problem under uncertainty. A decision-maker observes the state $s_t \in S$ of a system at the beginning of time-slot $t \in \{1, 2, \ldots, N\}$ and then chooses an action $a_t \in A$. The system then stochastically evolves to a state $s_{t+1} \in S$ by the beginning of slot $t+1$ with probability $T(s_{t+1}|s_t, a_t)$. As a result of this transition, the decision-maker collects a reward $R(s_{t+1}|s_t, a_t)$. This process of state observation, action selection, state evolution, and reward collection repeats until the end of slot N. A policy trajectory $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$ is a decision-rule that assigns actions $\pi_t(s_t) \in A$ to states $s_t \in S$, for $t = 1, 2, \ldots, N$. Note that the set $\mathcal{P}$ of such policy trajectories is finite. The decision-maker's objective is to find a policy trajectory $\pi = (\pi_1, \pi_2, \ldots, \pi_N) \in \mathcal{P}$ that maximizes the expected reward

$$J_\pi(s_1) = E\left[\sum_{t=1}^{N} R(s_{t+1}|s_t, \pi_t(s_t))\right].$$

It is assumed for simplicity of notation that no terminal reward is earned at the end of slot N.

The transition probability function is often unknown to the decision-maker at the outset. This calls for online learning of transition probabilities while the system evolves. For instance, in medical treatment planning, a doctor might not know the uncertain dose-response function of an individual at the beginning of a treatment course, but may want to adaptively make drug selection and dosing decisions over the treatment course (Kotas and Ghate, 2016). Similarly, a seller conducting a sequence of auctions may not know the bidder demand and willingness-to-pay distributions, but must adaptively make auction-design decision such as the minimum bid in each auction (Ghate, 2015). Such problems fall under the broad framework of MDPs under imperfect information, and can be seen as Bayesian adaptive MDPs (BAMDPs) or partially observable MDPs (POMDPs) in some cases (Bertsekas, 2005; Dreyfus and Law, 1977; Krishnamurthy, 2016; Kumar, 1985; Kumar and Varaiya, 2016).

In this chapter, an MDP with unknown transition probabilities is modeled as follows. The transition probability function is parameterized by a single parameter λ that takes values from a finite set Ω. The transition probability function characterized by parameter $\lambda \in \Omega$ is denoted by T_λ; the corresponding MDP is described by the tuple $\mathcal{M}_\lambda = (S, A, T_\lambda, R, N)$. For each $\lambda \in \Omega$, the decision-maker can (easily) solve $\mathcal{M}_\lambda$. The true parameter value is $\lambda^* \in \Omega$ and it is unknown to the decision-maker. If the decision-maker were an omniscient oracle, it would implement a policy trajectory π^* that is optimal for the MDP $\mathcal{M}_{\lambda^*}$, and this would maximize expected reward. Since this is not possible, the decision-maker instead pursues a Bayesian approach, wherein it starts with a prior belief probability mass function (pmf) $\alpha_1(\cdot)$ over λ^* and updates this belief as states sampled from T_{λ^*} are observed, depending on the sequence of actions chosen after starting in an initial state $s_1 \in S$.

The situation described above arises, for instance, in stochastic inventory control, where λ equals the mean of Poisson demand for a product. Possible values of λ could be $\lambda_1 < \lambda_2 < \lambda_3$ corresponding to low, medium, and high demand. This Poisson demand characterizes the transition probability function for stochastic inventory evolution. A similar setup arises in the aforementioned sequential auction problem. Finally, in the above medical treatment planning example, different values of λ may correspond to non-responders, low-responders, medium-responders, and good-responders.

Intuitively, good solutions to such problems call for balancing the trade-off between exploring policy trajectories to acquire sufficient information about the true parameter λ^*, and exploiting acquired knowledge to accumulate high rewards (Powell and Ryzhov, 2012). This chapter proposes an information theoretic approach to explicitly optimize this trade-off during run-time as the system evolves.

1.1.1 Existing literature

The above model of parametric uncertainty results in a simultaneous learning and optimization problem whose classical formulation is recalled next (Bertsekas, 2005; Dreyfus and Law, 1977; Kumar, 1985). The physical state $s_t \in S$ of the system is augmented with an information state $\alpha_t(\cdot)$ that equals the decision-maker's posterior belief pmf about λ^* at stage t. This posterior belief pmf is calculated using Bayes' Theorem. The physical- and information-state pair is a sufficient statistic for the MDP with parametric uncertainty. This allows the decision-maker to write the expected reward maximization problem as another MDP, which is called a BAMDP. In fact, this BAMDP formulation is equivalent to a POMDP where λ^* is viewed as an unobservable "state" (Krishnamurthy, 2016). Bellman's equations for the BAMDP can, in principle, be solved via the backward recursion algorithm of dynamic programming. Exact implementation of dynamic programming, however, is impossible, since the information state is continuous. Discretization of this information state for an approximate implementation is impractical because its dimension equals the cardinality of Ω, and thus could be large. Consequently, there has been a considerable interest in heuristic methods for approximately solving such BAMDP problems.

Examples of such heuristic procedures include (Duff, 2001, 2002); the BEETLE algorithm of Poupart et al. (2006); and the forward search methods from Castro and Precup (2007); Fonteneau et al. (2013); Gelly et al. (2012); Guez et al. (2012); Kocsis and Szepesvári (2006);

Poupart et al. (2006); Ross et al. (2008); Wang et al. (2005). These methods can be myopic in that they could settle for higher short term rewards by favoring more exploitation than exploration. The BEB algorithm from Kolter and Ng (2009), the VBRB algorithm of Sorg et al. (2012), the BOLT algorithm of Araya et al. (2012), and the POT procedure of Kawaguchi and Araya (2013) thus heuristically add an exploration bonus to their search process.

Strens (2000) proposed a simple approach using Thompson Sampling (Thompson, 1933). At each step of Thompson Sampling, a parameter λ is drawn from the posterior belief. An action prescribed by a policy that is optimal for $\mathcal{M}_\lambda$ is then implemented. After the resulting transition is observed, the posterior is updated, and the process repeats. Convergence of this method is, however, provably slow because of the optimistic nature of sampling (Guez et al., 2014). The BOSS algorithm of Asmuth et al. (2009) introduces a more complex version of Thompson Sampling that samples many MDPs to guide action selection. Salemi-Parizi and Ghate (2016) implemented a one-step-lookahead (Bertsekas, 2005) version of Thompson Sampling to determine lot-sizes in sequential auctions, and numerically compared it with semi-stochastic certainty equivalent control (Bertsekas, 2005) and with a one-step-lookahead version of the knowledge gradient method (Frazier et al., 2008; Ryzhov et al., 2012).

There has been a surge of interest in rigorous analyses of regret bounds for the basic Thompson Sampling method of Strens and its variants. Notable recent examples include (Gopalan and Mannor, 2015; Osband and Van Roy, 2014; Osband et al., 2013). This chapter attempts to extend the philosophy and methodology from these recent papers to a more sophisticated policy sampling method.

1.1.2 Contribution of this chapter

The main contribution of this chapter is that it introduces a provably efficient and empirically powerful algorithm that explicitly optimizes the exploitation versus exploration trade-off. The performance of this method, as characterized by a worst case regret bound, only depends on the number of possible parameter values and the horizon length; it exhibits a $\mathcal{O}(\sqrt{N|\Omega|\log|\Omega|})$ complexity. Specifically, we extend an action sampling method called Information Directed Sampling (IDS), which was originally devised in Russo and Van Roy (2014, 2017) for bandit problems, to MDPs with parametric uncertainty. In each step of IDS, the decision-maker in a bandit problem samples an action that minimizes a so-called information ratio. The numerator of this information ratio equals the square of the one-period regret of an action. The denominator equals the information gain, which is a concept from information theory. Roughly speaking, this gain calculates the information contained in an action regarding the true parameter of the problem. The idea is that the decision-maker should prefer actions with small one-period regret and large information gain. It thus makes intuitive sense to minimize the information ratio. In particular, the information ratio explicitly quantifies the trade-off between exploration (denominator) and exploitation (numerator).

Our algorithm is naturally termed Information Directed *Policy* Sampling (IDPS). In each step of IDPS, the decision-maker also minimizes the information ratio. Here, the numerator equals the squared multi-period regret of a policy trajectory, and the denominator equals the information gain of the system-state trajectory induced by this policy trajectory. Our theoretical analysis of IDPS employs a nontrivial extension of the proof technique from Russo and Van Roy (2014, 2016, 2017). Since the concept of a physical state and hence of a policy are irrelevant in bandit problems, significant additional mathematical challenges need to be carefully handled in our proof.

Section 1.2 describes the IDPS method. Section 1.3 provides bounds on the decision-maker's regret while implementing IDPS. Section 1.4 demonstrates the potential benefits of IDPS versus Thompson Sampling via a well-known toy example. Finally, additional numerical results are provided to illustrate the potential benefits of IDPS relative to Thompson Sampling on a sequential auction-design problem from Ghate (2015), and a response-guided dosing problem from Kim et al. (2009); Kotas and Ghate (2016); Maass and Kim (2017). The ideas presented in the next two sections can be extended to the case of multiple unknown parameters and parametric uncertainty in rewards. The chapter does not explicitly focus on these cases to keep notation to a minimum. Please note, however, that such more general problems are tackled in the numerical results in Section 1.4.

1.2 Information directed policy sampling

Recall that the pmf $\alpha_t(\cdot)$ denotes the decision-maker's posterior belief about λ^* at the beginning of slot t. Suppose the decision-maker observes state s_t at the beginning of slot t. Let $\pi^t = (\pi_t, \pi_{t+1}, \pi_N)$ denote the tail of any policy trajectory $\pi = (\pi_1, \pi_2, \ldots, \pi_N) \in \mathcal{P}$. The set of tail policy trajectories is denoted by $\mathcal{P}^t$. Also let $\pi^*_\lambda = (\pi^*_{1,\lambda}, \ldots, \pi^*_{N,\lambda})$ denote an optimal policy for MDP $\mathcal{M}_\lambda$. Let $V^*_t(\lambda) = \sum\limits_{\ell=t}^{N} R(s_{\ell+1}|s_\ell, \pi^*_{\ell,\lambda}(s_\ell))$ denote the random tail reward accumulated on implementing an optimal policy in MDP $\mathcal{M}_\lambda$. Similarly, $V_t(\lambda, \pi^t) = \sum\limits_{\ell=t}^{N} R(s_{\ell+1}|s_\ell, \pi_\ell(s_\ell))$ for any tail policy π^t. In both these expressions, state $s_{\ell+1}$ is drawn from T_λ given state s_ℓ and action $\pi^*_{\ell,\lambda}(s_\ell)$ or action $\pi_\ell(s_\ell)$, respectively.

Knowing s_t and α_t, the decision-maker minimizes the information ratio at the beginning of slot t. To define the information ratio, we need to first characterize the expected regret and information gain of tail policy trajectory π^t. In particular, the expected regret is defined as

$$\Delta_t(\pi^t|s_t, \alpha_t(\cdot)) = \mathop{E}_{\lambda\sim\alpha_t(\cdot)}\left[\mathop{E}_{\{s_{\ell+1}\sim T_\lambda(\cdot|s_\ell,\pi_\ell(s_\ell)):\ell=t,\ldots,N\}}\Big[V^*_t(\lambda) - V_t(\lambda, \pi^t)\Big]\right]. \tag{1.1}$$

This expression computes the expectation (with respect to the decision-maker's posterior $\alpha_t(\cdot)$) of the expected difference between the optimal value and the value of policy π^t. The inner expectation is taken with respect to the stochastic state trajectory from stage $t+1$ to the end of the planning horizon.

The information gain (or mutual information) between two random variables X and Y is given by $I(X;Y) = \sum\limits_{x,y} P(x,y) \ln \frac{P(y|x)}{P(y)}$, where the letter P denotes the appropriate joint, conditional, and marginal distributions. In our MDP context, X takes values in the parameter set Ω, while Y takes values in the observation set $\{s_{\ell+1} \sim T_{\lambda^*}(\cdot|s_\ell, \pi_\ell(s_\ell)) : \ell =$

$t, \cdots, N\}$. The information gain thus equals

$$g_t(\pi^t|s_t, \alpha_t(\cdot)) = \sum_{\lambda\in\Omega} \sum_{s_{t+1},\dots,s_N} \left(\prod_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell, \pi_\ell(s_\ell)) \right) \alpha_t(\lambda) \ln \left[\frac{\prod_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell, \pi_\ell(s_\ell))}{\sum_{\lambda\in\Omega} \prod_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell, \pi_\ell(s_\ell))\alpha_t(\lambda)} \right]. \tag{1.2}$$

Now let D^t denote the set of all pmfs over tail policies in $\mathcal{P}^t$. That is, pmf $\nu^t \in D^t$ assigns probability mass $\nu^t(\pi^t)$ to tail policy $\pi^t \in \mathcal{P}^t$. Then, given the pair $(s_t, \alpha_t(\cdot))$ at the beginning of slot t, the expected regret and expected information gain of pmf ν^t are given by

$$\Delta_t(\nu^t|s_t, \alpha_t(\cdot)) = \sum_{\pi^t\in\mathcal{P}^t} \nu^t(\pi^t)\Delta_t(\pi^t|s_t, \alpha_t(\cdot)), \ \forall \nu^t \in D^t, \tag{1.3}$$

and

$$g_t(\nu^t|s_t, \alpha_t(\cdot)) = \sum_{\pi^t\in\mathcal{P}^t} \nu^t(\pi^t)g_t(\pi^t|s_t, \alpha_t(\cdot)), \ \forall \nu^t \in D^t. \tag{1.4}$$

This notation uses operator overloading, since the operators $\Delta_t(\cdot|s_t, \alpha_t(\cdot))$ and $g_t(\cdot|s_t, \alpha_t(\cdot))$ take arguments from $\mathcal{P}^t$ as well as D^t. This is done in favor of parsimonious notation in the hope that the meaning should be clear from context. Such overloading may be used in the sequel for other quantities as well. Furthermore, the conditioning $(\cdot|s_t, \alpha_t(\cdot))$ is dropped in some expressions below for brevity; this conditioning is assumed to be implicit in such cases. For instance, the expected regret may simply be written as $\Delta_t(\nu^t)$ and the expected information gain as $g_t(\nu^t)$. The information ratio is now defined as

$$\Psi_t(\nu^t) = \frac{(\Delta_t(\nu^t))^2}{g_t(\nu^t)}. \tag{1.5}$$

The decision-maker finds a pmf over D^t by solving

$$\nu_*^t \in \operatorname*{argmin}_{\nu^t\in D^t} \Psi_t(\nu^t). \tag{1.6}$$

Let Ψ_t^* denote the optimal value $\min\limits_{\nu^t \in D^t} \Psi_t(\nu^t)$ of this problem. This is called the minimal information ratio. As in IDS, solution of the above problem is facilitated by the fact that $\Psi_t(\nu^t)$ is a convex function over $\{g_t(\nu^t) > 0\}$. Moreover, as in IDS, the above optimization problem has a two-point optimal solution. That is, there exists an optimal solution ν_*^t such that the cardinality of $\{\pi^t \in \mathcal{P}^t | \nu_*^t(\pi^t) > 0\}$ is at most two.

While implementing IDPS, a tail policy $\pi^t = (\pi_t, \ldots, \pi_N) \in \mathcal{P}^t$ is sampled from such a two-point optimal pmf ν_*^t and the action $\pi_t(s_t)$ prescribed by (the first time-component of) this policy is implemented in state s_t. The system then stochastically evolves to s_{t+1} according to $T_{\lambda^*}(\cdot|s_t, \pi_t(s_t))$. The posterior $\alpha_t(\cdot)$ is updated to $\alpha_{t+1}(\cdot)$ using Bayes' Theorem given that state s_{t+1} was observed after choosing action $\pi_t(s_t)$ in state s_t. This procedure is summarized in Algorithm 1 below. A theoretical performance analysis of IDPS is presented in the next section.

Algorithm 1 Information Directed Policy Sampling

Require: MDPs $\mathcal{M}_\lambda = \{S, A, T_\lambda, R, N\}$ for $\lambda \in \Omega$. Prior pmf $\alpha_1(\cdot)$. Initial state $s_1 \in S$.

```
function IDPS
    for episode $k = 1, 2, 3, \cdots$ do
        Set $t = 1$
        Initialize state $s_1$; and prior $\alpha_1(\cdot) \leftarrow \alpha_{N+1}(\cdot)$ if $k > 1$
        repeat
            Compute distribution $\nu_*^t = \operatorname*{argmin}_{\nu^t \in D^t} \Psi_t(\nu^t | s_t, \alpha_t(\cdot))$
            Sample $\pi^t = (\pi_t, \ldots, \pi_N) \sim \nu_*^t$
            Implement action $\pi_t(s_t)$
            Observe $s_{t+1}$ drawn from $T_{\lambda^*}(\cdot|s_t, \pi_t(s_t))$
            Update probability mass $\alpha_{t+1}(\lambda) \propto T_\lambda(s_{t+1}|s_t, \pi_t(s_t))\alpha_t(\lambda)$, for each $\lambda \in \Omega$
            t ← t+1
        until end of horizon $N$
    end for
end function
```

1.3 Theoretical analysis

This section provides an upper bound on the decision-maker's regret while implementing IDPS. Such regret bounds are often of interest for learning algorithms in MDPs (Gopalan and Mannor, 2015; Jaksch et al., 2010; Russo and Van Roy, 2016). Our method of proof is motivated by Russo and Van Roy (2014, 2016, 2017), with substantial modifications to handle the more complicated temporal dependence structure and policy sampling requirements of MDPs. It turns out that our regret bound does not explicitly depend on the cardinality of S or of A. It instead depends on the cardinality of Ω and on the Shannon entropy of the initial prior $\alpha_1(\cdot)$.

Let $\Lambda_1 : \Omega \mapsto \mathbb{R}$ denote a random variable on Ω with pmf $\alpha_1(\cdot)$ and $H(\Lambda_1|s_1)$ denote its Shannon entropy $-\sum\limits_{\lambda\in\Omega} \alpha_1(\lambda)\ln(\alpha_1(\lambda))$. A similar notation is also employed for other time-stages t whereby $H(\Lambda_t|s_t)$ denotes the Shannon entropy of the prior on Λ_t. Moreover, $H(\Lambda_t|\{s_{\ell+1} : \ell = t : N\}, \pi^t, s_t)$ is the Shannon entropy of the posterior on Λ_t having observed a trajectory $\{s_{\ell+1} : \ell = t : N\}$ while following a policy π^t starting in state s_t. The next two intermediate lemmas help later in the proof of regret bounds in Theorem 1.3.3.

Lemma 1.3.1. *The information gain $g_t(\nu^t|s_t, \alpha_t(\cdot))$ defined in (1.4) can be rewritten as*

$$g_t(\nu^t|s_t, \alpha_t(\cdot)) = H(\Lambda_t|s_t) - \underset{\pi^t\sim\nu^t}{E}\left[\underset{\{s_{\ell+1}:\ell=t:N\}}{E}[H(\Lambda_t|\{s_{\ell+1} : \ell = t : N\}, \pi^t, s_t)]\right].$$

Proof. The proof uses a property of information gain (see Lemma 5.5.6 in Gray (2011)), which states that for random variables X and Y, the information gain can be expressed as $I(X;Y) = H(X) - \sum\limits_{y\in Y} P(y)H(X|Y = y)$. Recall that, in our context, X takes values in the parameter set Ω, and Y takes values in the observation set $\{s_{\ell+1} \sim T_{\lambda^*}(\cdot|s_\ell, \pi_\ell(s_\ell)) : \ell = t :$

$N\}$. Thus, a direct application of the above property results in

$$\begin{aligned}
&g_t(\pi^t|s_t,\alpha_t(\cdot)) = I\left(\Lambda_t;\{s_{\ell+1}\sim T_{\lambda^*}(\cdot|s_\ell,\pi_\ell(s_\ell)):\ell=t:N\}\right)\\
&= H(\Lambda_t|s_t) - \sum_{\{s_{\ell+1}:\ell=t:N\}}\left(\sum_{\lambda\in\Omega}\prod_{\ell=t}^{N}T_\lambda(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))\alpha_t(\lambda)\right)H(\Lambda_t|\{s_{\ell+1}:\ell=t:N\},\pi^t,s_t)\\
&= H(\Lambda_t|s_t) - \underset{\{s_{\ell+1}:\ell=t:N\}}{E}\left[H(\Lambda_t|\{s_{\ell+1}:\ell=t:N\},\pi^t,s_t)\right].
\end{aligned}$$

Then, by taking expectation with respect to ν^t yields

$$\begin{aligned}
g_t(\nu^t|s_t,\alpha_t(\cdot)) &= \underset{\pi^t\sim\nu^t}{E}\Big[g_t(\pi^t|s_t,\alpha_t(\cdot))\Big]\\
&= \underset{\pi^t\sim\nu^t}{E}\left[H(\Lambda_t|s_t)\right] - \underset{\pi^t\sim\nu^t}{E}\left[\underset{\{s_{\ell+1}:\ell=t:N\}}{E}[H(\Lambda_t|\{s_{\ell+1}:\ell=t:N\},\pi^t,s_t)]\right]\\
&= H(\Lambda_t|s_t) - \underset{\pi^t\sim\nu^t}{E}\left[\underset{\{s_{\ell+1}:\ell=t:N\}}{E}[H(\Lambda_t|\{s_{\ell+1}:\ell=t:N\},\pi^t,s_t)]\right].
\end{aligned}$$

The last equality follows because $H(\Lambda_t|s_t)$ does not depend on π^t. This completes the proof. □

We use Ψ^* to denote $\max\limits_{t\in\{1,2,\ldots,N\}}(\Psi^*_t)$ for brevity, and $|\cdot|$ to denote set cardinalities.

Lemma 1.3.2. *Fix any arbitrary randomized tail policies $\nu^2\in D^2$, $\nu^3\in D^3$, …, $\nu^N\in D^N$. Then, for any $\epsilon>0$, we have,*

$$\begin{aligned}
&\sum_{t=1}^{N-1}\sqrt{\Psi^*_{t+1}\left(H(\Lambda_{t+1}|s_{t+1}) - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}[H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})]\right)} -\\
&\sum_{t=1}^{N-1}\underset{s_{t+1}}{E}\left[\sqrt{\Psi^*_{t+1}\left(H(\Lambda_{t+1}|s_{t+1}) - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}[H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})]\right)}\right]\\
&\leq(1+\sqrt{N-1})\sqrt{\epsilon\Psi^*\log(|\Omega|)},
\end{aligned} \tag{1.7}$$

with probability $\left(1-e^{-\frac{(c\epsilon-2)^2}{2N}}\right)\left(1-e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}\right)$*, where* $c > 2/\epsilon$ *and* $c_0 > 2/\sqrt{\epsilon}$ *are positive constants.*

Proof. The proof starts by simplifying and bounding the summations on the left hand side of Equation (2.3). We find an upper bound on the first term and a lower bound on the second term. We first consider the second term in Equation (2.3), which is given by

$$\begin{aligned}
&\sum_{t=1}^{N-1} \underset{s_{t+1}}{E}\left[\sqrt{\Psi^*_{t+1}\left(H(\Lambda_{t+1}|s_{t+1}) - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}[H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})]\right)}\right]\\
&\overset{a}{=}\sum_{t=1}^{N-1} \underset{s_{t+1}}{E}\left[\sqrt{\underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}\left[\Psi^*_{t+1}\left(H(\Lambda_{t+1}|s_{t+1}) - H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})\right)\right]}\right]\\
&\overset{b}{\geq}\sum_{t=1}^{N-1} \underset{s_{t+1}}{E}\underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}\left[\sqrt{\Psi^*_{t+1}\left(H(\Lambda_{t+1}|s_{t+1}) - H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})\right)}\right]\\
&\overset{c}{=}\sum_{t=1}^{N-1} \underset{\substack{\{s_{\ell+1}:\ell=t:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}\left[\sqrt{\Psi^*_{t+1}\left(H(\Lambda_{t+1}|s_{t+1}) - H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})\right)}\right]\\
&\overset{d}{=}\sum_{t=1}^{N-1} \underset{\substack{\{s_{\ell+1}:\ell=t:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}\left[Z_{t+1}\right]. \qquad (1.8)
\end{aligned}$$

Equality "a" follows because the entities Ψ^*_{t+1} and $H(\Lambda_{t+1}|s_{t+1})$ do not depend on the trajectory $\{s_{\ell+1}:\ell=t:N\}$. Inequality "b" follows from Jensen's inequality, which guarantees that for any random variable X, $\sqrt{(E[X])} \geq E[\sqrt{X}]$ (see (Boyd and Vandenberghe, 2004)). Equality "c" is a simple modification of notations. Equality "d" follows by defining

$$Z_{t+1} = \left(\sqrt{\Psi^*_{t+1}\left(H(\Lambda_{t+1}|s_{t+1}) - H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})\right)}\right)$$

for brevity.

Now consider the first term in Equation (2.3), which is given by

$$
\begin{aligned}
&\sum_{t=1}^{N-1} \sqrt{\Psi_{t+1}^* \left(H(\Lambda_{t+1}|s_{t+1}) - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E} [H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})] \right)} \\
&= \sum_{t=1}^{N-1} \Bigg\{ \Psi_{t+1}^* \Bigg(H(\Lambda_{t+1}|s_{t+1}) - H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1}) + \\
&\quad H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1}) - \\
&\quad \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E} [H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})] \Bigg) \Bigg\}^{1/2} \\
&\overset{a}{\leq} \sum_{t=1}^{N-1} \Bigg(\sqrt{\Psi_{t+1}^* \left(H(\Lambda_{t+1}|s_{t+1}) - H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1}) \right)} + \\
&\quad \Bigg\{ \Psi_{t+1}^* \Bigg(H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1}) - \\
&\quad \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E} [H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})] \Bigg) \Bigg\}^{1/2} \Bigg) \\
&= \sum_{t=1}^{N-1} \Bigg(Z_{t+1} + \Bigg\{ \Psi_{t+1}^* \Bigg(H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1}) - \\
&\quad \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E} [H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})] \Bigg) \Bigg\}^{1/2} \Bigg) \\
&\overset{b}{\leq} \sum_{t=1}^{N-1} Z_{t+1} + \\
&\quad \sqrt{N-1} \Bigg\{ \sum_{t=1}^{N-1} \Psi_{t+1}^* \Bigg(H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1}) - \\
&\quad \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E} [H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})] \Bigg) \Bigg\}^{1/2} \\
&\overset{c}{\leq} \sum_{t=1}^{N-1} Z_{t+1} + \sqrt{(N-1)\epsilon\Psi^* \log(|\Omega|)}, \text{ with probability } 1 - e^{-\frac{(c\epsilon-2)^2}{2N}}.
\end{aligned}
\tag{1.9}
$$

Inequality "a" follows from the algebraic relation $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$. Inequality "b" follows from Cauchy-Schwarz. Inequality "c" follows from an application of the generalized Hoeffding's inequality to Markov chains (Glynn and Ormoneit, 2002; Hoeffding, 1963). To see this, consider the sequence of random variables $Y_1, Y_2, \cdots, Y_N$ such that $Y_t = \Psi_t H(\Lambda_t | \{s_{\ell+1} : \ell = t : N\}, \pi^t, s_t)$. Here, the pair $\{\alpha_t(\cdot), s_t\}$ forms a Markov chain. Consider any $\mu > 0$. Then, by Glynn and Ormoneit (2002), we have,

$$\sum_{t=1}^{N-1} \left(Y_{t+1} - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\} \\ \pi^{t+1} \sim \nu^{t+1}}}{E} [Y_{t+1}] \right) \geq \mu,$$

with probability $\exp\left(-\frac{2}{N}\left(\frac{c\mu}{2||\Psi H||} - 1\right)^2\right)$, where $c > \frac{2||\Psi H||}{\mu}$ is a positive constant. The sup-norm $||\Psi H||$ is taken over $\{\lambda_t, \{s_{\ell+1} : \ell = t+1 : N\}, \pi^t, t\}$. The sup-norm can be further bounded above as $||\Psi H|| \leq ||\Psi|| ||H|| \leq \Psi^* \log(|\Omega|)$. Here, we have used the fact that Shannon entropy is maximum for a uniform distribution and takes the value $\log(|\Omega|)$. Substituting $\mu = \epsilon ||\Psi H||$ we get,

$$\begin{aligned}
&\sum_{t=1}^{N-1} \Big(\Psi^*_{t+1} H(\Lambda_{t+1} | \{s_{\ell+1} : \ell = t+1 : N\}, \pi^{t+1}, s_{t+1}) - \\
&\underset{\substack{\{s_{\ell+1}:\ell=t+1:N\} \\ \pi^{t+1} \sim \nu^{t+1}}}{E} \left[\Psi_{t+1} H(\Lambda_{t+1} | \{s_{\ell+1} : \ell = t+1 : N\}, \pi^{t+1}, s_{t+1}) \right] \Big) \\
&\leq \epsilon \Psi^* \log(|\Omega|),
\end{aligned}$$

with probability $1 - e^{-\frac{(c\epsilon-2)^2}{2N}}$, where $c > 2/\epsilon$ is a positive constant.

We now combine (1.8) and (1.9), and apply the Hoeffding's inequality for Markov chains one more time. Here, we consider the sequence of random variables $Z_1, Z_2, \cdots, Z_N$ such that $Z_t = \sqrt{\Psi^*_t \left(H(\Lambda_t | s_t) - H(\Lambda_t | \{s_{\ell+1} : \ell = t : N\}, \pi^t, s_t) \right)}$. Note again that the pair $\{\alpha_t(\cdot), s_t\}$

forms a Markov chain. Thus,

$$\sum_{t=1}^{N-1}\Bigg(\sqrt{\Psi^*_{t+1}\left(H(\Lambda_{t+1}|s_{t+1})-H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})\right)}$$
$$-\underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}\left[\sqrt{\Psi^*_{t+1}\left(H(\Lambda_{t+1}|s_{t+1})-H(\Lambda_{t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})\right)}\right]\Bigg)$$
$$\geq\sqrt{\epsilon}||\Psi H||,$$

with probability $e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}$, where $c_0>2/\sqrt{\epsilon}$ is a positive constant. This implies that

$$\sum_{t=1}^{N-1}Z_{t+1}-\underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}\left[\sum_{t=1}^{N-1}Z_{t+1}\right]\leq\sqrt{\epsilon\Psi^*\log(|\Omega|)},\tag{1.10}$$

with probability $1-e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}$, where $c_0>2/\sqrt{\epsilon}$ is a positive constant.

Now, putting together the results of (1.8), (1.9), and (1.10) yields the conclusion of the lemma. □

The next theorem bounds the decision-maker's cumulative expected regret over N stages, which is defined as

$$\text{Regret}(N)=\sum_{t=1}^{N}\left(\Delta_t(\nu^t_*|s_t)-\underset{s_{t+1}}{E}\left[\Delta_{t+1}(\nu^t_*|s_{t+1})\right]\right),\tag{1.11}$$

where $s_{t+1}\sim T_{\lambda^*}(\cdot|s_t,\pi^t(s_t))$ with $\pi^t\sim\nu^t_*$. Here,

$$\Delta_{t+1}(\nu^t_*|s_{t+1})=\underset{(\pi_t,\pi^{t+1})\sim\nu^t_*}{E}\left[\Delta_{t+1}(\pi^{t+1}|s_{t+1})\right],\tag{1.12}$$

with

$$\Delta_{t+1}(\pi^{t+1}|s_{t+1})=\sum_{\lambda\sim\alpha_t(\cdot)}\left[\underset{\{s_{\ell+1}:\ell=t+1:N\}}{E}\left[V^*_{t+1}(\lambda)-V_{t+1}(\lambda,\pi^{t+1})\right]\right].$$

Recall from (1.3) that $\Delta_t(\nu^t_*|s_t)$ is the expected regret if a policy π^t sampled according to

$\nu_*^t \in D^t$ is implemented in stages $t : N$ starting in state s_t. The term $\Delta_{t+1}(\nu_*^t | s_{t+1})$ defined in (1.12) above is the expected regret if a policy $\pi^t = (\pi_t, \pi^{t+1})$ sampled according to ν_*^t at time-step t is implemented in stages $t+1 : N$ starting in state s_{t+1}. Thus, the difference $\Delta_t(\nu_*^t | s_t) - \underset{s_{t+1}}{E}[\Delta_{t+1}(\nu_*^t | s_{t+1})]$ may be viewed as the one-step expected regret of ν_*^t. The cumulative expected regret over N stages as defined in (2.9) can therefore be interpreted as a sum of these one-step regrets, because the decision-maker recomputes the randomized policy ν_*^t at every time-step t.

Theorem 1.3.3. Worst Case Regret Bound: *Suppose there is a γ such that $\Psi^* \leq \gamma$. Then, for any $\epsilon > 0$, we have,*

$$\textit{Regret}(N) \leq ((1+\sqrt{N-1})\sqrt{\epsilon}+1)\sqrt{\gamma \log(|\Omega|)} \quad \textit{with probability}\ (1-e^{-\frac{(c\epsilon-2)^2}{2N}})(1-e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}),$$

where $c > 2/\epsilon$ and $c_0 > 2/\sqrt{\epsilon}$ are positive constants.

Proof. Define $g_{t+1}(\nu_*^t | s_{t+1})$ as the information gain when a policy sampled according to ν_*^t is implemented from $t+1$ onwards starting in state s_{t+1}. Then we get

$$\begin{aligned}
&\text{Regret}(N) = \sum_{t=1}^{N} \Big(\Delta_t(\nu_*^t | s_t) - \underset{s_{t+1}}{E}[\Delta_{t+1}(\nu_*^t | s_{t+1})] \Big) \\
&= \sum_{t=1}^{N} \Big(\sqrt{\Psi_t^* g_t(\nu_*^t | s_t)} - \underset{s_{t+1}}{E}[\sqrt{\Psi_{t+1} g_{t+1}(\nu_*^t | s_{t+1})}] \Big) \\
&\overset{a}{\leq} \sum_{t=1}^{N} \sqrt{\Psi_t^* \Big(H(\Lambda_t | s_t) - \underset{\substack{\{s_{\ell+1}:\ell=t:N\} \\ \pi^t \sim \nu^t}}{E}[H(\Lambda_t | \{s_{\ell+1} : \ell = t+1 : N\}, \pi^t, s_t)] \Big)} - \\
&\underset{s_{t+1}}{E}\left[\sum_{t=1}^{N} \sqrt{\Psi_{t+1}^* \left(H(\Lambda_{t+1} | s_{t+1}) - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\} \\ \pi^{t+1} \sim \nu^{t+1}}}{E}[H(\Lambda_t | \{s_{\ell+1} : \ell = t : N\}, \pi^{t+1}, s_{t+1})] \right)} \right] \\
&\overset{b}{=} \sum_{t=1}^{N-1} \sqrt{\Psi_{t+1}^* \left(H(\Lambda_{t+1} | s_{t+1}) - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\} \\ \pi^{t+1} \sim \nu^{t+1}}}{E}[H(\Lambda_t | \{s_{\ell+1} : \ell = t : N\}, \pi^{t+1}, s_{t+1})] \right)} -
\end{aligned}$$

$$\sum_{t=1}^{N-1} \underset{s_{t+1}}{E} \left[\sqrt{\Psi^*_{t+1} \left(H(\Lambda_{t+1}|s_{t+1}) - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\} \\ \pi^{t+1}\sim\nu^{t+1}}}{E} [H(\Lambda_t|\{s_{\ell+1} : \ell = t+1 : N\}, \pi^{t+1}, s_{t+1})] \right)} \right] +$$

$$\sqrt{\Psi^*_1 H(\Lambda_1|s_1) - \underset{\substack{\{s_{\ell+1}:\ell=1:N\} \\ \pi^{1}\sim\nu^{1}}}{E} [H(\Lambda_1|\{s_{\ell+1} : \ell = 1 : N\}, \pi^1, s_1)]} -$$

$$\sqrt{\Psi^*_{N+1} H(\Lambda_{N+1}|s_{N+1})}$$

$$\overset{\text{c}}{\leq} (1 + \sqrt{N-1})\sqrt{\epsilon\gamma\Psi^* \log(|\Omega|)} + \sqrt{\Psi^* H(\Lambda_1|s_1)}, \text{ with prob. } \left(1 - e^{-\frac{(c\epsilon-2)^2}{2N}}\right)\left(1 - e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}\right)$$

$$\leq (1 + \sqrt{N-1})\sqrt{\epsilon\gamma \log(|\Omega|)} + \sqrt{\gamma H(\Lambda_1|s_1)}$$

$$\overset{\text{d}}{\leq} ((1 + \sqrt{N-1})\sqrt{\epsilon} + 1)\sqrt{\gamma \log(|\Omega|)}.$$

Here, inequality “a” follows from two ideas. The first is an equality that follows from Lemma 1.3.1, and the second is that $\Psi_{t+1} \geq \Psi^*_{t+1}$. Equality “b” is just a rearrangement of entities inside the summation. Inequality “c” follows from Lemma 1.3.2. Inequality “d” holds because the Shannon entropy is maximum for a uniform distribution where it equals $\log(|\Omega|)$. □

The next result obtains a uniform upper bound on the minimal information ratio. This uniform bound is then inserted back into Theorem 1.3.3 to derive Corollary 1.3.6. Russo and Van Roy (Russo and Van Roy, 2014, 2017) used a similar approach to bound minimal information ratios in bandit problems. Our proof is its generalization to MDPs. We first recall the following fact from Russo and Van Roy (2016).

Fact 1.3.4. *Let P and Q be any distributions such that P is absolutely continuous with respect to Q. Consider any random variable $X : \Omega \to R$ such that $\sup(X) - \inf(X) \leq 1$. Then,*

$$E_P[X] - E_Q[X] \leq \sqrt{\frac{1}{2} D_{KL}(P||Q)},$$

where $D_{KL}(P||Q) = -\sum_x P(x) \log\left(\frac{Q(x)}{P(x)}\right)$ denotes Kullback-Leibler divergence (for discrete distributions) (Gray, 2011).

Proposition 1.3.5. Bounds on minimal information ratio: *The minimal information ratio is bounded above by $|\Omega|/2$.*

Proof. The information gain in (1.2) can be expressed as

$$g_t(\pi^t|s_t) = \underset{\lambda\sim\alpha_t(\cdot)}{E}\left\{D_{KL}\left[\left(\prod_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))\right)\middle\|\sum_{\lambda\in\Omega}\prod_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))\alpha_t(\lambda)\right]\right\}.$$

Define

$$U_t(\pi^t,\lambda) = \underset{\{s_{\ell+1}\sim T_\lambda(\cdot|s_\ell,\pi_\ell(s_\ell)):\ell=t,\ldots,N\}}{E}[V_t(\lambda,\pi^t)],$$

and

$$L_t(\pi^t,\lambda) = U_t(\pi^t,\lambda) - \underset{\lambda\sim\alpha_t(\cdot)}{E}\left[U_t(\pi^t,\lambda)\right].$$

Further define

$$\Psi_t^L(\nu^t) = \frac{\Delta_t(\nu^t)^2}{\underset{\lambda\sim\alpha_t(\cdot)}{E}[L_t(\nu^t,\lambda)^2]},$$

and

$$\nu_L^t = \underset{\nu^t\in D^t}{\operatorname{argmin}}\,\Psi_t^L(\nu^t).$$

At the optimal solution ν_{IDPS}^t to the information ratio minimization problem, we have, $\Psi_t(\nu_{IDPS}^t) \le \Psi_t(\nu_L^t)$.

$$\begin{aligned}
\Psi_t(\nu^t) &= \frac{\Delta_t(\nu^t)^2}{\underset{\lambda\sim\alpha_t(\cdot)}{E}\left\{D_{KL}\left[\left(\prod_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))\right)\middle\|\sum_{\lambda\in\Omega}\prod_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))\alpha_t(\lambda)\right]\right\}}\\
&\overset{\text{a}}{\le} \frac{\Delta_t(\nu^t)^2}{2\underset{\lambda\sim\alpha_t(\cdot)}{E}[U_t(\pi^t,\lambda) - \underset{\lambda\sim\alpha_t(\cdot)}{E}[U_t(\pi^t,\lambda)]]^2}\\
&=\frac{1}{2}\Psi_t^L(\nu^t).
\end{aligned}$$

Here, "a" follows from Fact 1.3.4 with distribution P identified as $P = \prod_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))$

and distribution Q as $Q = \sum\limits_{\lambda \in \Omega} \prod\limits_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell, \pi_\ell(s_\ell))\alpha_t(\lambda)$ and $X = U_t(\pi^t, \lambda)$. We assume that $\sup(U_t(\pi^t, \lambda)) - \inf(U_t(\pi^t, \lambda)) \leq 1$. For a finite MDP, stage-rewards and hence value functions are uniformly bounded. Thus, the assumption $\sup(U_t(\pi^t, \lambda)) - \inf(U_t(\pi^t, \lambda)) \leq 1$ holds without any loss of generality. This is because if $\sup(U_t(\pi^t, \lambda)) - \inf(U_t(\pi^t, \lambda)) > 1$, we can rescale the rewards such that the new rewards $R(s, a) \leftarrow R(s, a)/(\sup(U_t(\pi^t, \lambda)) - \inf(U_t(\pi^t, \lambda)))$. This rescaling does not affect any decision rules for the MDP but ensures that $\sup(U_t(\pi^t, \lambda)) - \inf(U_t(\pi^t, \lambda)) \leq 1$ making the theorem valid. Also note that the distribution P is absolutely continuous with respect to Q because $Q(\cdot) = 0$ implies that $P(\cdot) = 0$. By definition $Q = \sum\limits_{\lambda \in \Omega} \prod\limits_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell, \pi_\ell(s_\ell))\alpha_t(\lambda)$ and $\prod\limits_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell, \pi_\ell(s_\ell))\alpha_t(\lambda) \geq 0;\ \forall \lambda$. For the sum of non-negative components to be 0 it is required that each individual component should be 0. Thus, either $\prod\limits_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell, \pi_\ell(s_\ell)) = 0$ or $\alpha_t(\lambda) = 0$. Since this is true for any arbitrary α_t and $\sum_{\lambda \in \Omega} \alpha_t(\lambda) = 1$, it implies that $\prod\limits_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell, \pi_\ell(s_\ell)) = 0$, which corresponds to $P(\cdot) = 0$. This yields

$$\Psi_t(\nu^t_{IDPS}) \leq \Psi_t(\nu^t_L) \leq \Psi^L_t(\nu^t_L)/2. \tag{1.13}$$

Now let ν^t_{PS} be the distribution over the policies, where each policy is optimal with respect to the MDP $\mathcal{M}_\lambda$ obtained from sampling the corresponding parameter λ from the posterior distribution α_t. Hence by definition, the distribution ν^t_{PS} is same as α_t with the identification that the domain consists of optimal policies of M_λ, for all $\lambda \in \Omega$. Therefore,

$$\underset{\lambda \sim \alpha_t(\cdot)}{E}[L_t(\nu^t_{PS}, \lambda)^2] = \underset{\substack{\lambda \sim \alpha_t(\cdot) \\ \pi \sim \nu^t_{PS}}}{E}[L_t(\pi^t, \lambda)^2] \overset{\text{a}}{=} \underset{\substack{\lambda \sim \alpha_t(\cdot) \\ \pi \sim \alpha_t(\cdot)}}{E}[L_t(\pi^t, \lambda)^2]. \tag{1.14}$$

Equality "a" holds because of the aforementioned equivalence between sampling policies from ν^t_{PS} and from $\alpha_t(\cdot)$. Similarly,

$$\Delta_t(\nu^t_{PS}) = \underset{\pi^t \sim \nu^t_{PS}}{E}[\Delta_t(\pi^t)] = \underset{\pi^t \sim \alpha_t(\cdot)}{E}[\Delta_t(\pi^t)]$$

$$
\begin{aligned}
&= \underset{\pi^t \sim \alpha_t(\cdot)}{E}\left[\underset{\lambda \sim \alpha_t(\cdot)}{E}\left[\underset{\{s_{\ell+1}\sim T_\lambda(\cdot|s_\ell,\pi_\ell(s_\ell)):\ell=t,\ldots,N\}}{E}\Big[V_t^*(\lambda) - V_t(\lambda,\pi^t)\Big]\right]\right] \\
&\overset{\text{b}}{=} \underset{\pi^t \sim \alpha_t(\cdot)}{E}\left[\underset{\{s_{\ell+1}\sim T_\lambda(\cdot|s_\ell,\pi_\ell(s_\ell)):\ell=t,\ldots,N\}}{E}\Big[V_t(\lambda,\pi^t)\Big]\right. \\
&- \left.\underset{\lambda \sim \alpha_t(\cdot)}{E}\left[\underset{\{s_{\ell+1}\sim T_\lambda(\cdot|s_\ell,\pi_\ell(s_\ell)):\ell=t,\ldots,N\}}{E}\Big[V_t(\lambda,\pi^t)\Big]\right]\right] \\
&= \underset{\pi^t \sim \alpha_t(\cdot)}{E}[L_t(\pi^t,\lambda)].
\end{aligned}
$$

Equality “‘b”’ holds due to an extension of the logic behind equality “a”. Let $\mathcal{P}^*$ be the set of policies optimal for MDPs $\mathcal{M}_\lambda$, $\forall\lambda \in \Omega$. The above RHS is then bounded using Cauchy-Schwarz as

$$
\begin{aligned}
\underset{\pi^t \sim \alpha_t(\cdot)}{E}[L_t(\pi^t,\lambda)] &= \sum_{\pi^t \in \mathcal{P}_*^t} \alpha_t(\pi^t) L_t(\pi^t,\lambda) \qquad (1.15)\\
&\leq \sqrt{|\Omega| \sum_{\pi^t \in \mathcal{P}_*^t} (\alpha_t(\pi^t) L_t(\pi^t,\lambda))^2} \\
&= \sqrt{|\Omega|}\sqrt{\sum_{\pi^t,\lambda} \alpha_t(\pi^t)\alpha_t(\lambda) L_t(\pi^t,\lambda)^2} \\
&= \sqrt{|\Omega|}\sqrt{\underset{\substack{\lambda\sim\alpha_t(\cdot)\\ \pi\sim\alpha_t(\cdot)}}{E}[L_t(\pi^t,\lambda)^2]}.
\end{aligned}
$$

This implies that $\Delta_t(\nu_{PS}^t)^2 \leq |\Omega| \underset{\substack{\lambda\sim\alpha_t(\cdot)\\ \pi^t\sim\alpha_t(\cdot)}}{E}[L^t(\pi_t,\lambda)^2]$. Now recall that

$$
\Psi_t^L(\nu_{PS}^t) = \frac{\Delta_t(\nu_{PS}^t)^2}{\underset{\lambda\sim\alpha_t(\cdot)}{E}[L_t(\nu_{PS}^t,\lambda)^2]} = \frac{\Delta_t(\nu_{PS}^t)^2}{\underset{\substack{\lambda\sim\alpha_t(\cdot)\\ \pi\sim\alpha_t(\cdot)}}{E}[L_t(\pi^t,\lambda)^2]} \leq \frac{|\Omega| \underset{\substack{\lambda\sim\alpha_t(\cdot)\\ \pi\sim\alpha_t(\cdot)}}{E}[L_t(\pi^t,\lambda)^2]}{\underset{\substack{\lambda\sim\alpha_t(\cdot)\\ \pi\sim\alpha_t(\cdot)}}{E}[L_t(\pi^t,\lambda)^2]} = |\Omega|, \qquad (1.16)
$$

where the second equality follows from (1.14) and the inequality holds because of the above upper bound on the numerator. Also note that $\Psi_t^L(\nu_L^t) \leq \Psi_t^L(\nu_{PS}^t)$ by optimality of ν_L^t.

Using this in (2.13) and combining with the above upper bound on $\Psi_t^L(\nu_{PS}^t)$ yields

$$\Psi_t(\nu_{IDPS}^t) \leq \Psi_t^L(\nu_{PS}^t)/2 \leq |\Omega|/2. \tag{1.17}$$

This completes the proof. □

Corollary 1.3.6. *The cumulative expected regret is bounded as*

$$\mathit{Regret}(N) \leq \frac{((1+\sqrt{N-1})\sqrt{\epsilon}+1)}{\sqrt{2}}\sqrt{|\Omega|\log(|\Omega|)},$$

with probability $\left(1-e^{-\frac{(c\epsilon-2)^2}{2N}}\right)\left(1-e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}\right)$, *where* $c > 2/\epsilon$ *and* $c_0 > 2/\sqrt{\epsilon}$ *are positive constants.*

Proof. Follows from Theorem 1.3.3 and Proposition 1.3.5. □

A regret bound of $\mathcal{O}(S\sqrt{AN})$ is provided in (Osband et al., 2013) for Thompson Sampling. Our worst-case regret bound in the above corollary for IDPS only depends on $|\Omega|$ and N. This provides a better regret scaling when the state/policy space is large but where the set of uncertain parameters is relatively much smaller. In addition, it provides insight via explicit dependence on the prior information. Finally, note that the running average regret, that is Regret$(N)/N$, asymptotically approaches zero at the rate $(\epsilon/\sqrt{N}) + (1/N)$. This shows that IDPS is asymptotically optimal. In the next section, we supplement these theoretical guarantees with numerical experiments.

1.4 Numerical results

Our numerical results are split into two sections. Section 1.4.1 studies a well-known stylized example to gain insight into why IDPS might work well on some problems. Sections 1.4.2 and 1.4.3 implement IDPS on a business analytics problem and a medical treatment planning problem, respectively, and compare its performance against Thompson Sampling.

1.4.1 Illustrative example: 5-state MDP

In this section, we illustrate the potential advantage of using IDPS via the chain problem from (Dearden et al., 1998). The decision maker has 2 actions $\{a, b\}$ available, which cause transitions between 5 states of the chain shown in Figure 1.1. Action a takes the agent to the next state with a probability λ_1, unless it is the state 5, where it causes the agent to remain in the same state with same probability λ_1. Additionally, action a can cause the agent to transit to state 1 with probability λ_2. Action b also causes the same transitions but with reversed probabilities. Consider a discounted infinite horizon version of this problem with discount factor 0.9. The true parameter values are $\{\lambda_1^*, \lambda_2^*\} = \{0.8, 0.2\}$. The decision-maker knows that the true parameter value belongs to the set $\Omega = \{\{0.2, 0.8\}, \{0.8, 0.2\}\}$. The objective is to maximize the cumulative expected discounted reward while learning the true parameter value $\lambda^* = \{0.8, 0.2\}$. It is known that the optimal policy here is to follow action a in every state, and that learning algorithms tend to get stuck in loops of choosing action b (Dearden et al., 1998). In our experience, on the other hand, IDPS performed quite well on this problem. IDPS realized that the maximum information gain is obtained while choosing action a in all states. This action happens to be the optimal action in every state, hence the expected value accumulated while following the policy sampled by IDPS was higher than Thompson Sampling. The average cumulative reward on starting in state 1 was 150.86 for IDPS and 139.33 for Thompson Sampling with 10 episodes of $N = 50$ time-stages each. (The results were averaged over 50 independent simulations).

While IDPS performed well in the above example, the information structure of that example is perhaps too simple to justify executing IDPS. We therefore consider a modification of the 5-state MDP. Here, in addition to actions $\{a, b\}$, the decision maker has access to a third action c. Action c skips the next state and takes the system to the state after that, with probability γ; the system remains in the same state with probability $1-\gamma$. If the system is in state 4 or state 5, action c takes the system to state 1 or state 2, respectively, with probability γ. The decision-maker has to learn the true parameter value $l^* = \{\gamma^*, \lambda^*\} = \{0.6, 0.4, 0.6\}$,

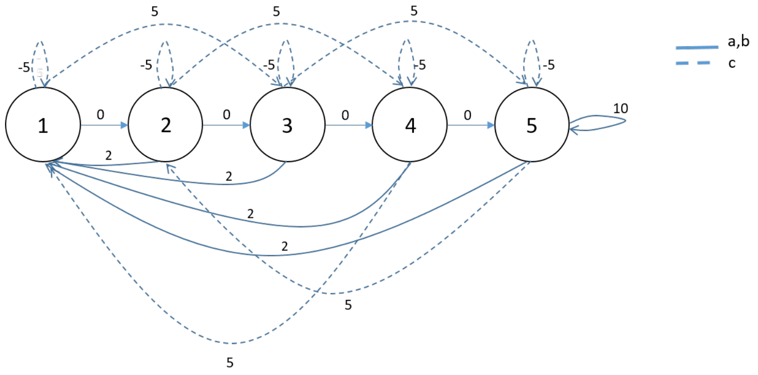


Figure 1.1: Schematic of the MDP with 5 states discussed in Section 1.4.1. Rewards are shown next to transition arcs.

while maximizing the cumulative reward. The decision maker knows that the true parameter value belongs to the set $\Omega = \{\{0.6, 0.4, 0.6\}, \{0.9, 0.4, 0.4\}\}$. For this problem, choosing action a for all states is optimal if the true parameter is $l_1 = \{0.6, 0.4, 0.6\}$. We call this policy π^a. Action b is optimal in all states if the true parameter is $l_2 = \{0.9, 0.4, 0.4\}$. We call this policy π^b. Policy π^c corresponds to choosing action c in all states. This policy is not optimal for l_1 or for l_2.

Table 1.1: Properties of three policies for the 5-state MDP in Section 1.4.1.

	Information gain	Expected Regret	Value Function
Policy π^a	0.0130	11.61	12.32
Policy π^b	0.0132	11.65	12.21
Policy π^c	0.0620	12.50	10.05

Table 1.1 shows the value functions for the three policies starting in state 1. If the decision-maker uses Thompson Sampling, it will never chose policy π^c because π^c is not optimal for any parameter combination from Ω. On the other hand, policy π^c provides information about the other policies. Indeed, Table 1.1 shows that policy π^c has a larger information gain compared to π^a, π^b. IDPS will thus select policy π^c in the beginning, acquire information about the optimal policy π^a and will select that policy in future epochs. Table 1.2

shows the average cumulative rewards for Thompson Sampling and IDPS starting in state 1, for 3 consecutive episodes each of $N = 50$ time-steps. The table shows that, although IDPS chooses a non-optimal policy π^c in the beginning owing to its higher information gain, it is still able to attain better rewards than Thompson Sampling in each of the three episodes.

Table 1.2: Cumulative rewards for the 5-state MDP discussed in Section 1.4.1.

	Episode 1	Episode 2	Episode 3
IDPS	4.5	8.5	10.9
Thompson Sampling	1.8	7.3	10.5

1.4.2 Optimizing minimum bids in sequential Vickery auctions with unknown demand

In this section, we implement and compare Thompson Sampling and IDPS on a learning problem in sequential Vickrey auctions that was originally studied by (Ghate, 2015). Consider a seller who initially holds $I \geq 1$ units in her inventory. The seller uses a sequence of online, single-unit auctions of equal durations to clear inventory and generate revenue. A single unit is put up for auction and the minimum bid requirement is announced. The seller uses the Vickrey, that is, the second price-sealed-bid mechanism. Remaining units are held in inventory during the auction. A cost of $h \geq 0$ is incurred per unit held in inventory over the duration of the auction and is charged at the beginning of the auction. At the end of each auction, inventory level either drops by one (if at least one bid is posted) or stays the same (if no bids are posted). This process continues until all inventory is cleared. The discount factor over the duration of each auction is $0 < \delta < 1$. The seller's goal is to find a policy for minimum bid decisions to maximize total discounted expected profit over all auctions. Bidders across different auctions are independent; each bidder has a private valuation for a single item; these private valuations of different bidders in any one auction and across auctions are independent and identically distributed (iid) random variables; and the numbers of potential bidders across auctions are iid random variables.

The random number of potential bidders in any auction is denoted by $\mathcal{N}$. Let $p(n)$ be the probability mass function of $\mathcal{N}$ with support $\{0, 1, 2, \cdots\}$ and mean $0 < \lambda < \infty$. This demand is Poisson distributed. The seller knows this but she does not know its mean λ. The seller has a prior belief distribution on λ and that she updates this belief over multiple auctions by observing the number of posted bids.

Let χ denote the distribution function of bidder valuations with a density function θ that satisfies $\theta(v) > 0$ for all $v \in (0, 1)$. The seller's task is to find the minimum bid $b \in [0, 1]$ in each auction after observing the remaining inventory to maximize expected total discounted profit. Bidders whose valuations are less than b will not post a bid, whereas others will. Denote the probability that no bid is posted by $q_\lambda(b)$, for any given value of λ. Recall from Ghate (2015) that

$$q_\lambda(b) = e^{-\lambda(1-\chi(b))},$$

and for linearly distributed private valuations with negative slope,

$$\chi(b) = 2b - b^2.$$

The transition probability function is given by

$$T_\lambda(s'|s, b) = \begin{cases} q_\lambda(b), & \text{for } s' = s \\ 1 - q_\lambda(b), & \text{for } s' = s - 1. \end{cases}$$

Following Ghate (2015), define $\phi(b)$ to be the expected revenue earned in one single-unit Vickrey auction with minimum bid b. This is given by

$$\phi(b) = 1 - \frac{3b-1}{2} q_\lambda(b) - \frac{3}{4}\sqrt{\frac{\pi}{\lambda}} erf\left(\sqrt{\lambda(1-b)}\right).$$

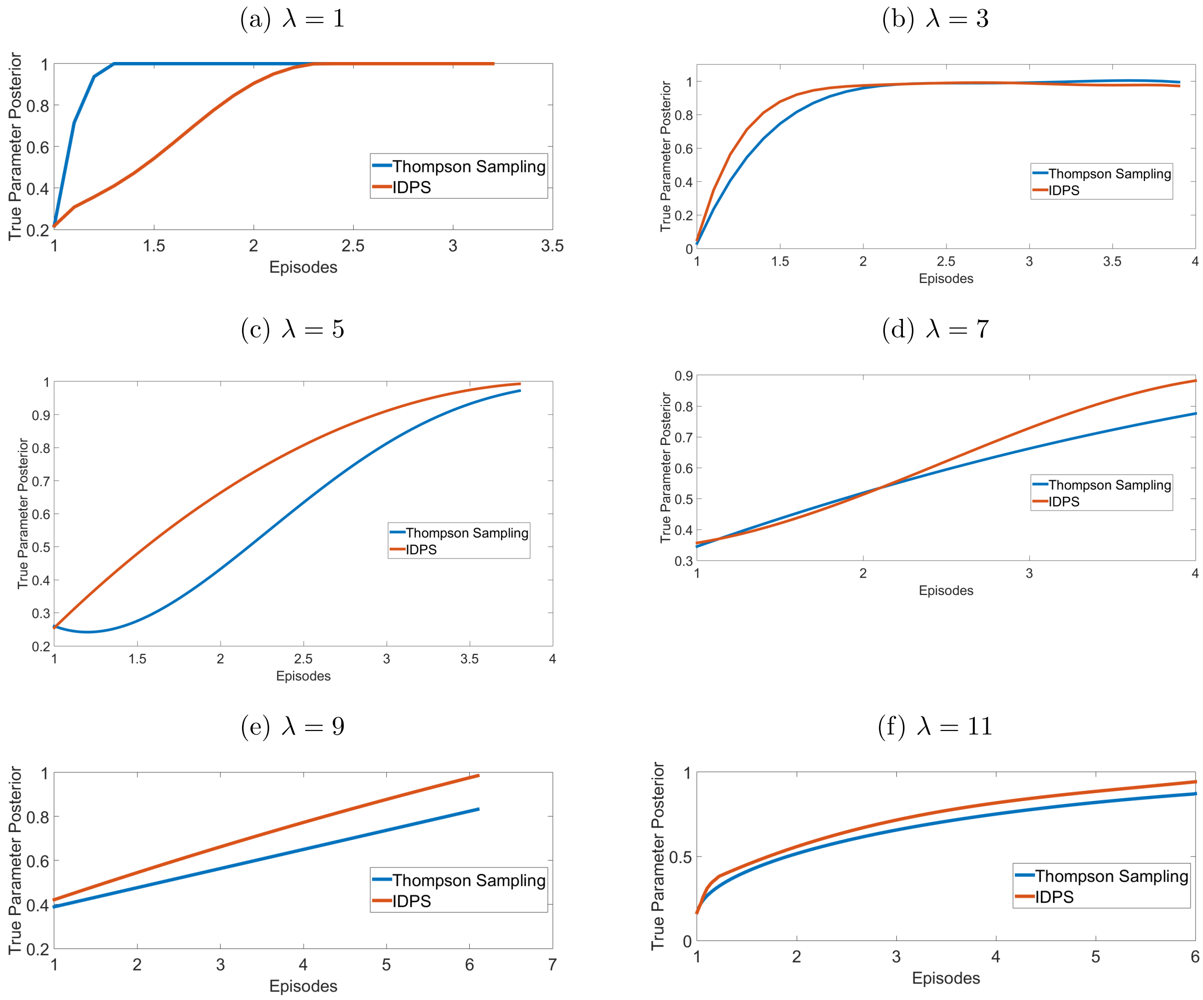
Even when λ is known, the auctioneer faces complicated economic trade-offs while dynamically deciding minimum bids b as a function of inventory levels across different auctions.

Too high a minimum bid would filter out a large portion of the bidding population thus reducing bidder-competition, but at the same time would ensure that the closing price is high as long as at least one bid is posted. A low minimum bid on the other hand allows several bidders to participate, thus potentially increasing competition, but the actual posted bids and hence the closing price might still be somewhat low. In addition, bidders in different auctions compete against each other through the opportunity cost of limited inventory. For large values of λ, the auctioneer can afford to be more aggressive in selecting minimum bids because, roughly speaking, a larger number of bidders is interested in participating. In contrast, for smaller λ values, the auctioneer may prefer to use lower minimum bid values to make the most out of the small demand. When λ is unknown, the auctioneer needs to balance such trade-offs associated with learning with the above trade-offs related to revenue maximization, thus further complicating solution.

This problem can be modeled as an MDP with parametric uncertainty in λ. A BAMDP formulation of this problem was presented in Ghate (2015). It was solved using heuristic approximate dynamic programming methods. We applied IDPS and compared it with Thompson Sampling on this problem.

Figure 1.2 plots the seller's posterior about the true parameter value for IDPS and Thompson Sampling. Observe that when the information structure is complex enough to benefit from more sophisticated exploration, IDPS learns the true parameter value better than Thompson sampling. This is the case when $\lambda^* = 3, 5, 7, 9, 11$. On the other hand, when $\lambda^* = 1$, Thompson Sampling seems to learn better. This can perhaps be attributed to a simpler information structure for smaller λ values; a small lambda value implies low demand and consequently insufficient impact of decisions on future observed states and rewards to justify sophisticated sampling. Figure 1.3 and Tables 1.3, 1.4 provide running average regrets and posterior beliefs for IDPS and Thompson Sampling. All results are averaged over 50 independent simulation runs. The results are consistent with Figure 1.2: IDPS seems to

Figure 1.2: Posterior belief about the true parameter value for the auction problem in Section 1.4.2.



(a) $\lambda = 1$

(b) $\lambda = 3$

(c) $\lambda = 5$

(d) $\lambda = 7$

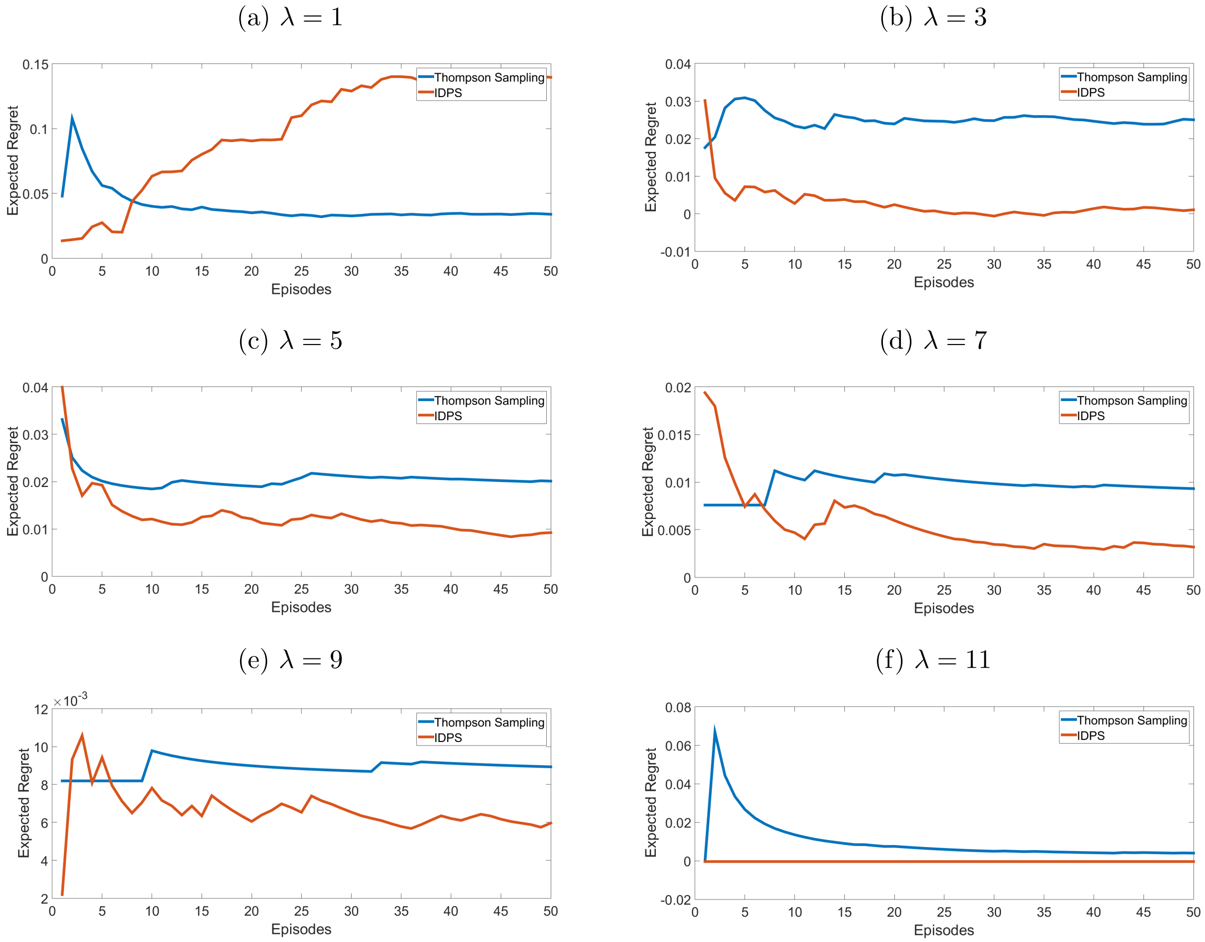
(e) $\lambda = 9$

(f) $\lambda = 11$

Table 1.3: Thompson Sampling: Posterior for the true parameter at the start of each episode for the auction problem.

Episode / Parameter	1	2	3	4	5	6
1	0.22	1	1	1	1	1
3	0.05	0.91	1	1	1	1
5	0.26	0.42	0.80	0.97	1	1
7	0.35	0.49	0.61	0.72	0.81	0.87
9	0.41	0.47	0.56	0.65	0.73	0.83
11	0.17	0.51	0.65	0.75	0.82	0.87

Figure 1.3: Running average expected regret for different true parameter values for the auction problem in Section 1.4.2.


(a) $\lambda = 1$

(b) $\lambda = 3$

(c) $\lambda = 5$

(d) $\lambda = 7$

(e) $\lambda = 9$

(f) $\lambda = 11$

Table 1.4: IDPS: Posterior for the true parameter at the start of each episode for the auction problem.

Episode Parameter	1	2	3	4	5	6
1	0.22	0.78	1	1	1	1
3	0.05	1	1	1	1	1
5	0.26	0.68	0.93	1	1	1
7	0.35	0.47	0.65	0.81	0.93	1
9	0.42	0.53	0.65	0.76	0.86	0.98
11	0.17	0.55	0.71	0.82	0.88	0.94

outperform Thompson Sampling in all cases except $\lambda = 1$.

In the next section, we apply our methodology to another problem where the learning versus optimization trade-off is crucial.

1.4.3 Response-guided dosing

This section considers an MDP formulation for response-guided dosing (RGD) in diseases that call for treatment courses with multiple sessions. The objective in RGD is to tailor drug-doses or various treatment modalities to the stochastic evolution of each individual patient's disease condition over the treatment course, in order to trade-off the patient's aversion to doses with disease control. Here, a patient's aversion to dose may stem from ill-effects of treatment such as high cost, side effects on health, and logistical inconvenience of administering/receiving treatment (Kim et al., 2009; Kotas and Ghate, 2016; Maass and Kim, 2017).

Here, at the beginning of each treatment session, the system state includes a numerical score of the patient's disease condition, and also a numerical score of the treatment's side effect. Disease condition examples include cholesterol level for heart disease; viral load for hepatitis C; blood pressure; DAS28 scores for rheumatoid arthritis; or tumor size for cancer patients. Side effect state examples include toxicity of radiation on healthy tissue for cancer radiotherapy or platelet levels for chemotherapy. The decisions correspond to the doses administered to the patient in each session after observing the state. The immediate cost is given by a disutility function that models patients' aversion to doses. The disease conditions and side effects evolve according to a stochastic dose-response function of the dose level. The decision-maker's goal is to minimize the total expected disutility of the doses given to the patient over the treatment course and that of the disease condition reached at the end of the course. The fundamental trade-off in this problem is that high dose levels are likely to achieve better disease control, but at the same time, may induce worse side effects. A good dosing strategy calls for adapting doses to the stochastic evolution of each individual patient's disease condition. This in turn requires learning a key parameter of the individual patient's dose-response function over the treatment course while simultaneously selecting doses.

The model here is an extension of the framework in Kim et al. (2009); Maass and Kim (2017), where the authors focus only on the perfect information special case. We consider a treatment course with N sessions wherein disease condition and side effect measurements are made and then a drug dose is administered at the beginning of sessions $t = 1, 2, \dots, N$. The disease condition in session t is denoted by X_t, patient's side effect is denoted by Y_t, and the dose chosen for this session after measuring $s_t = \{X_t, Y_t\}$ is denoted by d_t. Disease state X_t is integer value and belongs to the interval $\mathcal{X} = [0, m]$ with $X_t = 0$ representing the best disease state and $X_t = m$ denotes the worst disease state. Patient's side effect state Y_t is also integer valued and belongs to the interval $\mathcal{Y} = [0, n]$ with $Y_t = 0$ representing no side effect and $Y_t = n$ representing the worst side effect. Doses d_t are also integer valued and belong to the interval $D = [0, \bar{d}]$, where $\bar{d} < \infty$ is the maximum permissible dose in one session.

The disease condition and treatment's side effects evolve according to a probability distribution. The transition probabilities between disease states is denoted by $P^X(X_{t+1}|X_t, d_t)$, transition probability between side effects is denoted by $P^Y(Y_{t+1}|Y_t, d_t)$, and the state transition probability is given by $P(s_{t+1}|s_t, d_t) = P^X(X_{t+1}|X_t, d_t) \times P^Y(Y_{t+1}|Y_t, d_t)$. For simplicity, we assume that state variables can only change by at most one unit between two successive treatment periods, and that disease progression only improves after treatment dose while side effect only improves after a session with zero dose. Qualitatively, the disease state has a higher probability of improving, characterized by decreasing state value, with higher doses. When the dose equals zero, the disease state has a nonzero probability of getting worse. The side effect state has a higher chance of getting worse with higher doses and has a nonzero probability of improving when no dose is given. The probabilistic evolution of side effects is characterized by a parameter λ. The planner does not know the true value of this parameter. Specifically, for our numerical simulation, we employed the state transition probabilities described in Table 1.5 for non-boundary states. For boundary states, we used

$$P^X(0|0, d_t) = 1 \text{ and } P^Y(n|n, d_t) = 1, \text{ for } d_t \neq 0;$$

and

$$P^X(m|m, d_t) = 1 \text{ and } P^Y(0|0, d_t) = 1, \text{ for } d_t = 0.$$

The patient's utility function was assumed to take the form

$$r(s_{t+1}|s_t, d_t) = c_X(X_{t+1}, q_X) + c_Y(Y_{t+1}, q_Y) - cd_t,$$

where

$$c_X(X_t, q_X) = \frac{1}{m^{q_X}}(m^{q_X} - X_t^{q_X}),$$

and

$$c_Y(Y_t, q_Y) = \frac{1}{n^{q_Y}}(n^{q_Y} - Y_t^{q_Y}).$$

The function $c_X(X_t, q_x)$ represents the patients utility while being in disease state X_t; the function $c_Y(Y_t, q_Y)$ represents the patient utility with side effect Y_t. The higher disease/side effect states imply lower utility for the patient, as can be observed in the function forms. The last term cd_t represents the patient's disutility on taking higher doses. For our numerical simulation, we employed $m = n = 6$, $c = 6$, $q_X = q_Y = 2$, and $\bar{d} = 3$.

Table 1.5: State transition probabilities $P^X(X_{t+1}|x_t, d_t)$ and $P^Y(Y_{t+1}|Y_t, d_t)$. The decision maker does not know λ, which is a parameter that characterizes the transition probabilities for side effects.

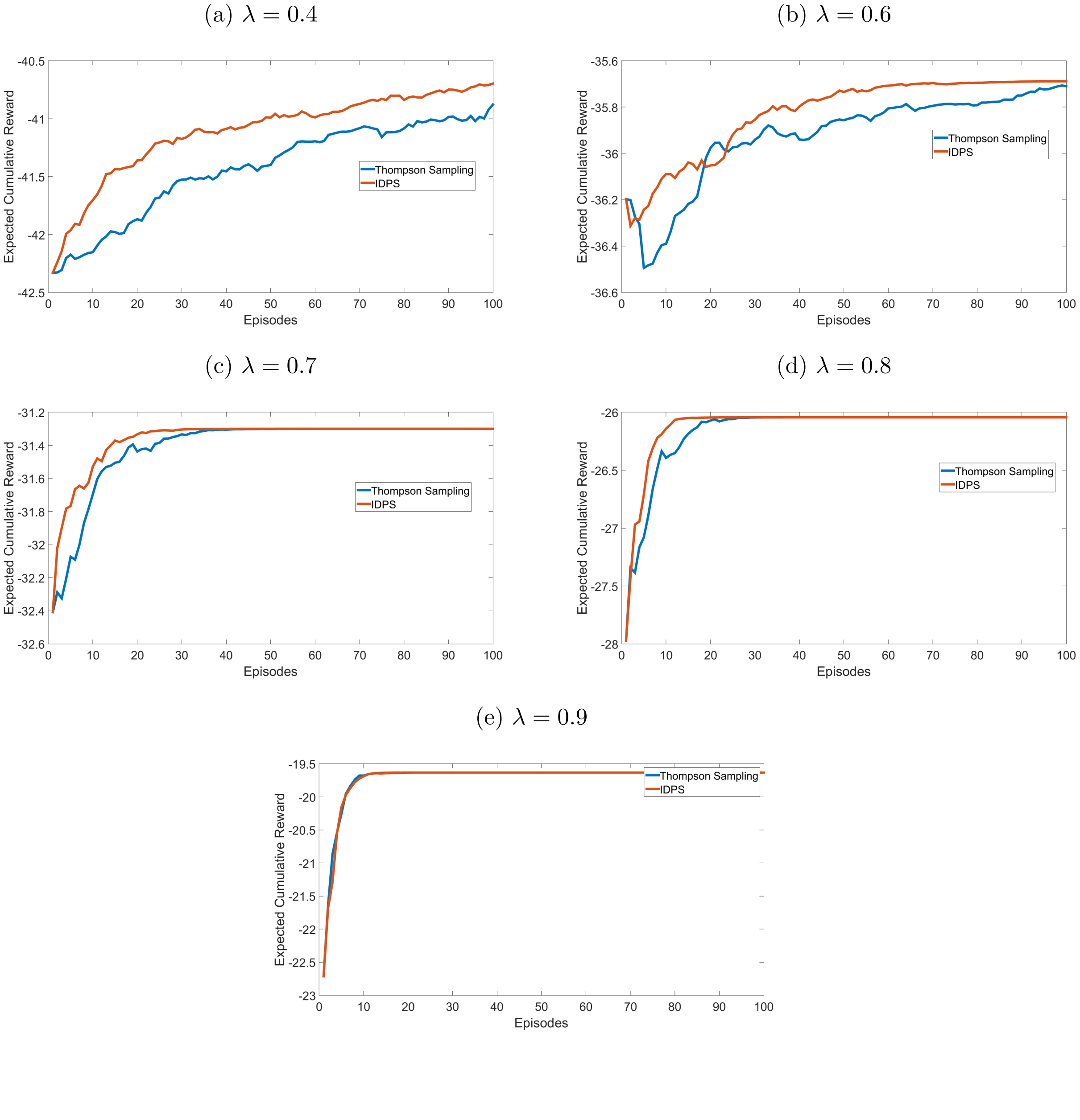
Dose d_t	Probability distribution over disease states X_{t+1} $P^X(X_{t+1} \mid X_t, d_t)$			Probability distribution over side effects Y_{t+1} $P^Y(Y_{t+1} \mid Y_t, d_t)$		
	$X_{t+1} = X_t - 1$	$X_{t+1} = X_t$	$X_{t+1} = X_t + 1$	$Y_{t+1} = Y_t - 1$	$Y_{t+1} = Y_t$	$Y_{t+1} = Y_t + 1$
$d_t > 0$	$0.7 + 0.3\frac{d_t}{d}$	$0.3 - 0.3\frac{d_t}{d}$	0	0	$1 - \lambda\frac{d_t}{d}$	$\lambda\frac{d_t}{d}$
$d_t = 0$	0	0.7	0.3	0.7	0.3	0

Figure 2.2 plots the cumulative reward averaged over 50 runs for IDPS and Thompson Sampling. Figure 1.5 plots the posterior of the true parameter values for IDPS and Thomposon Sampling. We observe that IDPS learns faster than Thompson Sampling approaches which is consistent with the objective of this research. An interesting observation is that the gap between learning rates for IDPS and Thompson Sampling decreases as parameter λ approaches 1. This is because as $\lambda \to 1$, the effect of dose level on side effect increases and it becomes prohibitive to use higher doses, irrespective of the disease condition. As such, the decision maker is better-off selecting the lowest nonzero dose in all states for both methods.

1.5 *Conclusion*

This chapter introduced an information theoretic sampling method for MDPs with parametric uncertainty. Problems with complex information structure requiring careful balancing of the trade-off between exploration and exploitation could potentially benefit from IDPS. We were able to extend the IDS analysis on bandit problems from Russo and Van Roy (2014, 2017) to the more difficult case of IDPS for MDPs. In particular, we were able to obtain a worst-case regret bound. The specific form of this bound provides insight to the decision-maker. Our numerical experiments suggest that IDPS could be better at learning and reward maximization as compared to Thompson Sampling on some problems. This chapter focused on a theoretical analysis of an idealized version of IDPS. Its exact implementation may be intractable when the set $\mathcal{P}$ of policies is large. Future work could investigate methods for

Figure 1.4: Expected cumulative rewards for different true parameter values for response-guided dosing from Section 1.4.3.



(a) $\lambda = 0.4$

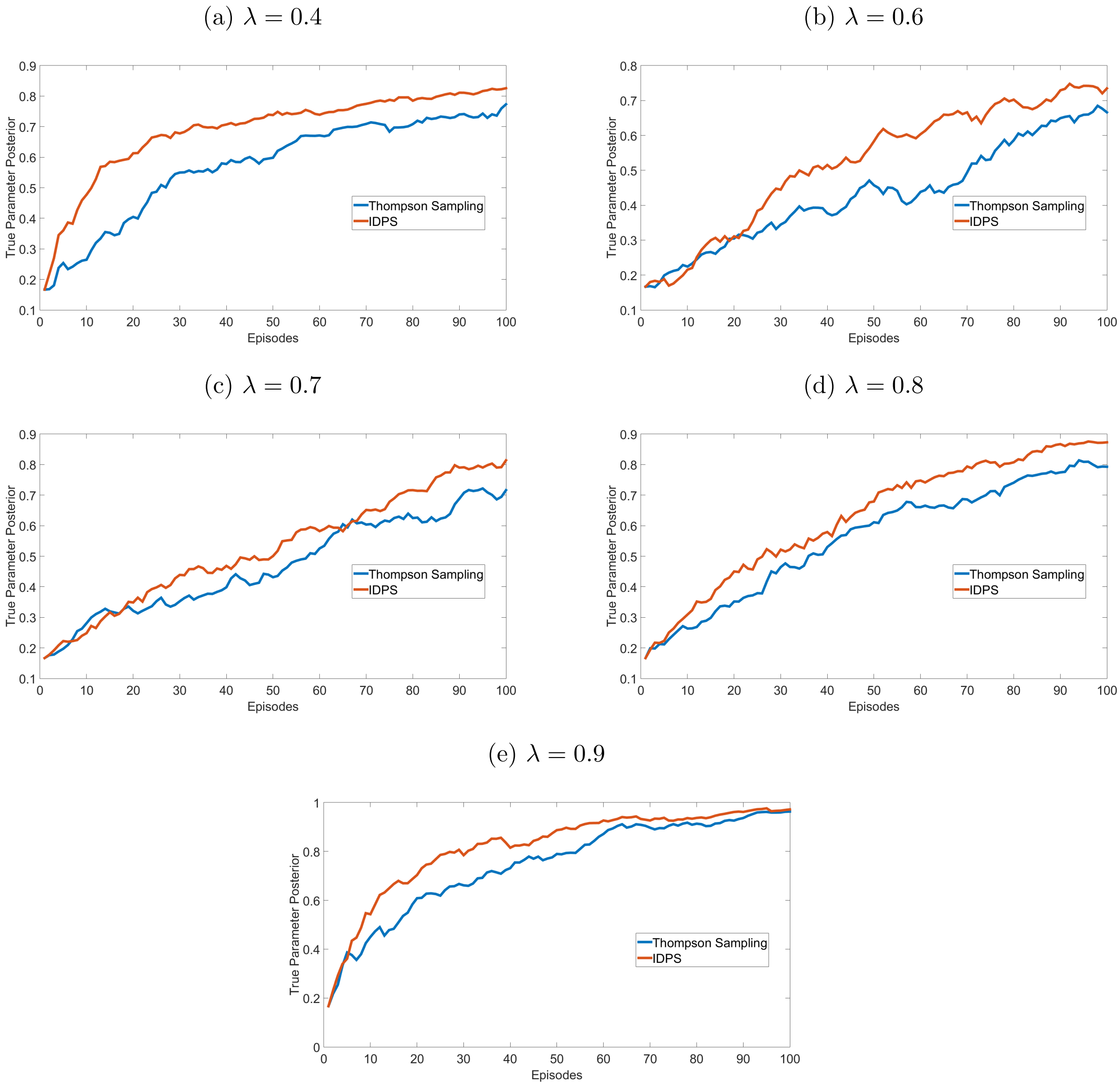
(b) $\lambda = 0.6$

(c) $\lambda = 0.7$

(d) $\lambda = 0.8$

(e) $\lambda = 0.9$

Figure 1.5: Posterior belief about the true parameter value for the response-guided dosing problem in Section 1.4.3.



(a) $\lambda = 0.4$

(b) $\lambda = 0.6$

(c) $\lambda = 0.7$

(d) $\lambda = 0.8$

(e) $\lambda = 0.9$

approximate implementation in such cases.

Chapter 2

LEARNING IN MDPS WITH HIERARCHICAL PARAMETRIC UNCERTAINTY

2.1 *Introduction*

I again consider MDPs of the form $Q = (S, A, T, R, N)$ as in Chapter 1. The parametric uncertainty in the transition probability function T here takes a hierarchical form. Let Θ be a finite set whose elements index a family T_θ, for $\theta \in \Theta$, of possible transition functions. This is the top-level parametrization in my MDP. For each $\theta \in \Theta$, the transition function T_θ is further parametrized by a parameter $\lambda \in \Lambda_\theta$, where Λ_θ is also a finite set. This is the bottom-level parametrization in my MDP. The true transition function is thus characterized by a pair (θ^*, λ^*) such that $\theta^* \in \Theta$ and $\lambda^* \in \Lambda_{\theta^*}$. The decision-maker knows that the true transition function is parametrized in this hierarchical fashion. It also knows all relevant finite sets of possible parameters, but does not know the pair (θ^*, λ^*) that identifies the true transition function. The family of all possible MDPs is denoted by tuples $\mathcal{M}_{\theta,\lambda} = (S, A, T_{\theta,\lambda}, R, N)$ with $\theta \in \Theta$ and $\lambda \in \Lambda_\theta$. I assume, as in Chapter 1, that the decision-maker can (easily) solve each MDP $\mathcal{M}_{\theta,\lambda}$. The decision-maker begins with a prior belief pmf $h_1(\cdot)$ on θ and conditional prior belief pmfs $\ell_1(\cdot|\theta)$, for $\theta \in \Theta$, on λ. These belief pmfs are updated via Bayes' Theorem as states drawn from the true transition function T_{θ^*,λ^*} are observed starting from an initial state $s_1 \in S$. The decision-maker's objective is to simultaneously learn the true transition function while maximizing expected reward.

2.1.1 *Existing literature*

Hierarchical Bayesian approaches have been employed for modeling and learning in several applications. Examples include veterinary studies (Stryhn and Christensen, 2014); sea level studies (Cahill et al., 2015); navigation tasks (Wilson et al., 2007); human performance studies (Kruschke and Vanpaemel, 2015); biotechnology (Broët et al., 2002); and operations management (Yeh, 1985; Qian and Wu, 2008). In some situations, characteristics of the particular application naturally lend themselves to a hierarchical structure. Furthermore, Bayesian methods can leverage hierarchical model structure for better learning as compared to flat models. (Chipman and McCulloch, 2000) assert that hierarchical models provide more representation power with efficient estimation performance. Hierarchical structures are often used to model complex problems by composing simple models together. Such interleaving can facilitate more complex model-fitting with the same data without over-fitting (Schmid and Brown, 2000; Allenby et al., 2005).

Hierarchical models have been used in an entirely different context within the MDP literature. That work focuses on temporal abstraction to utilize the hierarchical structure of the problem (Barto and Mahadevan, 2003; Dietterich, 2000; Hauskrecht et al., 1998; Sutton et al., 1999; He et al., 2011; Lim et al., 2011; Pineau and Thrun, 2002; Theocharous and Kaelbling, 2003; Cao and Ray, 2012; Vien et al., 2016). These works show a reduction in the difficulty of finding an optimal policy. The only other work I know is by Wilson et al. (2007), where a hierarchical Bayesian learning approach is employed to transfer knowledge between MDPs. That research showed how hierarchical models can speed up convergence to an optimal policy.

Many sequential decision-making and learning problems naturally call for a hierarchical modeling framework. For instance, in the MDP for response-guided dosing in Kotas and Ghate (2016), a patient's response to a therapeutic drug dose may be characterized by one of several options: the Michaels-Menten function, the exponential linear-quadratic function, the power law function, etc. Each such function in turn has a key parameter characterizing (say) low responders, moderate responders, and good responders. A doctor may wish to sequentially select dose levels while learning the response function and its parameter for an individual patient or a cohort of patients in a clinical trial (Kotas and Ghate, 2015). Similarly, in dynamic pricing problems, the price-demand function comes from several standard possibilities, and each function has a key parameter (Bertsimas and Perakis, 2006). To the best of my knowledge, there is currently no research that formulates or solves hierarchical learning problems in MDPs.

2.1.2 Contribution of this chapter

The first contribution of this chapter is to describe and formulate an MDP with hierarchical parametric uncertainty. The second contribution is the development of a Bayesian framework for updating the prior based on state observations in this hierarchical setting. The third contribution utilizes this Bayesian framework to extended my IDPS framework from Chapter 1 to the hierarchical setting. This uses formulas for the information ratio that I developed by exploiting the hierarchical structure. I have also devised a Thompson Sampling algorithm for the hierarchical framework. I have implemented Thompson Sampling and IDPS on a response-guided dosing problem from Ghate (2015). My numerical experiments on Thompson Sampling and IDPS hint at the potential benefit of using a hierarchical modeling framework as opposed to a "flat" one. I also provide a theoretical analyses of these learning algorithms to provably demonstrate this advantage. Specifically, I rederive the Regret bounds for the hierarchical case for IDPS and show that it is much stronger than the equivalent flat case.

2.2 Learning algorithms

In this section, I propose two algorithms for learning in MDPs with hierarchical parametric uncertainty. The first is a simple and natural extension of Thompson Sampling from Gopalan and Mannor (2015) and Strens (2000). The second is an extension of my IDPS framework from Chapter 1. Before I delve into the details of these two algorithms, I now describe how Bayes' Theorem can be applied to update priors $h_1(\cdot)$ and $\ell_1(\cdot|\theta)$ at the top and bottom levels, respectively.

Let $P(\theta, \lambda|s, a, s', h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\})$ denote the decision-maker's posterior belief that the true MDP is $\mathcal{M}_{\theta,\lambda}$ after the decision-maker observes new state s' by choosing action a in state s, given that the decision-maker's prior beliefs at the top and bottom levels were $h_t(\cdot)$ and $\{l_t(\cdot|\theta)|\theta \in \Theta\}$. Note that

$$\begin{aligned}
&P(\theta, \lambda | s, a, s', h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\}) \\
&\ = P(\lambda|\theta, s, a, s', h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\}) P(\theta | s, a, s', h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\}).
\end{aligned}$$

The first term in the above RHS is the decision-maker's posterior belief $\ell_{t+1}(\lambda|\theta)$ at the bottom level. The second term equals the posterior belief $h_{t+1}(\theta)$ at the top level. These two terms are individually calculated below. For the first term,

$$\begin{aligned}
\ell_{t+1}(\lambda|\theta) &= P(\lambda|\theta, s, a, s', h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\}) \\
&\propto P(s'|\theta, \lambda, s, a, h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\}) P(\lambda|\theta, s, a, h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\}) \\
&= P(s'|\theta, \lambda, s, a) P(\lambda|\theta, \{l_t(\cdot|\theta)|\theta \in \Theta\}) \\
&= T_{\theta,\lambda}(s'|s, a) l_t(\lambda|\theta).
\end{aligned}$$

Now for the second term,

$$\begin{aligned}
h_{t+1}(\theta) &= P(\theta|s, a, s', h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\}) \\
&\propto P(s'|\theta, s, a, h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\}) P(\theta|s, a, h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\}) \\
&= P(s'|\theta, s, a, h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\}) P(\theta|h_t(\cdot)) \\
&= h_t(\theta) P(s'|\theta, s, a, h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\}) \\
&= h_t(\theta) \sum_{\lambda \in \Lambda_\theta} P(s'|\theta, \lambda, s, a, h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\}) P(\lambda|\theta, s, a, h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\}) \\
&= h_t(\theta) \sum_{\lambda \in \Lambda_\theta} P(s'|\theta, \lambda, s, a) P(\lambda|\theta, \{l_t(\cdot|\theta)|\theta \in \Theta\}) \\
&= h_t(\theta) \sum_{\lambda \in \Lambda_\theta} T_{\theta,\lambda}(s'|s, a) \ell_t(\lambda|\theta).
\end{aligned}$$

This derivation shows that the tuple $(s, h_t(\cdot), \{l_t(\cdot|\theta)|\theta \in \Theta\})$ is a sufficient statistic for this problem of imperfect information in the sense of Bertsekas (2005). Although, algorithmically speaking, l_{t+1} is available while computing h_{t+1}. I will hence use l_{t+1} instead of l_t. This can be developed using formal algebraic computation but I will skip this derivation for brevity. Although qualitatively, this can be derived along the same lines as the above derivation except we consider one step lookahead as well. Also note that exact calculation of posterior beliefs $h_{t+1}(\cdot)$ and $\ell_{t+1}(\cdot|\theta)$ would in general be difficult owing to the normalization constants in the denominators that are not displayed in the above derivation. Note, however, that since I have assumed S, Θ, and Λ_θ to be finite sets, these exact calculations can be performed in principle.

2.2.1 Thompson Sampling

In this version of Thompson Sampling, an MDP $\mathcal{M}_{\theta_t,\lambda_t}$ is obtained in state s_t at stage t by sampling θ_t from $h_t(\cdot)$ and then sampling λ_t from $\ell(\cdot|\theta_t)$. Let π_t be a policy that is optimal in stage t for $\mathcal{M}_{\theta_t,\lambda_t}$. Then, an action $\pi_t(s_t)$ is implemented and state s_{t+1} is observed according to the true transition function $T_{\theta^*,\lambda^*}(\cdot|s_t,\pi_t(s_t))$. The priors are updated to $h_{t+1}(\cdot)$ and $\ell_{t+1}(\cdot|\theta)$ for $\theta \in \Theta$ according to the Bayes' formulas derived above. This is repeated until the end of stage N. This procedure is summarized in Algorithm 2 below.

2.2.2 IDPS

To extend IDPS to the hierarchical setting, define expected regrets, information gains, and information ratios at the top and bottom levels separately. Expressions are defined in a manner similar to Chapter 1.

For each fixed $\theta \in \Theta$, the bottom-level expected regret is defined as

$$\Delta_\theta(\pi|s_t,\ell(\cdot|\theta)) = \underset{\lambda\sim\ell(\cdot|\theta)}{E}\left[\underset{\{s_{\ell+1}\sim T_{\theta,\lambda}(\cdot|s_\ell,\pi_\ell(s_\ell)):\ell=t,\ldots,N\}}{E}\Big[V^*(\theta,\lambda)-V(\theta,\lambda,\pi)\Big]\right]. \tag{2.1}$$

Algorithm 2 Thompson Sampling for MDPs with hierarchical parametric uncertainty

Require: MDPs $\mathcal{M}_{\theta,\lambda} = \{S, A, T_{\theta,\lambda}, R, N\}$ for $\theta \in \Theta$, and $\lambda \in \Lambda_\theta$. Prior pmf $h_1(\cdot)$ for the top-level and and prior bottom-level conditional pmfs $\ell_1(\cdot|\theta)$ for $\theta \in \Theta$. Initial state $s_1 \in S$.

```
function HIERARCHICAL-TS
    for episode $k = 1, 2, 3, \cdots$ do
        Set $t = 1$
        Initialize state $s_1$; $h_1(\cdot) \leftarrow h_{N+1}(\cdot)$ and $\ell_1(\cdot|\theta) \leftarrow \ell_{N+1}(\cdot|\theta)$ for $\theta \in \Theta$, if $k > 1$
        repeat
            Sample $\theta_t \sim h_t(\cdot)$ and $\lambda_t \sim \ell_t(\cdot|\theta_t)$
            Let $\pi_t$ = policy optimal to $\mathcal{M}_{\theta_t,\lambda_t}$ in stage $t$.
            Observe $s_{t+1}$ drawn from the true distribution $T_{\theta^*,\lambda^*}(\cdot|s_t, \pi_t(s_t))$
            Update $\ell_{t+1}(X_\lambda|\theta) \propto T_{\theta,\lambda}(s_{t+1}|s_t, \pi_t(s_t))\ell_t(\lambda|\theta)$
            Update $h_{t+1}(\theta) \propto h_t(\theta) \sum_{\lambda\in\Lambda_\theta} T_{\theta,\lambda}(s_{t+1}|s_t, \pi_t(s_t))\ell_{t+1}(X_\lambda|\theta)$
            t ← t+1
        until end of horizon $N$
    end for
end function
```

Similarly, for each fixed $\theta \in \Theta$, the bottom-level information gain is given by

$$g_\theta(\pi|s_t, \ell_t(\cdot|\theta)) = \sum_{\lambda\in\Lambda_\theta}\sum_{s_t,\dots,s_N}\left(\prod_{\ell=t}^{N} T_\lambda(s_{\ell+1}|s_\ell, \pi_\ell(s_\ell))\right) \ell_t(\lambda|\theta)\ln\left[\frac{\prod\limits_{t=1}^{N} T_{\theta,\lambda}(s_{t+1}|s_t, \pi_t(s_t))}{\sum\limits_{\lambda\in\Lambda_\theta}\prod\limits_{t=1}^{N} T_{\theta,\lambda}(s_{t+1}|s_t, \pi_t(s_t))\ell_1(\lambda|\theta)}\right]. \tag{2.2}$$

After taking expectations with respect to pmf $\nu \in \mathcal{N}$ over $\mathcal{P}$, where $\mathcal{N}$ is the set of all possible pmfs at the lower levels. I obtain

$$\Delta_\theta(\nu) = \sum_{\pi\in\mathcal{P}} \nu(\pi)\Delta_\theta(\pi|s, \ell(\cdot|\theta)), \tag{2.3}$$

and

$$g_\theta(\nu) = \sum_{\pi \in \mathcal{P}} \nu(\pi) g_\theta(\pi | s, \ell(\cdot|\theta)). \tag{2.4}$$

For brevity and without loss of any generality, I am suppressing the dependence on $\ell(\cdot|\theta)$ and s_ℓ.

Similar to Chapter 1, the bottom-level information ratio is given by

$$\Psi_\theta(\nu) = \frac{(\Delta_\theta(\nu))^2}{g_\theta(\nu)}. \tag{2.5}$$

At the top-level, I define μ as the distribution over the lower level pmfs $\nu \in \mathcal{N}$ the expected regret and information gain are defined as

$$\Delta(\mu) = \sum_{\gamma \in \Theta} \mu(\nu_\gamma^*) \underset{\theta \sim h(\cdot)}{E} \left[\Delta_\theta(\nu_\gamma^*)\right], \tag{2.6}$$

and

$$g(\mu) = \sum_{\gamma \in \Theta} \mu(\nu_\gamma^*) \underset{\theta \sim h(\cdot)}{E} \left[g_\theta(\nu_\gamma^*)\right], \tag{2.7}$$

where $\nu_\gamma^* \in \{\arg\min_\nu \Psi_\gamma(\nu) \; \forall \gamma \in \Theta\}$,is defined as the randomized policy that optimizes the lower level information ratio for each element in Θ. Quantities $\Delta_\theta(\cdot)$ and $g_\theta(\cdot)$ are obtained from (2.3), (2.4), respectively. This yields the top-level information ratio

$$\Psi(\mu) = \frac{(\Delta(\mu))^2}{g(\mu)}. \tag{2.8}$$

IDPS separately minimizes the above two information ratios to sample policies at each stage. The resulting procedure is summarized in Algorithm 3 below.

The next theorem bounds the decision-maker's cumulative expected regret over N stages for hierarchical IDPS.

The rest of the entities are defined similar to the flat case in Theorem 1.3.3. For the

Algorithm 3 Information Directed Policy Sampling for MDPs with hierarchical parametric uncertainty

Require: MDPs $\mathcal{M}_{\theta,\lambda} = \{S, A, T_{\theta,\lambda}, R, N\}$ for $\theta \in \Theta$, and $\lambda \in \Lambda_\theta$. Prior pmf $h_1(\cdot)$ for the top-level and and prior bottom-level conditional pmfs $\ell_1(\cdot|\theta)$ for $\theta \in \Theta$. Initial state $s_1 \in S$.

```
function Hierarchical-IDPS
    for episode $k = 1, 2, 3, \cdots$ do
        Set $t = 1$
        Initialize state $s_1$; $h_1(\cdot) \leftarrow h_{N+1}(\cdot)$ and $\ell_1(\cdot|\theta) \leftarrow \ell_{N+1}(\cdot|\theta)$ for $\theta \in \Theta$, if $k > 1$
        repeat
            Compute the distribution $\nu_\theta^t = \text{argmin}_\nu \Psi_\theta^t(\nu|s)$ using (2.5) where $\Delta_{\theta,t}(\nu|s)$ and $g_{\theta,t}(\nu|s)$ are defined in (2.3) $\forall\theta$
            Compute the distribution $\mu_t = \text{argmin}_\mu \Psi(\mu)$ using (2.8) where $\Delta(\mu|s)_t$ and $g(\mu|s)_t$ are defined in (2.6) $\forall\theta$
            Sample $\nu_t \sim \mu_t$
            Sample $\pi_t \sim \nu_t$
            Observe $s_{t+1}$ drawn from the true distribution $T_{\lambda^*}(\cdot|s_t, \pi_t(s_t))$
            Update $\ell_{t+1}(\lambda|\theta) \propto T_{\theta,\lambda}(s_{t+1}|s_t, \pi_t(s_t))\ell_t(\lambda|\theta)$
            Update $h_{t+1}(\theta) \propto h_t(\theta) \sum_{\lambda \in \Lambda_\theta} T_{\theta,\lambda}(s_{t+1}|s_t, \pi_t(s_t))\ell_{t+1}(\lambda|\theta)$
            t $\leftarrow$ t+1
        until end of horizon $N$
    end for
end function
```

proof, I follow a similar structure as in Theorem 1.3.3; The main challenges that I overcame for the hierarchical case are a) Propagation of state observations to the entities defined on top levels, and b) Incorporating the hierarchical structure in the Bayesian learning part of the algorithm to bound overall regret.

Regret(N) for hierarchical MDP is defined as

$$Regret(N) = \sum_{t=1}^{N} \left(\underset{\substack{\{s_{\ell+1}:\ell=t+1:N\} \\ \theta,\lambda}}{E} \left[V^*(\theta,\lambda) - V(\pi^t,\theta,\lambda) - \underset{s_{t+1}}{E} \left[V^*(\theta,\lambda) - V(\pi^{t+1},\pi_t,\theta,\lambda) \right] \right] \right) \tag{2.9}$$

where $V^*(\theta,\lambda) = \sum_{\ell=t}^{N} R(s_{\ell+1}|s_\ell,\pi^*_{\ell,\theta,\lambda}(s_\ell))$ denote the random tail reward accumulated on implementing an optimal policy in MDP $\mathcal{M}_{\theta,\lambda}$. Similarly, $V(\theta,\lambda,\pi^t) = \sum_{\ell=t}^{N} R(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))$, for $s_{t+1} \sim T_{\theta^*,\lambda^*}(\cdot|s_t,\pi^t(s_t))$ and $\pi^t \sim \nu^t_*$. Here $\nu^t_* \in \{\arg\min_\nu \Psi_\gamma(\nu)|_t \ \forall\gamma \in \Theta\}$, and $\nu^t_* \sim \mu^t_*$, where $\mu^t_* = \arg\min_\mu \Psi(\mu)$. Also note that $\pi^t = \{\pi_t, \pi^{t+1}\}$. For the next proof, I also define $\Omega_{\theta_{max}} = \max_\theta\{\Omega_\theta \ \forall\theta \in \Theta\}$, which is the the largest parameter set at the lower level.

Theorem 2.2.1. Regret Bound: *Suppose there is a γ such that $\Psi^* \leq \gamma$. Then, for any $\epsilon > 0$, we have,*

$$Regret(N) \leq ((1+\sqrt{N-1})\sqrt{\epsilon}+1)\sqrt{\gamma\log(|\Omega_{\theta_{max}}|)} \quad \textit{with probability}\ (1-e^{-\frac{(c\epsilon-2)^2}{2N}})(1-e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}),$$

where $c > 2/\epsilon$ and $c_0 > 2/\sqrt{\epsilon}$ are positive constants.

Proof. This result is derived by building on Theorem 1.3.3 in Chapter 1. Consider

$$\begin{aligned}
Regret(N) = \sum_{t=1}^{N} \Bigg(& \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\} \\ \theta,\lambda}}{E} \Bigg[V^*(\theta,\lambda) - V(\pi^t,\theta,\lambda) \\
& - \underset{s_{t+1}}{E} \left[V^*(\theta,\lambda) - V(\pi^{t+1},\pi_t,\theta,\lambda) \right] \Bigg] \Bigg) \\
& \overset{a}{=} \sum_{t=1}^{N} \Bigg(\Delta_t(\mu_*^t|s_t) - \underset{s_{t+1}}{E} \Delta_{t+1}(\mu_*^t|s_{t+1}) \Bigg) \\
& \overset{b}{\leq} \sum_{t=1}^{N} \Bigg(\sqrt{\Psi_t^* \sum_{\gamma} \mu_t(\nu_\gamma^*) \underset{\theta}{E}[g_\theta(\nu_\gamma^*)]} - \underset{s_{t+1}}{E} \Bigg[\sqrt{\Psi_{t+1}^* \sum_{\gamma} \mu_t(\nu_\gamma^*) \underset{\theta}{E}[g_\theta(\nu_\gamma^*)]} \Bigg] \Bigg) \\
& \overset{c}{=} \sum_{t=1}^{N} \Bigg(\sqrt{\Psi_t^* \sum_{\gamma} \mu_t(\nu_\gamma^*) \underset{\theta}{E} \Bigg[H(\Lambda_{\theta,t}|s_t) - \underset{\substack{\{s_{\ell+1}:\ell=t:N\} \\ \pi^t\sim\nu^t}}{E} [H(\Lambda_{\theta,t}|\{s_{\ell+1} : \ell = t+1 : N\},\pi^t,s_t)] \Bigg]} \\
& - \underset{s_{t+1}}{E} \Bigg[\sqrt{\Psi_t^* \sum_{\gamma} \mu_t(\nu_\gamma^*) \underset{\theta}{E} \Bigg[H(\Lambda_{\theta,t}|s_{t+1}) - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\} \\ \pi^{t+1}\sim\nu^{t+1}}}{E} [H(\Lambda_{\theta,t}|\{s_{\ell+1} : \ell = t : N\},\pi^{t+1},s_{t+1})] \Bigg]} \Bigg] \\
& \overset{d}{=} \sum_{t=1}^{N} \Bigg(\sqrt{\Psi_{t+1}^* \sum_{\gamma} \mu_t(\nu_\gamma^*) \underset{\theta}{E} \Bigg[H(\Lambda_{\theta,t+1}|s_t) - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\} \\ \pi^t\sim\nu^t}}{E} [H(\Lambda_{\theta,t+1}|\{s_{\ell+1} : \ell = t : N\},\pi^t,s_t)] \Bigg]} \\
& - \underset{s_{t+1}}{E} \Bigg[\sqrt{\Psi_t^* \sum_{\gamma} \mu_t(\nu_\gamma^*) \underset{\theta}{E} \Bigg[H(\Lambda_{\theta,t+1}|s_{t+1}) - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\} \\ \pi^{t+1}\sim\nu^{t+1}}}{E} [H(\Lambda_{\theta,t}|\{s_{\ell+1} : \ell = t : N\},\pi^{t+1},s_{t+1})] \Bigg]} \Bigg] \\
& + \sqrt{\Psi_1^* \sum_{\gamma} \mu_t(\nu_\gamma^*) \underset{\theta}{E} \Bigg[H(\Lambda_{\theta,1}|s_1) - \underset{\substack{\{s_{\ell+1}:\ell=1:N\} \\ \pi^1\sim\nu^1}}{E} [H(\Lambda_{\theta,1}|\{s_{\ell+1} : \ell = 1 : N\},\pi^1,s_1)] \Bigg]} \\
& - \sqrt{\Psi_{N+1}^* \sum_{\gamma} \mu_t(\nu_\gamma^*) \underset{\theta}{E} \left[H(\Lambda_{\theta,N+1}|s_{N+1}) \right]},
\end{aligned}$$

where 'a' follows from the definition of expected regrets $\Delta_t(\mu_*^t|s_t)$ and $\Delta_{t+1}(\mu_*^t|s_{t+1})$. Inequality '$b$' follows from 2 ideas; The first is the equality that comes from the definition of the upper level regrets and the second one is the inequality that comes from the optimality condition $\Psi_{t+1} > \Psi_{t+1}^*$. Equality '$c$' follows from Lemma 1.3.1. Equality 'd' is just rearrangement of terms. Now the rest of the proof bound the first and the second part of the last equation and combine them together to arrive at the result. Consider the second part of the last equation first:

$$
\begin{aligned}
&\sum_{t=1}^{N-1} \underset{s_{t+1}}{E} \left[\sqrt{\Psi^*_{t+1} \sum_{\gamma} \mu_t(\nu^*_\gamma) \underset{\theta}{E}\Big[H(\Lambda_{\theta,t+1}|s_{t+1}) - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}[H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})]\Big]} \right] \\
&\overset{a}{=} \sum_{t=1}^{N-1} \underset{s_{t+1}}{E} \left[\sqrt{\underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E} \left[\Psi^*_{t+1} \sum_{\gamma} \mu_t(\nu^*_\gamma) \underset{\theta}{E}\Big[H(\Lambda_{\theta,t+1}|s_{t+1}) - H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})\Big]\right]} \right] \\
&\overset{b}{\geq} \sum_{t=1}^{N-1} \underset{s_{t+1}}{E} \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E} \left[\sqrt{\Psi^*_{t+1} \sum_{\gamma} \mu_t(\nu^*_\gamma) \underset{\theta}{E}\Big[H(\Lambda_{\theta,t+1}|s_{t+1}) - H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})\Big]} \right] \\
&\overset{c}{=} \sum_{t=1}^{N-1} \underset{\substack{\{s_{\ell+1}:\ell=t:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E} \left[\sqrt{\Psi^*_{t+1} \sum_{\gamma} \mu_t(\nu^*_\gamma) \underset{\theta}{E}\Big[H(\Lambda_{\theta,t+1}|s_{t+1}) - H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})\Big]} \right] \\
&\overset{d}{=} \sum_{t=1}^{N-1} \underset{\substack{\{s_{\ell+1}:\ell=t:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E} \left[Z_{t+1}\right], \qquad (2.10)
\end{aligned}
$$

where equality 'a' follows due to independence of $H(\Lambda_{\theta,t+1}|s_{t+1})$ with respect to the future trajectory $s_{\ell+1} : \ell = t+1 : N\}$, and policy $\pi^{t+1} \sim \nu^{t+1}$. Inequality 'b' follows from the Jensen's inequality and concavity of the square root function. Equality 'c' is just simplifying the expectation terms, and equality 'd' follows from definition of Z_{t+1}.

Now consider the first part of the above equation for $Regret(N)$:

$$
\begin{aligned}
&\sum_{t=1}^{N-1}\sqrt{\Psi^*_{t+1}\sum_{\gamma}\mu_t(\nu^*_\gamma)\underset{\theta}{E}\Big[H(\Lambda_{\theta,t+1}|s_{t+1})-\underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}[H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})]\Big]}\\
&=\sum_{t=1}^{N-1}\Bigg\{\Psi^*_{t+1}\sum_{\gamma}\mu_t(\nu^*_\gamma)\underset{\theta}{E}\Bigg[H(\Lambda_{\theta,t+1}|s_{t+1})-H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})+\\
&\quad H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})-\\
&\quad\underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}[H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})]\Bigg)\Bigg\}^{1/2}\\
&\overset{a}{\leq}\sum_{t=1}^{N-1}\Bigg(\sqrt{\Psi^*_{t+1}\sum_{\gamma}\mu_t(\nu^*_\gamma)\underset{\theta}{E}\Big[H(\Lambda_{\theta,t+1}|s_{t+1})-H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})\Big]}+\\
&\quad\Bigg\{\Psi^*_{t+1}\Big(H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})-\\
&\quad\underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}[H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})]\Big)\Bigg\}^{1/2}\Bigg)\\
&=\sum_{t=1}^{N-1}\Bigg(Z_{t+1}+\Bigg\{\Psi^*_{t+1}\sum_{\gamma}\mu_t(\nu^*_\gamma)\underset{\theta}{E}\Big[H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})-\\
&\quad\underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}[H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})]\Big]\Bigg\}^{1/2}\Bigg)\\
&\overset{b}{\leq}\sum_{t=1}^{N-1}Z_{t+1}+\\
&\quad\sqrt{N-1}\Bigg\{\sum_{t=1}^{N-1}\Psi^*_{t+1}\sum_{\gamma}\mu_t(\nu^*_\gamma)\underset{\theta}{E}\Bigg[H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})-\\
&\quad\underset{\substack{\{s_{\ell+1}:\ell=t+1:N\}\\ \pi^{t+1}\sim\nu^{t+1}}}{E}[H(\Lambda_{\theta,t+1}|\{s_{\ell+1}:\ell=t+1:N\},\pi^{t+1},s_{t+1})]\Bigg]\Bigg\}^{1/2}\\
&\quad\overset{c}{\leq}\sum_{t=1}^{N-1}Z_{t+1}+\sqrt{(N-1)|\epsilon\Psi^*_{t+1}\sum_{\gamma}\mu_t(\nu^*_\gamma)\underset{\theta}{E}[H(\Lambda_\theta)]|},\ \text{with probability } 1-e^{-\frac{(c\epsilon-2)^2}{2N}}\\
&\quad\overset{d}{\leq}\sum_{t=1}^{N-1}Z_{t+1}+\sqrt{(N-1)|\epsilon\Psi^*_{t+1}|\ |\sum_{\gamma}\mu_t(\nu^*_\gamma)\underset{\theta}{E}[H(\Lambda_\theta)]|},\ \text{with probability } 1-e^{-\frac{(c\epsilon-2)^2}{2N}}\\
&\overset{e}{\leq}\sum_{t=1}^{N-1}Z_{t+1}+\sqrt{(N-1)\epsilon\Psi^*\log(|\Omega_{\theta_{max}}|)},\ \text{with probability } 1-e^{-\frac{(c\epsilon-2)^2}{2N}}, \qquad (2.11)
\end{aligned}
$$

where 'a' follows from a simple algebraic inequality, 'b' follows from Cauchy-Schwarz inequality. Inequality 'c' follows from Hoeffding's inequality for Markov Chains, and inequality 'd' follows from the algebraic inequality $|ab| \leq |a||b|$.

Now combining Equations (2.11) and (2.10), we get

$$\begin{aligned}
&\sum_{t=1}^{N-1} Z_{t+1} + \sqrt{(N-1)\epsilon\Psi^* \log(|\Omega_{\theta_{max}}|)} - \sum_{t=1}^{N-1} \underset{\substack{\{s_{\ell+1}:\ell=t:N\} \\ \pi^{t+1}\sim\nu^{t+1}}}{E}[Z_{t+1}] \\
&= \sqrt{\epsilon||\Psi^*_{t+1} \sum_{\gamma} \mu_t(\nu^*_\gamma) \underset{\theta}{E}[H(\Lambda_\theta)]} \\
&+ \sqrt{(N-1)\epsilon\Psi^* \log(|\Omega_{\theta_{max}}|)}, \text{ with probability } (1 - e^{-\frac{(c\epsilon-2)^2}{2N}})(1 - e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}) \\
&= \sqrt{\epsilon\Psi^* \log(|\Omega_{\theta_{max}}|)} + \sqrt{(N-1)\epsilon\Psi^* \log(|\Omega_{\theta_{max}}|)}, \\
&\text{with probability } (1 - e^{-\frac{(c\epsilon-2)^2}{2N}})(1 - e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}).
\end{aligned}$$

Now combining all of this together, the regret bounds for hierarchical IDPS can be derived as:

$$\begin{aligned}
Regret(N) = \sum_{t=1}^{N} \Bigg(&\sqrt{\Psi^*_{t+1} \sum_{\gamma} \mu_t(\nu^*_\gamma) \underset{\theta}{E} \left[H(\Lambda_{\theta,t+1}|s_t) - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\} \\ \pi^t\sim\nu^t}}{E}[H(\Lambda_{\theta,t+1}|\{s_{\ell+1} : \ell = t : N\}, \pi^t, s_t)] \right]} \\
&- \underset{s_{t+1}}{E} \left[\sqrt{\Psi^*_t \sum_{\gamma} \mu_t(\nu^*_\gamma) \underset{\theta}{E} \left[H(\Lambda_{\theta,t+1}|s_{t+1}) - \underset{\substack{\{s_{\ell+1}:\ell=t+1:N\} \\ \pi^{t+1}\sim\nu^{t+1}}}{E}[H(\Lambda_{\theta,t}|\{s_{\ell+1} : \ell = t : N\}, \pi^{t+1}, s_{t+1})] \right]} \right] \\
&+ \sqrt{\Psi^*_1 \sum_{\gamma} \mu_t(\nu^*_\gamma) \underset{\theta}{E} \left[H(\Lambda_{\theta,1}|s_1) - \underset{\substack{\{s_{\ell+1}:\ell=1:N\} \\ \pi^1\sim\nu^1}}{E}[H(\Lambda_{\theta,1}|\{s_{\ell+1} : \ell = 1 : N\}, \pi^1, s_1)] \right]} \\
&- \sqrt{\Psi^*_{N+1} \sum_{\gamma} \mu_t(\nu^*_\gamma) \underset{\theta}{E}[H(\Lambda_{\theta,N+1}|s_{N+1})]} \\
&\leq (1 + \sqrt{N-1})\sqrt{\epsilon\Psi^* \log(|\Omega_{\theta_{max}}|)} + \sqrt{\gamma H(\Lambda_{\theta,1}|s_1)} \\
&\overset{\text{a}}{\leq} ((1 + \sqrt{N-1})\sqrt{\epsilon} + 1)\sqrt{\Psi^* \log(|\Omega_{\theta_{max}}|)}
\end{aligned}$$

with probability $(1 - e^{-\frac{(c\epsilon-2)^2}{2N}})(1 - e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}})$. □

Here inequality 'a' follows by bounding Shanon entropy and choosing the maximum possible bound from the lower levels. The next step is to find the upper bound for the optimal top level information ratio Ψ^*. I derive that next in Proposition 2.2.2.

Proposition 2.2.2. Bounds on minimal information ratio: *The minimal information ratio is bounded above by $|\Omega_{\theta_{max}}|/2$.*

Proof. The information gain in (1.2) can be expressed as

$$g_t(\pi^t|s_t) = \underset{\lambda\sim\alpha_t(\cdot)}{E}\left\{D_{KL}\left[\left(\prod_{\ell=t}^{N} T_{\theta,\lambda}(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))\right)\middle|\middle|\sum_{\lambda\in\Omega}\prod_{\ell=t}^{N} T_{\theta,\lambda}(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))\alpha_t(\lambda)\right]\right\}.$$

Define

$$U_t(\pi^t,\theta,\lambda) = \underset{\{s_{\ell+1}\sim T_{\theta,\lambda}(\cdot|s_\ell,\pi_\ell(s_\ell)):\ell=t,\ldots,N\}}{E}[V_t(\theta,\lambda,\pi^t)],$$

and

$$L_t(\pi^t,\theta,\lambda) = U_t(\pi^t,\theta,\lambda) - \underset{\lambda\sim l_{\theta,t}(\cdot)}{E}\left[U_t(\pi^t,\theta,\lambda)\right].$$

Further define

$$\Psi_t^L(\mu^t) = \frac{\Delta_t(\mu_{PS}^t)^2}{\underset{\nu_\gamma\sim\mu_{PS}^t(\cdot)}{E}\ \underset{\theta\sim h_t(\cdot)}{E}\ \underset{\pi\sim\nu_\gamma(\cdot)}{E}\ \underset{\lambda\sim l_{\theta,t}(\cdot)}{E}[L_t(\pi^t,,\theta,\lambda)^2]},$$

and

$$\nu_L^t = \underset{\mu^t\in D^t}{\operatorname{argmin}}\,\Psi_t^L(\mu^t).$$

At the optimal solution ν_{IDPS}^t to the information ratio minimization problem, we have, $\Psi_t(\nu_{IDPS}^t) \leq \Psi_t(\nu_L^t)$.

$$\begin{aligned}\Psi_t(\mu^t) &= \frac{\Delta_t(\mu^t)^2}{\underset{\nu_\gamma\sim\mu^t(\cdot)}{E}\underset{\theta\sim h_t(\cdot)}{E}\underset{\pi\sim\nu_\gamma(\cdot)}{E}\underset{\lambda\sim l_{\theta,t}(\cdot)}{E}\left\{D_{KL}\left[\left(\prod_{\ell=t}^{N} T_{\theta,\lambda}(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))\right)\middle\|\sum_{\lambda\in\Omega}\prod_{\ell=t}^{N} T_{\theta,\lambda}(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))\alpha_t(\lambda)\right]\right\}}\\ &\overset{\text{a}}{\leq}\frac{\Delta_t(\mu^t)^2}{2\underset{\nu_\gamma\sim\mu^t(\cdot)}{E}\underset{\theta\sim h_t(\cdot)}{E}\underset{\pi\sim\nu_\gamma(\cdot)}{E}\underset{\lambda\sim l_{\theta,t}(\cdot)}{E}[U_t(\pi^t,\theta,\lambda)-\underset{\lambda\sim l_{\theta,t}(\cdot)}{E}[U_t(\pi^t,\theta,\lambda)]]^2}\\ &=\frac{1}{2}\Psi_t^L(\mu^t). \end{aligned} \tag{2.12}$$

Here, "a" follows from Fact 1.3.4 with distribution P identified as $P = \prod_{\ell=t}^{N} T_{\theta,\lambda}(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))$ and distribution Q as $Q = \sum_{\lambda\in\Omega}\prod_{\ell=t}^{N} T_{\theta,\lambda}(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))\alpha_t(\lambda)$ and $X = U_t(\pi^t,\theta,\lambda)$. We assume that $\sup(U_t(\pi^t,\theta,\lambda)) - \inf(U_t(\pi^t,\theta,\lambda)) \leq 1$. For a finite MDP, stage-rewards and hence value functions are uniformly bounded. Thus, the assumption $\sup(U_t(\pi^t,\theta,\lambda)) - \inf(U_t(\pi^t,\theta,\lambda)) \leq 1$ holds without any loss of generality. This is because if $\sup(U_t(\pi^t,\theta,\lambda)) - \inf(U_t(\pi^t,\theta,\lambda)) > 1$, we can rescale the rewards such that the new rewards $R(s,a) \leftarrow R(s,a)/(\sup(U_t(\pi^t,\theta,\lambda)) - \inf(U_t(\pi^t,\theta,\lambda)))$. This rescaling does not affect any decision rules for the MDP but ensures that $\sup(U_t(\pi^t,\theta,\lambda)) - \inf(U_t(\pi^t,\theta,\lambda)) \leq 1$ making the theorem valid. Also note that the distribution P is absolutely continuous with respect to Q because $Q(\cdot) = 0$ implies that $P(\cdot) = 0$. By definition $Q = \sum_{\lambda\in\Omega}\prod_{\ell=t}^{N} T_{\theta,\lambda}(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))\alpha_t(\lambda)$ and $\prod_{\ell=t}^{N} T_{\theta,\lambda}(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell))\alpha_t(\lambda) \geq 0;\ \forall\lambda$. For the sum of non-negative components to be 0 it is required that each individual component should be 0. Thus, either $\prod_{\ell=t}^{N} T_{\theta,\lambda}(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell)) = 0$ or $\alpha_t(\lambda) = 0$. Since this is true for any arbitrary α_t and $\sum_{\lambda\in\Omega}\alpha_t(\lambda) = 1$, it implies that $\prod_{\ell=t}^{N} T_{\theta,\lambda}(s_{\ell+1}|s_\ell,\pi_\ell(s_\ell)) = 0$, which corresponds to $P(\cdot) = 0$. This yields

$$\Psi_t(\nu^t_{IDPS}) \leq \Psi_t(\nu^t_L) \leq \Psi^L_t(\nu^t_L)/2. \tag{2.13}$$

Now let μ^t_{PS} be the distribution over the policies, where each policy is optimal with respect to the MDP $\mathcal{Q}_\lambda$ obtained from sampling the corresponding parameter λ from the posterior distribution α_t. Hence by definition, the distribution ν^t_{PS} is same as α_t with the identification that the domain consists of optimal policies of M_λ, for all $\lambda \in \Omega$. Therefore,

$$\underset{\nu_\gamma \sim \mu^t(\cdot)}{E}\ \underset{\theta \sim h_t(\cdot)}{E}\ \underset{\pi \sim \nu_\gamma(\cdot)}{E}\ \underset{\lambda \sim l_{\theta,t}(\cdot)}{E}[L_t(\nu^t_{PS}, \lambda)^2] = \underset{\substack{\lambda \sim \alpha_t(\cdot) \\ \pi \sim \nu^t_{PS}}}{E}[L_t(\pi^t, \theta, \lambda)^2] \overset{\text{a}}{=} \underset{\substack{\lambda \sim \alpha_t(\cdot) \\ \pi \sim \alpha_t(\cdot)}}{E}[L_t(\pi^t, \theta, \lambda)^2]. \tag{2.14}$$

Equality “a” holds because of the aforementioned equivalence between sampling policies from ν^t_{PS} and from $\alpha_t(\cdot)$. Similarly,

$$\begin{aligned}
\Delta_t(\mu^t_{PS}) &= \underset{\nu_\gamma \sim \mu^t(\cdot)}{E}\ \underset{\theta \sim h_t(\cdot)}{E}\ \underset{\pi \sim \nu_\gamma(\cdot)}{E}[\Delta_{\theta,t}(\pi^t)] \\
&= \underset{\nu_\gamma \sim \mu^t_{PS}(\cdot)}{E}\ \underset{\theta \sim h_t(\cdot)}{E}\ \underset{\pi \sim \nu_\gamma(\cdot)}{E}\ \underset{\lambda \sim l_{\theta,t}(\cdot)}{E}\ \underset{\{s_{\ell+1} \sim T_{\theta,\lambda}(\cdot|s_\ell, \pi_\ell(s_\ell)) : \ell = t,\ldots,N\}}{E}\Big[V^*_t(\theta, \lambda) - V_t(\theta, \lambda, \pi^t)\Big] \\
&\overset{\text{b}}{=} \underset{\nu_\gamma \sim h_t(\cdot)}{E}\ \underset{\theta \sim h_t(\cdot)}{E}\ \underset{\pi \sim \nu_\gamma(\cdot)}{E}\left[\underset{\{s_{\ell+1} \sim T_{\theta,\lambda}(\cdot|s_\ell, \pi_\ell(s_\ell)) : \ell = t,\ldots,N\}}{E}\Big[V_t(\theta, \lambda, \pi^t)\Big]\right. \\
&\left. - \underset{\lambda \sim l_{\theta,t}(\cdot)}{E}\Big[\underset{\{s_{\ell+1} \sim T_{\theta,\lambda}(\cdot|s_\ell, \pi_\ell(s_\ell)) : \ell = t,\ldots,N\}}{E}\Big[V_t(\theta, \lambda, \pi^t)\Big]\Big]\right] \\
&\overset{\text{c}}{=} \underset{\theta \sim h_t(\cdot)}{E}\ \underset{l_{\theta,t} \sim h_t(\cdot)}{E}\ \underset{\pi \sim l_{\theta,t}(\cdot)}{E}\left[\underset{\{s_{\ell+1} \sim T_{\theta,\lambda}(\cdot|s_\ell, \pi_\ell(s_\ell)) : \ell = t,\ldots,N\}}{E}\Big[V_t(\theta, \lambda, \pi^t)\Big]\right. \\
&\left. - \underset{\lambda \sim l_{\theta,t}(\cdot)}{E}\Big[\underset{\{s_{\ell+1} \sim T_{\theta,\lambda}(\cdot|s_\ell, \pi_\ell(s_\ell)) : \ell = t,\ldots,N\}}{E}\Big[V_t(\theta, \lambda, \pi^t)\Big]\Big]\right] \\
&= \underset{\theta \sim h_t(\cdot)}{E}\ \underset{l_{\theta,t} \sim h_t(\cdot)}{E}\ \underset{\pi \sim l_{\theta,t}(\cdot)}{E}[L_t(\pi^t, \theta, \lambda)] \\
&\overset{\text{d}}{\leq} \underset{\theta \sim h_t(\cdot)}{E}\ \underset{l_{\theta,t} \sim h_t(\cdot)}{E}\sqrt{|\Omega_\theta|\left[\sum_{\pi \sim l_{\theta,t}} (l_{\theta,t} L_t(\pi^t, \theta, \lambda))^2\right]} \\
&= \underset{\theta \sim h_t(\cdot)}{E}\sqrt{|\Omega_\theta|}\ \underset{l_{\theta,t} \sim h_t(\cdot)}{E}\sqrt{\underset{\substack{\pi \sim l_{\theta,t}(\cdot) \\ \lambda \sim l_{\theta,t}}}{E}[L_t(\pi^t, \theta, \lambda)]^2}
\end{aligned}$$

Equality “‘b”’ holds due to an extension of the logic behind equality “a”. Let $\mathcal{P}^*$ be the set of policies optimal for MDPs $\mathcal{M}_\lambda$, $\forall \lambda \in \Omega$. Now recall that

$$\Psi^L_t(\mu^t) = \frac{\Delta_t(\mu^t_{PS})^2}{\underset{\nu_\gamma \sim \mu^t_{PS}(\cdot)}{E}\ \underset{\theta \sim h_t(\cdot)}{E}\ \underset{\pi \sim \nu_\gamma(\cdot)}{E}\ \underset{\lambda \sim l_{\theta,t}(\cdot)}{E}[L_t(\pi^t, , \theta, \lambda)^2]}$$

$$\begin{aligned}
&= \frac{\Delta_t(\mu^t_{PS})^2}{\underset{\nu_\gamma\sim h_t(\cdot)}{E}\underset{\theta\sim h_t(\cdot)}{E}\underset{\pi\sim\nu_\gamma(\cdot)}{E}\underset{\lambda\sim l_{\theta,t}(\cdot)}{E}[L_t(\pi^t,,\theta,\lambda)^2]} \\
&= \frac{\left(\underset{\theta\sim h_t(\cdot)}{E}\sqrt{|\Omega_\theta|}\underset{l_{\theta,t}\sim h_t(\cdot)}{E}\sqrt{\underset{\substack{\pi\sim l_{\theta,t}(\cdot)\\ \lambda\sim l_{\theta,t}}}{E}[L_t(\pi^t,\theta,\lambda)]^2}\right)^2}{\underset{\nu_\gamma\sim h_t(\cdot)}{E}\underset{\theta\sim h_t(\cdot)}{E}\underset{\pi\sim\nu_\gamma(\cdot)}{E}\underset{\lambda\sim l_{\theta,t}(\cdot)}{E}[L_t(\pi^t,,\theta,\lambda)^2]} \\
&\leq |\Omega_{\theta_{max}}|\frac{\left(\underset{\theta\sim h_t(\cdot)}{E}\underset{l_{\theta,t}\sim h_t(\cdot)}{E}\sqrt{\underset{\substack{\pi\sim l_{\theta,t}(\cdot)\\ \lambda\sim l_{\theta,t}}}{E}[L_t(\pi^t,\theta,\lambda)]^2}\right)^2}{\underset{\nu_\gamma\sim h_t(\cdot)}{E}\underset{\theta\sim h_t(\cdot)}{E}\underset{\pi\sim\nu_\gamma(\cdot)}{E}\underset{\lambda\sim l_{\theta,t}(\cdot)}{E}[L_t(\pi^t,,\theta,\lambda)^2]} \\
&\leq |\Omega_{\theta_{max}}|\frac{\left(\sqrt{\underset{\theta\sim h_t(\cdot)}{E}\underset{l_{\theta,t}\sim h_t(\cdot)}{E}\underset{\substack{\pi\sim l_{\theta,t}(\cdot)\\ \lambda\sim l_{\theta,t}}}{E}[L_t(\pi^t,\theta,\lambda)]^2}\right)^2}{\underset{\nu_\gamma\sim h_t(\cdot)}{E}\underset{\theta\sim h_t(\cdot)}{E}\underset{\pi\sim\nu_\gamma(\cdot)}{E}\underset{\lambda\sim l_{\theta,t}(\cdot)}{E}[L_t(\pi^t,,\theta,\lambda)^2]} \\
&= |\Omega_{\theta_{max}}|\frac{\left(\sqrt{\underset{\theta\sim h_t(\cdot)}{E}\underset{l_{\theta,t}\sim h_t(\cdot)}{E}\underset{\substack{\pi\sim l_{\theta,t}(\cdot)\\ \lambda\sim l_{\theta,t}}}{E}[L_t(\pi^t,\theta,\lambda)]^2}\right)^2}{\underset{\theta\sim h_t(\cdot)}{E}\underset{l_{\theta,t}\sim h_t(\cdot)}{E}\underset{\pi\sim l_{\theta,t}(\cdot)}{E}\underset{\lambda\sim l_{\theta,t}(\cdot)}{E}[L_t(\pi^t,,\theta,\lambda)^2]} \\
&= |\Omega_{\theta_{max}}|
\end{aligned} \tag{2.15}$$

Note that $\Psi^L_t(\mu^t_L) \leq \Psi^L_t(\mu^t_{PS})$ by optimality of ν^t_L. Using this in (2.12) and combining with the above upper bound on $\Psi^L_t(\nu^t_{PS})$ yields

$$\Psi_t(\mu^t_{IDPS}) \leq \Psi^L_t(\mu^t_{PS})/2 \leq |\Omega_{\theta_{max}}|/2. \tag{2.16}$$

This completes the proof. □

Corollary 2.2.3. *The cumulative expected regret is bounded as*

$$Regret(N) \leq \frac{((1+\sqrt{N-1})\sqrt{\epsilon}+1)}{\sqrt{2}}\sqrt{|\Omega_{\theta_{max}}|\log(|\Omega_{\theta_{max}}|)},$$

with probability $\left(1 - e^{-\frac{(c\epsilon-2)^2}{2N}}\right)\left(1 - e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}\right)$, *where* $c > 2/\epsilon$ *and* $c_0 > 2/\sqrt{\epsilon}$ *are positive constants.*

The above result gives a stronger bound as compared to a flat MDPs.The regret bound for an equivalent flat MDP can be calculated using Corollary 1.3.6. The equivalent size of the parameter set for the flat MDP would be $\prod_{\theta}|\Omega_\theta|$, therefore using Corollary 1.3.6 the regret upper bound is

$$\frac{((1+\sqrt{N-1})\sqrt{\epsilon}+1)}{\sqrt{2}}\sqrt{\prod_{\theta}|\Omega_\theta|\log(\prod_{\theta}|\Omega_\theta|)},$$

which is much higher as compared to the hierarchical bound derived above, provided $|\Lambda_\theta|\forall\theta, |\Theta| > 1$, which always holds. In the next section, I implement Thompson Sampling and IDPS for a response-guided dosing problem. In each case I compare the performance of the algorithm in the hierarchical setting versus a flat setting.

2.3 Numerical results for response-guided dosing

Kotas and Ghate (2016) introduced a stochastic DP formulation for response-guided dosing in diseases that call for treatment courses with multiple sessions. Their objective was to tailor drug-doses to the stochastic evolution of each individual patient's disease condition over the treatment course, in order to trade-off the patient's aversion to doses with disease control. Patient's aversion to dose models ill-effect of treatment such as high cost, side effects, and logistical inconvenience of receiving treatment.

The system state equals to a numerical scores of the disease condition for the patient at the beginning of each treatment session. Higher scores correspond to worse disease conditions. Examples include cholesterol level for heart disease, viral load for hepatitis C, blood pressure, or DAS28 scores for rheumatoid arthritis. The decisions correspond to the doses administered to the patient in each session. The immediate cost is given by a disutility function that models patients' aversion to doses. The disease conditions evolve according to a dose-response function. The decision-maker's goal is to minimize the total expected disutility of the doses given to the patient over the treatment course and that of the disease condition reached at the end of the course.

In this section, I explore an extension of the problem in Kotas and Ghate (2016) under response-function uncertainty and function-parameter uncertainty. This problem naturally possesses a hierarchical information structure. I applied the framework developed in this chapter to make dosing decisions that try to attain a higher reward under this hierarchical parametric uncertainty.

I consider a treatment course with N sessions wherein disease condition measurements are made at the beginning of session $t = 1, 2, \ldots, N$ and a drug dose is administered. The disease condition in session t is denoted by X_t and the dose chosen for this session after measuring X_t is denoted by d_t. Doses d_t belong to the interval $D = [0, \bar{d}]$, where $\bar{d} < \infty$ is the maximum permissible dose in one session.

The treatment planner knows that the disease condition evolves according to one of two dose-response functions: one derived from the Michaelis-Menten formula and the other from power law. The planner is also uncertain about the specific parameters of these response-functions. As described in Kotas and Ghate (2016), the Michaelis-Menten response is written as

$$\ln(X_{t+1}) = \ln(X_t) + \ln(k_2) - \ln(k_1 + k_2 + d_t) + \mu_1. \tag{2.17}$$

Here, k_1 is a fixed parameter, and μ_1 is a known stochastic component that follows a zero-mean Normal distribution with standard deviation 5; k_2 is an unknown parameter that the planner must learn over the treatment course. Also, as described in Kotas and Ghate (2016), the power law response is given by

$$\ln(X_{t+1}) = \ln(X_t) - k \ln(1 + d_t) + \mu_2. \tag{2.18}$$

Here, k is an unknown parameter and μ_2 is a known stochastic component that I assume follows a zero mean Normal distribution with standard deviation 1.

For algebraic convenience, I modeled the state of the system as $s_t = \ln(X_t)$, and the actions a_t as the dose d_t in session t. These actions can take any value between $[0, 10]$

corresponding to $\bar{d} = 10$. The family of transition functions is parameterized by $\theta \in \{1, 2\}$, where $\theta = 1$ corresponds to Michaelis-Menten, and $\theta = 2$ corresponds to power law. For $\theta = 1$, the unknown model parameter λ_θ corresponds to k_2 and takes values in the interval $[5, 50]$. For $\theta = 2$, the unknown model parameter λ_θ corresponds to k and takes values in the interval $[0.5, 5]$. This yields transition functions

$$T_1(s'|s, a) = \mathcal{N}(s + \ln(k_2) - \ln(k_1 + k_2 + a), 5^2),$$

and

$$T_2(s'|s, a) = \mathcal{N}(s - k \ln(1 + a), 1).$$

I used a reward function similar to Kotas and Ghate (2016), given by

$$R(s'|s, a) = -e^s - ca,$$

where $c = 0.0285$.

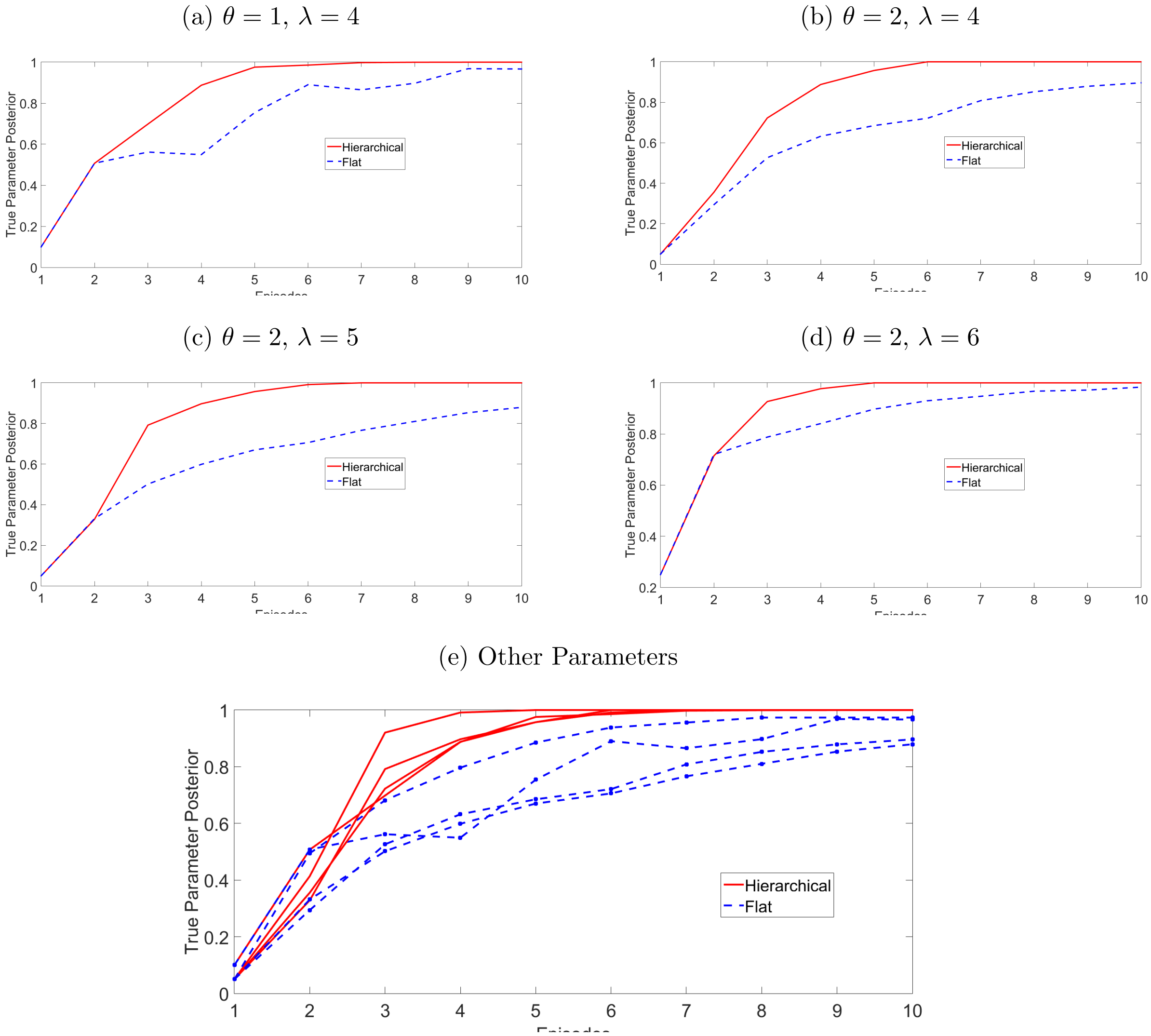
Figure 2.1 plots the posterior of the true parameter values and compares against the equivalent flat model both using Thomposon Sampling. Figure 2.2 plots the cumulative reward averaged over 50 runs for Thompson Sampling.

Similarly, Figure 2.3 plots learning curves for the posterior of the true parameter values and compares against the equivalent flat model while using IDPS. Figure 2.4 plots the cumulative reward averaged over 50 runs for IDPS. All plots show that my learning algorithms work better in the hierarchical setting than in the flat one.

2.4 Conclusion and future work

This chapter explores Bayesian learning in MDPs with hierarchical parametric uncertainty. This work is motivated by sequential decision problems where the decision-maker is uncertain

Figure 2.1: Thompson Sampling: Posterior for true parameter values.



(a) $\theta = 1,\ \lambda = 4$

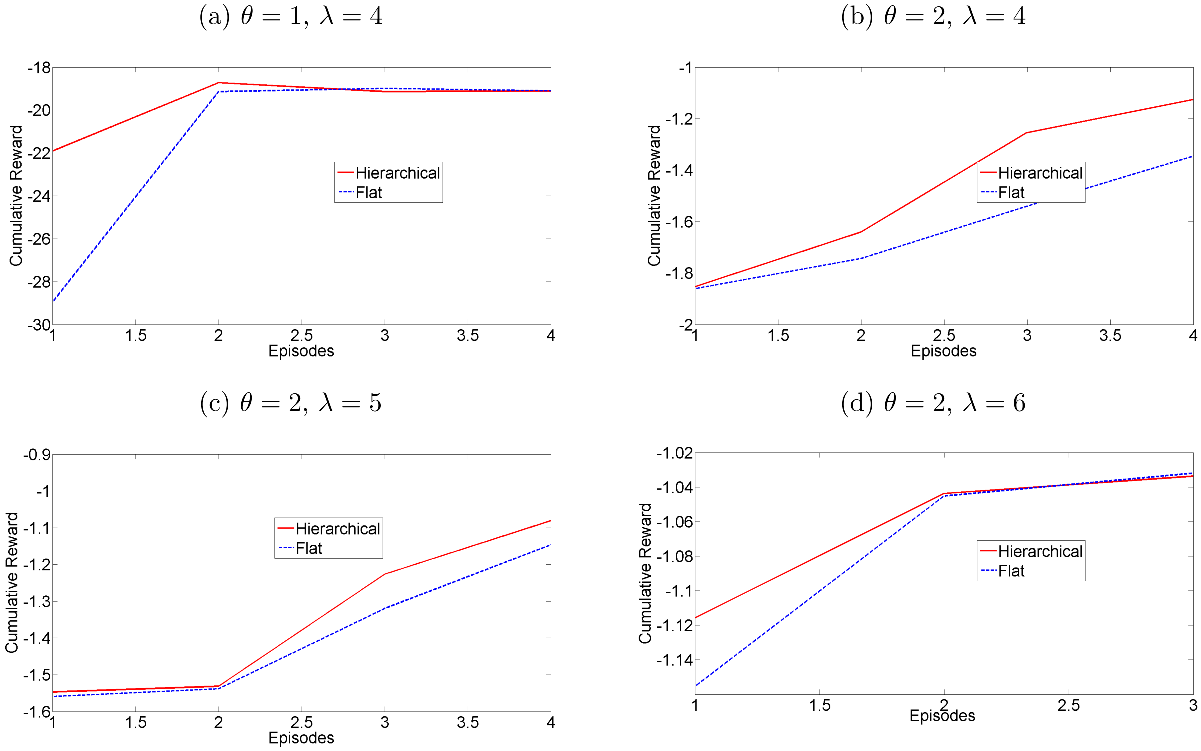
(b) $\theta = 2,\ \lambda = 4$

(c) $\theta = 2,\ \lambda = 5$

(d) $\theta = 2,\ \lambda = 6$

(e) Other Parameters

Figure 2.2: Thompson Sampling: Reward accumulated over multiple episodes.


(a) $\theta = 1,\ \lambda = 4$

(b) $\theta = 2,\ \lambda = 4$

(c) $\theta = 2,\ \lambda = 5$

(d) $\theta = 2,\ \lambda = 6$

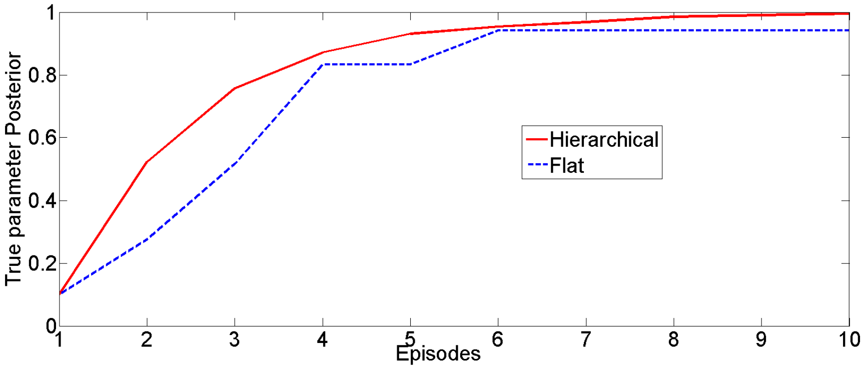


Figure 2.3: Information Directed Policy Sampling: Posterior for the true parameter.

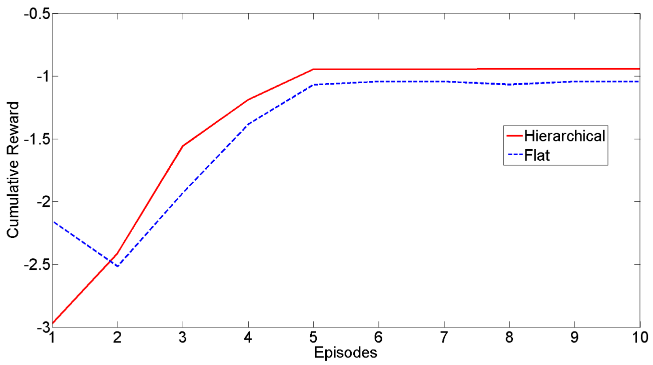


Figure 2.4: Information Directed Policy Sampling: Reward accumulated over multiple episodes.

about the model of stochastic system dynamics as well as about the parameter values for each model. In order to make good decisions, one must learn the system model as well as the model parameters during run-time as the system evolves. This can be viewed as simultaneous model selection and parameter estimation problem in the MDP context. The decision-maker maintains a prior on the system model, and a conditional prior on the parameters of each model. These priors are updated as state observations are made. I derived hierarchical prior update formulas using Bayes' rule and incorporated them into two learning algorithms: Thompson Sampling and IDPS. I provided numerical results on a response-guided dosing problem. In these results, a hierarchical modeling framework performed better in terms of both learning and reward accumulation. In addition, I derived theoretical performance bounds to demonstrate benefits of hierarchical modeling.

As in the previous chapter, the framework here assumes that the decision maker has access to (or can compute) the information gain and expected regret quantities. In practice, these quantities are computationally expensive to calculate. This work can benefit from developing efficient algorithms that can compute such quantities in an online manner. I believe temporal difference learning methods can provide a good opportunity to make this work more efficient in practice. It will also be interesting to see how model selection criterion (such as the Bayesian information criterion or Deviance information criterion) can be used to compare the advantages of hierarchical models and flat models without the specifics of algorithmic details.

Chapter 3

INFORMATION DIRECTED POLICY SAMPLING WITH PARTIALLY OBSERVABLE MARKOV DECISION PROCESS

In many MDPs, the state of the system is not directly observable. The decision maker only has access to partial or indirect observations of the system. A typical example is in medical treatment planning where the doctor makes certain measurements (observations) that provide information about a patient's health (state). In such cases, the system model also includes a distribution over observations conditioned on the state, in addition to the usual transition distribution, and reward function. This problem with imperfect information is called a Partially Observable MDP (POMDP) (Krishnamurthy, 2016). A POMDP is described by a tuple $\mathcal{Q} = (S, A, T, R, N, O, Z)$, where (S, A, T, R, N) is the same as in an MDP, O is the set of possible observations, and Z is a probability distribution over the observation of the systems conditioned on the state. This problem with imperfect information can be reformulated as an MDP (with perfect information), where the redefined state equals the decision-maker's probabilistic belief about the actual system state. This redefined state is often called "information state" and is a sufficient statistic for the decision-making problem. The decision-maker updates this belief using Bayes' Theorem, as the system evolves and observations are made over time. In fact, there is a natural equivalence between BAMDPs and POMDPs that has been discussed in the existing literature (Bertsekas, 2005; Duff, 2002). Thus, POMDPs with parametric uncertainty form a rich class of problems potentially amenable to information theoretic learning methods.

This chapter starts by providing background on POMDPs in Section 3.1. Section 3.2 develops IDPS for POMDPs with parameteric uncertainty. Section 3.3 provides theoretical bounds for this IDPS procedure. Section 3.4 starts by developing a generic POMDP model for response-guided dosing. This model is employed to demonstrate the benefits of IDPS for POMDPs. The rest of that section is organized into three parts, which provide algorithms and numerical results for three different special cases of the generic response-guided dosing problem. The chapter then closes with conclusion and some discussion on open problems.

3.1 Background

The MDP formulation for POMDP follows a belief MDP framework. I define the belief state as $x_t = (x_t^1, \cdots, x_t^{|S|})$, where $x_t^i = P(s_t = i)$, the probability that the system is in state $i \in S$ at time t.

$$X = \left\{ x \in \mathbb{R}_+^{|S|} \middle| \sum_i x^i = 1 \right\}$$

The transition between the belief states is given by a transition function $\phi : X \times A \times O \mapsto \phi$. I define $\phi_{s_{t+1}}(x_t, a_t, o_{t+1})$ as the probability that the next state is s_{t+1} given that the action a_t was chosen in belief state x_t and observation o_{t+1} was made. With some algebraic manipulations and using the Markovian property, we get

$$\phi_{s_{t+1}}(x_t, a_t, o_{t+1}) = \frac{Z(o_{t+1}|s_{t+1}, a_t) \sum_{s_t \in S} T(s_{t+1}|s_t, a_t) x_t(s_t)}{\sum_{s_{t+1}} Z(o_{t+1}|s_{t+1}, a_t) \sum_{s_t \in S} T(s_{t+1}|s_t, a_t) x_t(s_t)},$$

and

$$\phi(x_t, a_t, o_{t+1}) = \{\phi_{s_{t+1}}(x_t, a_t, o_{t+1}), \forall s_{t+1} \in S\}. \tag{3.1}$$

For a POMDP, this implies $x_{t+1}(.|x_t, a_t, o_{t+1}) = \phi(x_t, a_t, o_{t+1})$. The Bellman equation for the POMDP is given by

$$\begin{aligned} U_t^*(x_t) = \max_{a_t \in A} \Bigg(\sum_{s_t \in S} \Bigg(\sum_{s_{t+1} \in S} T(s_{t+1}|s_t, a_t) r(s_{t+1}|s_t, \pi(x_t)) \Bigg) \\ + \alpha \sum_{o_{t+1} \in O} P(o_{t+1}|x_t, a_t) U_{t+1}^*(x_{t+1}) \Bigg), \end{aligned}$$

where $P(o_{t+1}|x_t, \pi_t(o_t)) = \sum_{s_{t+1}} Z(o_{t+1}|s_{t+1}, \pi_t(x_t)) \sum_{s_t \in S} T(s_{t+1}|s_t, \pi_t(x_t))x_t(s_t)$.

3.2 Information directed sampling in POMDPs

I utilize the equivalence between a BAMDP and a POMDP to re-derive (or re-state) theoretical results for IDPS applied to POMDPs. I supplement these with numerical results on POMDPs. Following the same theme as in previous chapters, I use Thompson Sampling as a benchmark to evaluate the performance of IDPS on POMDPs.

Consider the parametric uncertainty in the transition distribution T and the probability distribution Z. Let Ω be the finite set whose elements, $\lambda \in \Omega$, index the transition distributions T_λ, and let $\mathcal{Q}$ be the finite set whose elements, $\mu \in \mathcal{Q}$ index the probability distribution Z. The family of possible POMDPS is given by $\mathcal{Q}_{\lambda,\mu} = \{S, A, T_\lambda, R, N, O, Z_\mu\}$. I assume, as in Chapter 1, that the decision maker can easily solve the POMDP $\mathcal{Q}_{\lambda,\mu}, \forall \mu \in \mathcal{Q}, \lambda \in \Omega$. The decision maker begins with a prior belief $\alpha_1(.)$ on λ and $\beta_1(.)$ on μ. These belief pmfs are updated via Bayes' Theorem as states drawn from the true transition function T_{λ^*}, and observations drawn from the true distribution function Z_{μ^*} are observed starting from an initial state $s_1 \in S$. This enables the decision to keep track of the posterior joint distribution $\delta_t(\lambda, \mu)$. The decision-maker's objective is to simultaneously learn the true transition function and the true observation distribution function while maximizing expected reward.

To extend IDPS to POMDP, I redefine some quantities. Suppose the decision-maker records the observation o_t at the beginning of slot t. Let $\pi^t = (\pi_t, \pi_{t+1}, \cdots, \pi_N)$ denote the tail of any policy trajectory $\pi = (\pi_1, \pi_2, \ldots, \pi_N) \in \mathcal{P}$. The set of tail policy trajectories is denoted by $\mathcal{P}^t$. Also let $\pi^*_{\lambda,\mu} = (\pi^*_{1,\lambda,\mu}, \ldots, \pi^*_{N,\lambda,\mu})$ denote an optimal policy for POMDP $\mathcal{M}_{\lambda,\mu}$. Let $x_{t,\lambda,\mu}(s)$ be the probability of being in state $s \in S$ at time t for parameters λ and μ. Knowing o_{t+1}, and the distributions T_λ, Z_μ, the decision-maker estimates $x_{t+1,\lambda,\mu}(s_{t+1}) =$

$\phi_{s_{t+1},\lambda,\mu}(x_t, a_l, o_{t+1})$ as in equation 3.1.

$$\phi_{s_{t+1},\lambda,\mu}(x_t, a_t, o_{t+1}) = \frac{Z_\mu(o_{t+1}|s_{t+1}, a_t) \sum_{s_t \in S} T_\lambda(s_{t+1}|s_t, a_t) x_t(s_t)}{\sum_{s_{t+1}} Z_\mu(o_{t+1}|s_{t+1}, a_t) \sum_{s_t \in S} T_\lambda(s_{t+1}|s_t, a_t) x_t(s_t)}. \tag{3.2}$$

I define a quantity $R'(o_{\ell+1}|x_\ell, \pi_\ell(x_\ell))$ which denotes the reward observed when the current belief state is x_ℓ, and an action $\pi_\ell(x_\ell)$ is applied which results in observing $o_{\ell+1}$. This quantity is defined to enable a more intuitive interpretation of value functions and other related quantities described later in this section. A decision maker never observes the state of the system directly, therefore the quantity $R(s_{\ell+1}|s_\ell, \pi_\ell(s_\ell))$ is not observed directly. Later in this section, I express $R'(o_{\ell+1}|x_\ell, \pi_\ell(x_\ell))$ in terms of the model of the system, T, Z and R.

Now let $V_t^*(\lambda, \mu) = \sum_{\ell=t}^{N} R'(o_{\ell+1}|x_\ell, \pi^*_{\ell,\lambda,\mu}(x_\ell))$ denote the random tail reward accumulated on implementing an optimal policy in MDP $\mathcal{M}_{\lambda,\mu}$. Similarly, $V_t(\lambda, \mu, \pi^t) = \sum_{\ell=t}^{N} R'(o_{\ell+1}|x_\ell, \pi_\ell(x_\ell))$ for any tail policy π^t. In both these expressions, observation o_{t+1} is drawn from a distribution derived from T_λ and Z_μ given observation o_t, a belief state x_t, and action $\pi^*_{t,\lambda,\mu}(x_t)$ or action $\pi_t(x_t)$, respectively. Now consider

$$\begin{aligned}
U_t(\lambda, \mu, \pi^t) &= \underset{\{o_{\ell+1} \sim P_{\lambda,\mu}(\cdot|x_\ell, \pi_{\ell,\lambda,\mu}(x_\ell)) : \ell = t,\ldots,N\}}{E} [V_t(\lambda, \mu, \pi^t)] \\
&= \underset{\{o_{\ell+1} \sim P_{\lambda,\mu}(\cdot|x_\ell, \pi_{\ell,\lambda,\mu}(x_\ell)) : \ell = t,\ldots,N\}}{E} \left[\sum_{\ell=t}^{N} R'(o_{\ell+1}|x_\ell, \pi_{\ell,\lambda,\mu}(x_\ell)) \right] \\
&= \sum_{o_{\ell+1}:\ell=t,\ldots,N} P_{\lambda,\mu}(\cdot|x_\ell, \pi_{\ell,\lambda,\mu}(x_\ell)) \left[\sum_{\ell=t}^{N} \sum_{s_{\ell+1} \in S} Z_\mu(o_{\ell+1}|s_{\ell+1}, \pi_{\ell,\lambda,\mu}(x_\ell)) \right. \\
&\left. \sum_{s_\ell \in S} T_\lambda(s_{\ell+1}|s_\ell, \pi_{\ell,\lambda,\mu}(x_\ell)) x_\ell(s_\ell) R(s_{\ell+1}|s_\ell, \pi_{\ell,\lambda,\mu}(x_\ell)) \right].
\end{aligned}$$

Similarly,

$$U_t^*(\lambda, \mu) = \underset{\{o_{\ell+1} \sim P_{\lambda,\mu}(\cdot|x_\ell, \pi^*_{\ell,\lambda,\mu}(x_\ell)) : \ell = t,\ldots,N\}}{E} [V_t^*(\lambda, \mu)]$$

$$= \underset{\{o_{\ell+1}\sim P_{\lambda,\mu}(\cdot|x_\ell,\pi^*_{\ell,\lambda,\mu}(x_\ell)):\ell=t,\ldots,N\}}{E}\left[\sum_{\ell=t}^{N} R'(o_{\ell+1}|x_\ell,\pi^*_{\ell,\lambda,\mu}(x_\ell))\right]$$

$$= \sum_{o_{\ell+1}:\ell=t,\ldots,N}\left[\sum_{\ell=t}^{N}\sum_{s_{\ell+1}\in S} Z_\mu(o_{\ell+1}|s_{\ell+1},\pi^*_{\ell,\lambda,\mu}(x_\ell))\right.$$

$$\left.\sum_{s_\ell\in S} T_\lambda(s_{\ell+1}|s_\ell,\pi^*_{\ell,\lambda,\mu}(x_\ell))x_\ell(s_\ell)R(s_{\ell+1}|s_\ell,\pi^*_{\ell,\lambda,\mu}(x_\ell))\right].$$

To define the information ratio, we need to first characterize the expected regret and information gain of tail policy trajectory π^t. In particular, the expected regret is defined as

$$\Delta_t(\pi^t|o_t,x_t,\delta_t(\cdot,\cdot)) = \underset{\{\lambda,\mu\}\sim\delta_t(\cdot,\cdot)}{E}\left[U_t^*(\lambda,\mu) - U_t(\lambda,\mu,\pi^t)\right]. \tag{3.3}$$

This expression computes the expectation (with respect to the decision-maker's posterior $\delta_t(\cdot)$) of the expected difference between the optimal value and the value of policy π^t.

Please recall that the information gain (or mutual information) between two random variables X and Y is given by $I(X;Y) = \sum\limits_{x,y} P(x,y)\ln\frac{P(y|x)}{P(y)}$, where the letter P denotes the appropriate joint, conditional, and marginal distributions. In our POMDP context, X takes values in the parameter set $\Omega\times\mathcal{Q}$, while Y takes values in the observation set $\{o_{\ell+1}\sim P_{\lambda^*,\mu^*}(\cdot|x_\ell,\pi_\ell(x_\ell)) : \ell = t,\cdots,N\}$. The information gain thus equals

$$g_t(\pi^t|o_t,x_t,\delta_t(\cdot)) = \sum_{\substack{\lambda\in\Omega\\ \mu\in\mathcal{Q}}}\sum_{o_{t+1},\ldots,o_N}\left(\prod_{\ell=t}^{N} P_{\lambda,\mu}(o_{\ell+1}|x_\ell,\pi_\ell(x_\ell))\right)$$

$$\delta_t(\lambda,\mu)\ln\left[\frac{\prod\limits_{\ell=t}^{N} P_{\lambda,\mu}(o_{\ell+1}|x_\ell,\pi_\ell(x_\ell))}{\sum\limits_{\substack{\lambda\in\Omega\\ \mu\in\mathcal{Q}}}\prod\limits_{\ell=t}^{N} P_{\lambda,\mu}(o_{\ell+1}|x_\ell,\pi_\ell(x_\ell))\delta_t(\lambda,\mu)}\right], \tag{3.4}$$

where $P_{\lambda,\mu}(o_{\ell+1}|x_\ell,\pi_\ell(x_\ell)) = \sum_{s_{\ell+1}} Z_\mu(o_{\ell+1}|s_{\ell+1},\pi_\ell(x_\ell))\sum_{s_\ell\in S} T_\lambda(s_{\ell+1}|s_\ell,\pi_\ell(x_\ell))x_\ell(s_\ell)$. Now let D^t denote the set of all probability distributions over tail policies in $\mathcal{P}^t$. That is, probability distribution $\nu^t\in D^t$ assigns probability $\nu^t(\pi^t)$ to tail policy $\pi^t\in\mathcal{P}^t$. Then, given the pair $(x_t, \delta_t(\cdot,\cdot))$ at the beginning of slot t, the expected regret and expected information gain of probability distribution ν^t are given by

$$\Delta_t(\nu^t|o_t, x_t, \delta_t(\cdot,\cdot)) = \underset{\pi^t\sim\nu^t}{E}\left[\Delta_t(\pi^t|x_t, \delta_t(\cdot,\cdot))\right], \ \forall \nu^t \in D^t, \tag{3.5}$$

and

$$g_t(\nu^t|o_t, x_t, \delta_t(\cdot,\cdot)) = \underset{\pi^t\sim\nu^t}{E}\left[g_t(\pi^t|x_t, \delta_t(\cdot,\cdot))\right], \ \forall \nu^t \in D^t. \tag{3.6}$$

As in Chapter 1 the information ratio is defined as

$$\Psi_t(\nu^t) = \frac{(\Delta_t(\nu^t))^2}{g_t(\nu^t)}. \tag{3.7}$$

The decision-maker finds a probability distribution over D^t by solving

$$\nu_*^t \in \underset{\nu^t\in D^t}{\operatorname{argmin}}\, \Psi_t(\nu^t). \tag{3.8}$$

Let Ψ_t^* denote the optimal value $\min\limits_{\nu^t\in D^t} \Psi_t(\nu^t)$ of this problem.

3.3 Regret bounds

The next theorem bounds the decision-maker's cumulative expected regret over N stages, which is defined as

$$\text{Regret}(N) = \sum_{t=1}^{N}\left(\Delta_t(\nu_*^t|o_t, x_t) - \underset{o_{t+1}}{E}[\Delta_{t+1}(\nu_*^t|o_{t+1}, x_{t+1})]\right), \tag{3.9}$$

where $o_{t+1} \sim P_{\lambda^*,\mu^*}(\cdot|x_t, \pi^t(x_t))$. The decision maker estimates the belief state by using an expectation on $x_{t+1,\lambda,\mu}$ in 3.2, and

$$x_{t+1} = \underset{\lambda,\mu}{E}\left[\left\{\frac{Z_\mu(o_{t+1}|s_{t+1}, a_t)\sum_{s_t\in S} T_\lambda(s_{t+1}|s_t, a_t)x_t(s_t)}{\sum_{s_{t+1}} Z_\mu(o_{t+1}|s_{t+1}, a_t)\sum_{s_t\in S} T_\lambda(s_{t+1}|s_t, a_t)x_t(s_t)}, \forall s_{t+1}\right\}\right] \tag{3.10}$$

with $\pi^t \sim \nu_*^t$. Here,

$$\Delta_{t+1}(\nu_*^t|o_{t+1}, x_{t+1}) = \underset{(\pi_t, \pi^{t+1}) \sim \nu_*^t}{E} \left[\Delta_{t+1}(\pi^{t+1}|o_{t+1}, x_{t+1})\right], \tag{3.11}$$

with

$$\Delta_{t+1}(\pi^{t+1}|o_{t+1}, x_{t+1}) = \sum_{\lambda, \mu \sim \delta_t(\cdot)} \left[U_{t+1}^*(\lambda, \mu) - U_{t+1}(\lambda, \mu, \pi^{t+1})\right].$$

Recall from (3.5) that $\Delta_t(\nu_*^t|o_t, x_t)$ is the expected regret if a policy π^t sampled according to $\nu_*^t \in D^t$ is implemented in stages $t : N$ starting in belief state x_t with observation o_t. The term $\Delta_{t+1}(\nu_*^t|o_{t+1}, x_{t+1})$ defined in (1.12) above is the expected regret if a policy $\pi^t = (\pi_t, \pi^{t+1})$ sampled according to ν_*^t at time-step t is implemented in stages $t+1 : N$ starting in state x_{t+1} and observing o_{t+1}. Thus, the difference $\Delta_t(\nu_*^t|o_t, x_t) - \underset{o_{t+1}}{E}[\Delta_{t+1}(\nu_*^t|o_{t+1}, x_{t+1})]$ may be viewed as the one-step expected regret of ν_*^t. The cumulative expected regret over N stages as defined in (2.9) can therefore be interpreted as a sum of these one-step regrets, because the decision-maker recomputes the randomized policy ν_*^t at every time-step t.

Theorem 3.3.1. Worst Case Regret Bound: *Suppose there is a γ such that $\Psi^* \leq \gamma$. Then, for any $\epsilon > 0$, we have,*

$$\textit{Regret}(N) \leq ((1+\sqrt{N-1})\sqrt{\epsilon}+1)\sqrt{\gamma \log(|\mathcal{Q}||\Omega|)} \quad \textit{with probability}\ (1-e^{-\frac{(c\epsilon-2)^2}{2N}})(1-e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}),$$

where $c > 2/\epsilon$ and $c_0 > 2/\sqrt{\epsilon}$ are positive constants.

Proof. The proof of this theorem is a straightforward generalization of the flat IDPS Regret bounds, which can be derived using some minor algebraic manipulations on the proof of Theorem 1.3.3. It starts by finding an equivalent MDP for the POMDP. The equivalent MDP $M_{\lambda,\mu}^Q = \{X, A, P_{\lambda,\mu}(x'|x, \pi(x)), \underset{X \times X}{E}[R], N\}$ for the POMDP Q, where $P_{\lambda,\mu}(x'|x, \pi(x)) = f_{\lambda,\mu}(T, Z, x, x', \pi)$. Let the cardinality of P be $|\Omega_Q| = |\mathcal{Q}\Omega|$. Now applying Theorem 1.3.3

for IDPS on the MDP M^Q yields

$$\begin{aligned}\text{Regret}(N) &\leq ((1+\sqrt{N-1})\sqrt{\epsilon}+1)\sqrt{\gamma\log(|\Omega_Q|)} \quad \text{w. p. } (1-e^{-\frac{(c\epsilon-2)^2}{2N}})(1-e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}) \\ &\leq ((1+\sqrt{N-1})\sqrt{\epsilon}+1)\sqrt{\gamma\log(|\mathcal{Q}||\Omega|)} \quad \text{w. p. } (1-e^{-\frac{(c\epsilon-2)^2}{2N}})(1-e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}).\end{aligned}$$

A point to note is that even though the MDP in Theorem 1.3.3 had discrete state space S, the proof of theorem does not assume anything about the continuity of the state space except that the space should allow a definition of a probability measure. Therefore Theorem 1.3.3 is also valid for continuous state space, which enables me to apply Theorem 1.3.3 to the equivalent continuous state space MDP M_Q. □

The following corollary is a straightforward application of the Theorem 3.3.1 and the bound on information ratio.

Corollary 3.3.2. *The cumulative expected regret is bounded as*

$$Regret(N) \leq \frac{((1+\sqrt{N-1})\sqrt{\epsilon}+1)}{\sqrt{2}}\sqrt{|\mathcal{Q}||\Omega|\log(|\mathcal{Q}||\Omega|)},$$

with probability $\left(1-e^{-\frac{(c\epsilon-2)^2}{2N}}\right)\left(1-e^{-\frac{(c_0\sqrt{\epsilon}-2)^2}{2N}}\right)$, *where* $c > 2/\epsilon$ *and* $c_0 > 2/\sqrt{\epsilon}$ *are positive constants.*

The proof of this theorem is a straightforward generalization of the flat IDPS Regret bounds, which can be derived using some minor algebraic manipulations on the proof of Theorem 1.3.3.

3.4 Algorithms and numerical results

This section evaluates IDPS on POMDP on a generic version of responde-guided dosing for specific cases. I start here by describing an example that is a more complex case of the responde-guided dosing problem (the simpler case described in Section 1.4.3) due to the partial observability and the corresponding uncertainties.

3.4.1 Response-guided dosing

This section considers a POMDP formulation for response-guided dosing (RGD) in diseases that call for treatment courses with multiple sessions. This problem is a generalization of the problem described in Section 1.4.3. Again, the objective in RGD is to tailor drug-doses or select various treatment modalities to the stochastic evolution of each individual patient's disease condition over the treatment course, in order to trade-off the patient's aversion to doses with disease control. In this case, the decision maker does not observe the disease state directly, in fact, the decision maker makes an observation based on a measurement of the disease condition. At the beginning of each treatment session, the decision maker makes an observation about the patient's state which includes a numerical score of the patient's disease condition, and also a numerical score of the treatment's side effect. These numerical score probabilistically represents the patient's overall score. Based on these observations, the decision maker estimates the probability of patient's state $x(s) \forall s$. The decision maker then makes a treatment decision, the decisions correspond to the doses administered to the patient in each session after observing the numerical scores. The immediate cost is given by a disutility function that models patients' aversion to doses. The disease conditions and side effects evolve according to a stochastic dose-response function of the dose level. The decision-maker's goal is to minimize the total expected disutility of the doses given to the patient over the treatment course and that of the disease condition reached at the end of the course. A good dosing strategy calls for adapting doses to the stochastic evolution of each individual patient's disease condition. This in turn requires learning the key parameters of the individual patient's dose-response function over the treatment course, and the distribution function over measurement outcomes, while simultaneously selecting doses.

The model here is an extension of the framework in Section 1.4.3. We consider a treatment course with N sessions wherein disease condition and side effect measurements are made and then a drug dose is administered at the beginning of sessions $t = 1, 2, \ldots, N$. The disease condition in session t is denoted by X_t, patient's side effect is denoted by Y_t, and dose

d_t is chosen for this session after measuring $o_t = \{o_{X,t}, o_{Y,t}\}$. Disease state X_t is integer value and belongs to the interval $[0, m]$ with $X_t = 0$ representing the best disease condition and $X_t = m$ denotes the worst disease condition. The measured disease scores $o_{X,t}$ takes values in $[0, m_o]$. Patient's side effect state Y_t is also integer valued and belongs to the interval $[0, n]$ with $Y_t = 0$ representing no side effect and $Y_t = n$ representing the worst side effect. Patient's side effect toxicity scores $o_{Y.t}$ takes values in $[0, n_o]$. Doses d_t are also integer valued and belong to the interval $D = [0, \bar{d}]$, where $\bar{d} < \infty$ is the maximum permissible dose in one session. The patient's state $s_t \in S$ represents the patient's overall health status which is patient's disease state and toxicity state due to dosing. The state space $S = \{S_L, S_D\}$ consists of living states S_L and an absorbing state S_D which corresponds to patient death. The living states S_L represents patients overall health and constitutes disease states S_X, such as good, bad and critical, and toxicity states due to dosing S_Y, such as, high toxicity, low toxicity and critical toxicity . Let $S = S_X \times S_Y = [0, y] \times [0, z]$, where $\{y, z\}$ represents the absorbing state S_D, and $\{0, 0\}$ represents the best health possible.

The disease condition and treatment's side effects evolve according to a probability distribution, as in Section 1.4.3. For our numerical simulation, we employed $m = n = 6$, $c = 6$, $q_X = q_Y = 2$, and $\bar{d} = 3$.

Additionally, the treatment planner does not observe the states directly but knows that the measurement outcomes are observed according to a probability distribution. The measurement outcome probability of disease score is denoted by $P^X(O_{X,t+1}|s_{t+1}, d_t)$, measurement outcome probability of side effect score is denoted by $P^Y(O_{Y,t+1}|s_{t+1}, d_t)$, and the measurement outcome probability is given by $P(o_{t+1}|s_{t+1}, d_t) = P^X(O_{X,t+1}|s_{t+1}, d_t) \times P^Y(O_{Y,t+1}|s_{t+1}, d_t)$, where $s_t = \{X_t, Y_t\}$.

3.4.2 Unknown transition distribution and known observation distribution

Consider the case where the transition distribution of the MDP is parametrized as in Chapters 1 or 2, while the observation distribution is known. In this case, $\lambda \in \Omega$ and $\mu = \mu^*$, which is known. Therefore, $\delta(\lambda, \mu) = \delta(\lambda, \mu^*) = \alpha(\lambda|\mu^*)$. Algorithm 4 presents the algorithm for this case.

Algorithm 4 Information Directed Policy Sampling

```
Require: MDPs $\mathcal{Q}_\lambda = \{S, A, T_\lambda, R, N, O, Z_{\mu^*}\}$ for $\lambda \in \Omega$. Prior pmf $\alpha_1(\cdot)$. Initial belief
    state $x_1(s), s \in S$.
$x_1(s) = p(s), \forall s \in S$

function IDPS
    for episode $k = 1, 2, 3, \cdots$ do
        Set $t = 1$
        Initialize $x_1(s)$; and prior $\alpha_1(\cdot) \leftarrow \alpha_{N+1}(\cdot)$ if $k > 1$
        repeat
            Compute distribution $\nu_*^t = \underset{\nu^t \in D^t}{\operatorname{argmin}} \Psi_t(\nu^t | o_t, x_t, \alpha_t(\cdot))$
            Sample $\pi^t = (\pi_t, \ldots, \pi_N) \sim \nu_*^t$
            Implement action $\pi_t(x_t)$
            Observe $o_{t+1}$ drawn from $P_{\lambda^*,\mu^*}(\cdot|x_t, \pi_t(x_t))$
            Estimate $x_{t+1}$ as in equation 3.10
            Update probability mass $\alpha_{t+1}(\lambda) \propto P_{\lambda,\mu^*}(o_{t+1}|x_t, \pi_t(x_t))\alpha_t(\lambda)$, for each $\lambda \in \Omega$
            t ← t+1
        until end of horizon $N$
    end for
end function
```

Numerical experiments for response-guided dosing

In this case the planner is uncertain about the parameter value λ that characterizes the patient's evolution of side effect states but knows the distribution characterizing the patient's disease evolution, and also knows the measurement outcome parameter μ^*. The measurement outcome probability for the disease scores is $Z^X(O_{x,t+1}|X_{t+1}, d_t)$

$$Z^X(o_{x,t+1}|x_{t+1}, d_t) = |\mathcal{F}(o_{x,t+1} + 1 - \frac{m_o}{m}x_{t+1} + \mu^* d_t) - \mathcal{F}(o_{x,t+1} - \frac{m_o}{m}x_{t+1} + \mu^* d_t)|.$$

Similarly, the measurement outcome probabilities for the toxicity score is given by

$$Z^Y(o_{y,t+1}|y_{t+1},d_t) = |\mathcal{F}(o_{y,t+1} + 1 - \frac{n_o}{n} y_{t+1} - \mu d_t) - \mathcal{F}(o_{y,t+1} - \frac{n_o}{n} y_{t+1} - \mu d_t)|.$$

Therefore, the measurement outcome probability is given as

$$Z(o_{t+1}|s_{t+1},d_t) = Z^X(o_{x,t+1}|x_{t+1},d_t) \times Z^Y(o_{y,t+1}|y_{t+1},d_t).$$

For this section the decision maker knows that $\mu = \mu^* = 0.125$. Figure 3.2 plots the cumulative reward averaged over 50 runs for IDPS and Thompson Sampling. Figure 3.1 plots the posterior of the true parameter values for IDPS and Thompson Sampling. We observe that for MDPs with partial observability IDPS learns faster than Thompson Sampling approaches which is consistent with the objective of this research.

3.4.3 Unknown measurement outcome distribution with known transition distribution

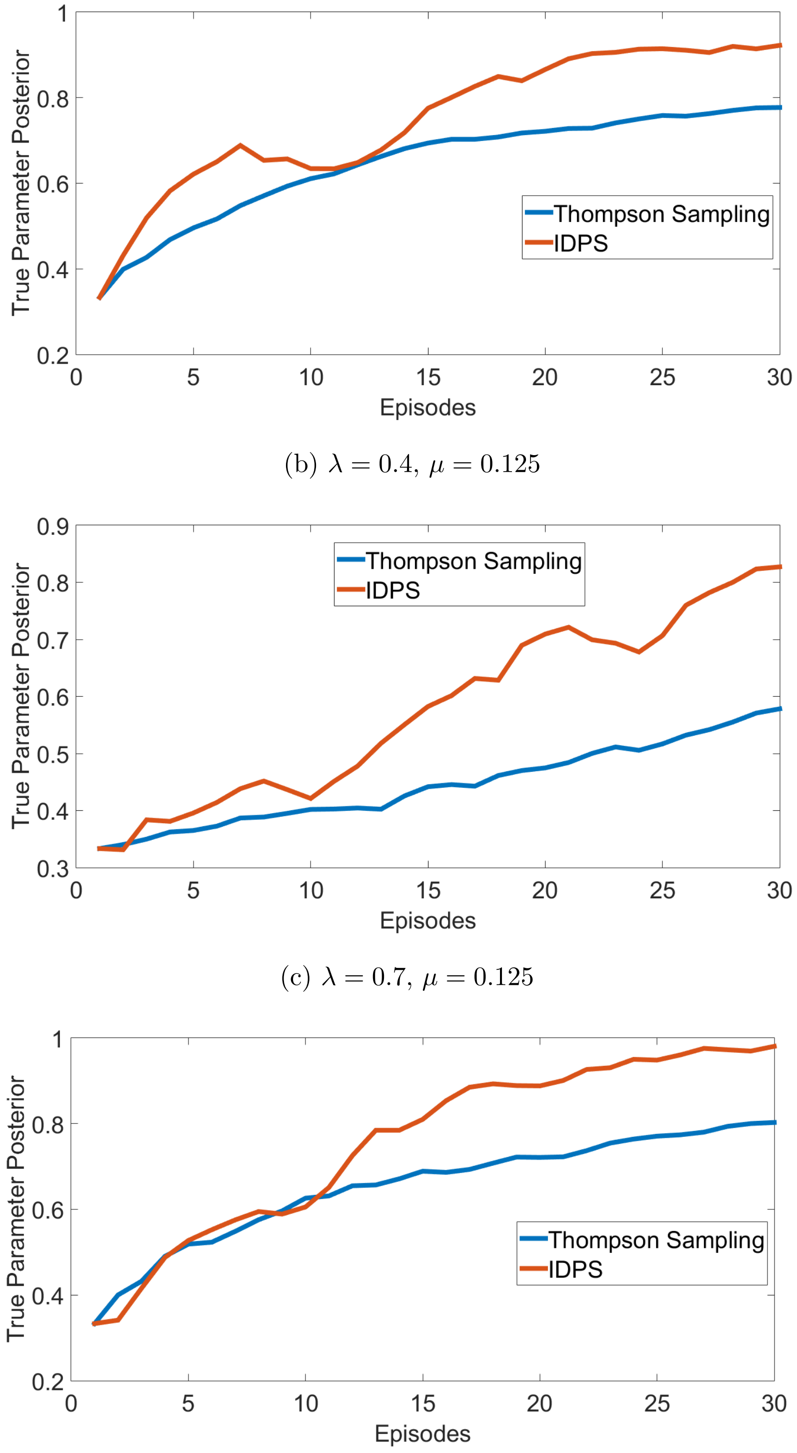
In this subtask, I investigate the case where the decision-maker is uncertain about the observation distribution, but knows the transition distribution. I extend my previous information theoretic learning methods to this setting. In this case, $\lambda = \lambda^*$, which is known and $\mu \in \mathcal{Q}$. Therefore, $\delta(\lambda, \mu) = \delta(\lambda^*, \mu) = \beta(\mu|\lambda^*)$. Algorithm 5 presents the algorithm for this case. I demonstrate this on the responde-guided dosing application as in section 3.4.1.

Numerical experiments for response-guided dosing

This case uses the same problem as in Section 3.4.1. Except now the planner is uncertain about the parameter value μ that characterizes the patient's disease and toxicity level measurement outcomes, but knows the transition distribution characterizing the patient's disease evolution and side effect evolution $\lambda = \lambda^* = 0.7$. The disease maker also know that $\mu \in \mathcal{Q} = \{0.05, 0.25, 0.45\}$. Figure 3.4 plots the cumulative reward averaged over 50 runs for

Figure 3.1: Posterior for true parameter values.

(a) $\lambda = 0.1$, $\mu = 0.125$

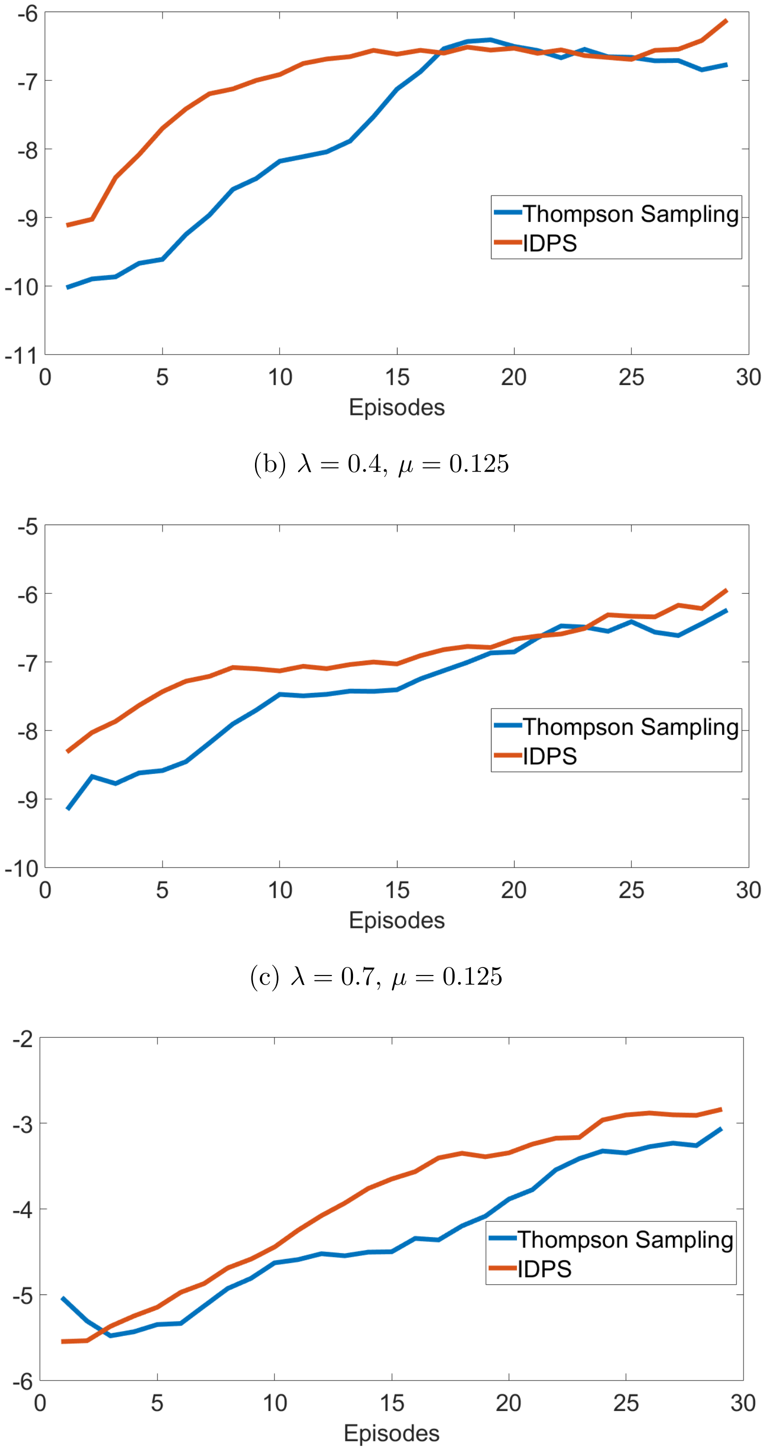


(b) $\lambda = 0.4$, $\mu = 0.125$

(c) $\lambda = 0.7$, $\mu = 0.125$

Figure 3.2: Averaged Cumulative Reward

(a) $\lambda = 0.1$, $\mu = 0.125$



(b) $\lambda = 0.4$, $\mu = 0.125$

(c) $\lambda = 0.7$, $\mu = 0.125$

Algorithm 5 Information Directed Policy Sampling

```
Require: MDPs $\mathcal{Q}_\mu = \{S, A, T_{\lambda^*}, R, N, O, Z_\mu\}$ for $\lambda \in \Omega$. Prior pmf $\beta_1(\cdot)$. Initial belief
    state $x_1(s), s \in S$.
$x_1(s) = p(s), \forall s \in S$

function IDPS
    for episode $k = 1, 2, 3, \cdots$ do
        Set $t = 1$
        Initialize $x_1(s)$; and prior $\beta_1(\cdot) \leftarrow \beta_{N+1}(\cdot)$ if $k > 1$
        repeat
            Compute distribution $\nu_*^t = \underset{\nu^t \in D^t}{\operatorname{argmin}} \Psi_t(\nu^t | o_t, x_t, \beta_t(\cdot))$
            Sample $\pi^t = (\pi_t, \ldots, \pi_N) \sim \nu_*^t$
            Implement action $\pi_t(x_t)$
            Observe $o_{t+1}$ drawn from $P_{\lambda^*,\mu^*}(\cdot | x_t, \pi_t(x_t))$
            Estimate $x_{t+1}$ as in equation 3.10
            Update probability mass $\beta_{t+1}(\mu) \propto P_{\lambda^*,\mu}(o_{t+1} | x_t, \pi_t(x_t))\beta_t(\mu)$, for each $\mu \in \mathcal{Q}$
            t ← t+1
        until end of horizon $N$
    end for
end function
```

IDPS and Thompson Sampling. Figure 3.3 plots the posterior of the true parameter values for IDPS and Thompson Sampling. IDPS learns faster than Thompson Sampling approaches which is consistent with the objective of this research.

3.4.4 Unknown measurement outcome distribution and unknown transition distribution

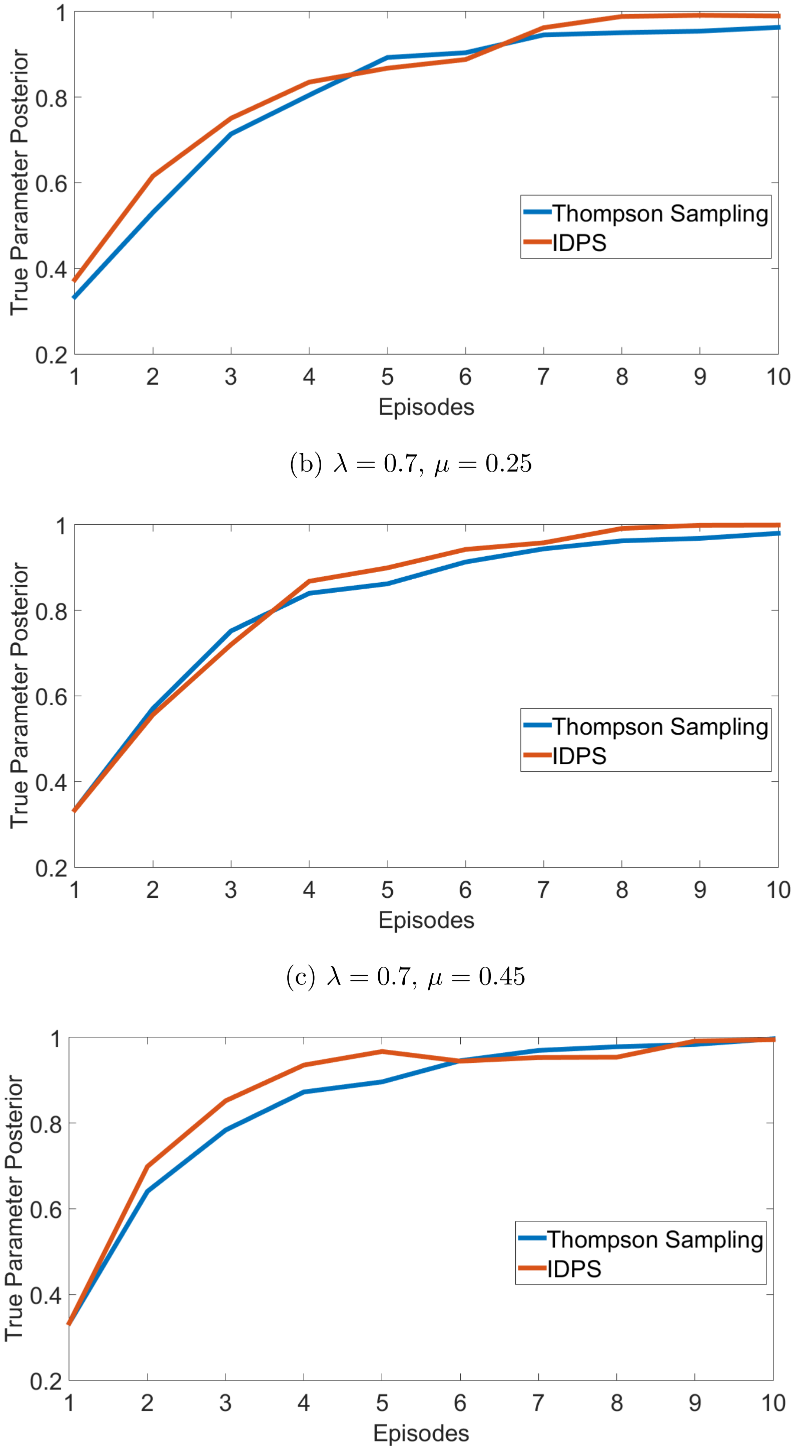
In this section, the decision-maker is uncertain about the measurement outcome distribution *and* the transition distribution. This is a generalization of the scenarios in Section 3.4.2 and 3.4.3. In this case, $\lambda \in \Omega$, and $\mu \in \mathcal{Q}$. Therefore, $\delta(\lambda, \mu) = \delta(\lambda, \mu)$. Algorithm 6 presents the algorithm for this case. I demonstrate this on the responde-guided dosing application as in section 3.4.1.

Algorithm 6 Information Directed Policy Sampling

```
Require: MDPs $\mathcal{Q}_\mu = \{S, A, T_\lambda, R, N, O, Z_\mu\}$ for $\lambda \in \Omega$. Prior pmf $\delta_1(\cdot,\cdot) = \alpha(\cdot)\beta(\cdot)$.
    Initial belief state $x_1(s), s \in S$.
$x_1(s) = p(s), \forall s \in S$

function IDPS
    for episode $k = 1, 2, 3, \cdots$ do
        Set $t = 1$
        Initialize $x_1(s)$; and prior $\delta_1(\cdot,\cdot) \leftarrow \delta_{N+1}(\cdot,\cdot)$ if $k > 1$
        repeat
            Compute distribution $\nu_*^t = \operatorname*{argmin}_{\nu^t \in D^t} \Psi_t(\nu^t | o_t, x_t, \delta_t(\cdot,\cdot))$
            Sample $\pi^t = (\pi_t, \ldots, \pi_N) \sim \nu_*^t$
            Implement action $\pi_t(x_t)$
            Observe $o_{t+1}$ drawn from $P_{\lambda^*,\mu^*}(\cdot | x_t, \pi_t(x_t))$
            Estimate $x_{t+1}$ as in equation 3.10
            Update probability mass $\delta_{t+1}(\lambda, \mu) \propto P_{\lambda,\mu}(o_{t+1} | x_t, \pi_t(x_t))\delta_t(\lambda, \mu)$, for each $\mu \in$
    $\mathcal{Q}$, and $\lambda \in \Omega$
            t ← t+1
        until end of horizon $N$
    end for
end function
```

Figure 3.3: Posterior for true parameter values.

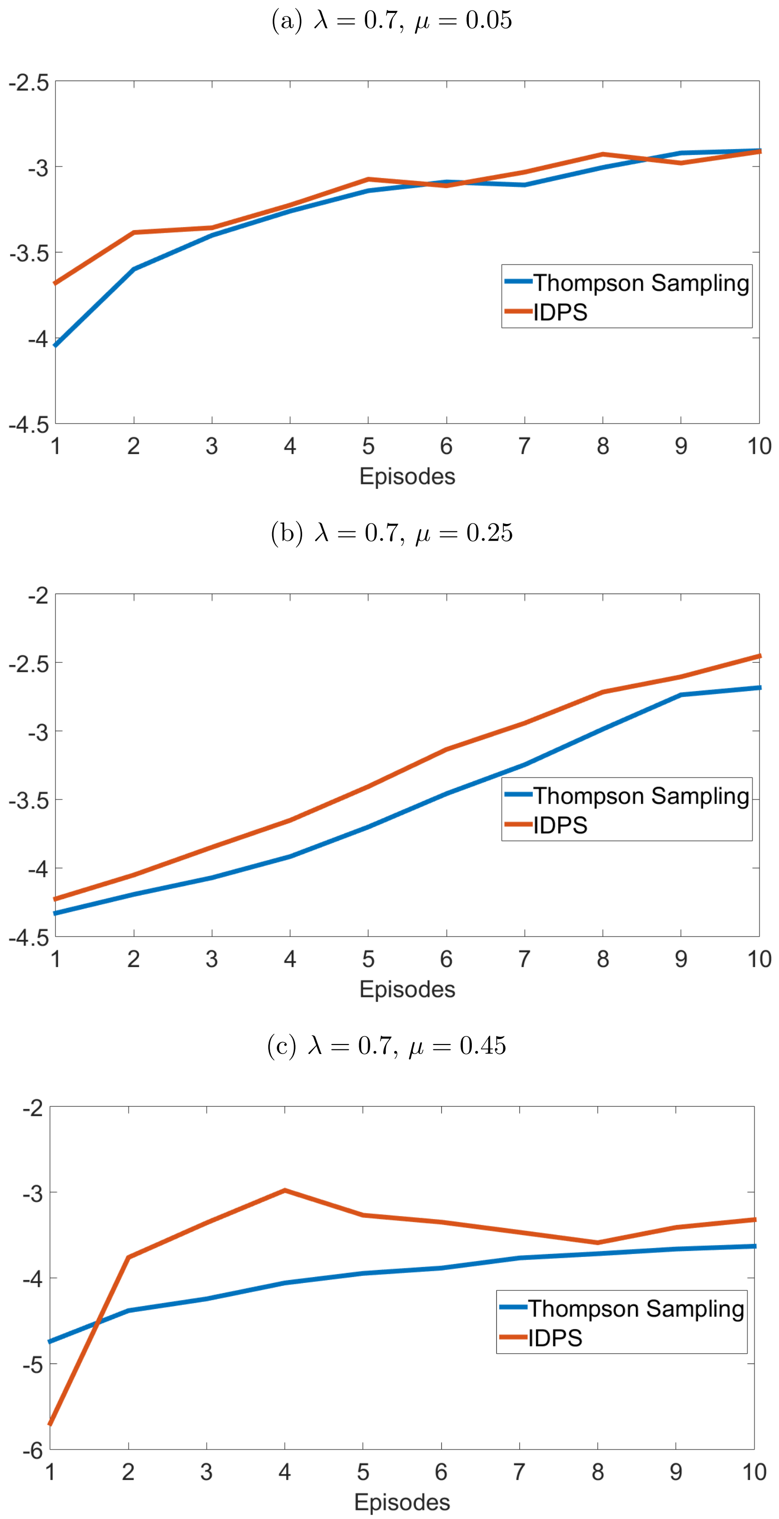


(a) $\lambda = 0.7,\ \mu = 0.05$

(b) $\lambda = 0.7,\ \mu = 0.25$

(c) $\lambda = 0.7,\ \mu = 0.45$

Figure 3.4: Averaged Cumulative Reward



(a) $\lambda = 0.7$, $\mu = 0.05$

(b) $\lambda = 0.7$, $\mu = 0.25$

(c) $\lambda = 0.7$, $\mu = 0.45$

Numerical experiments for response-guided dosing

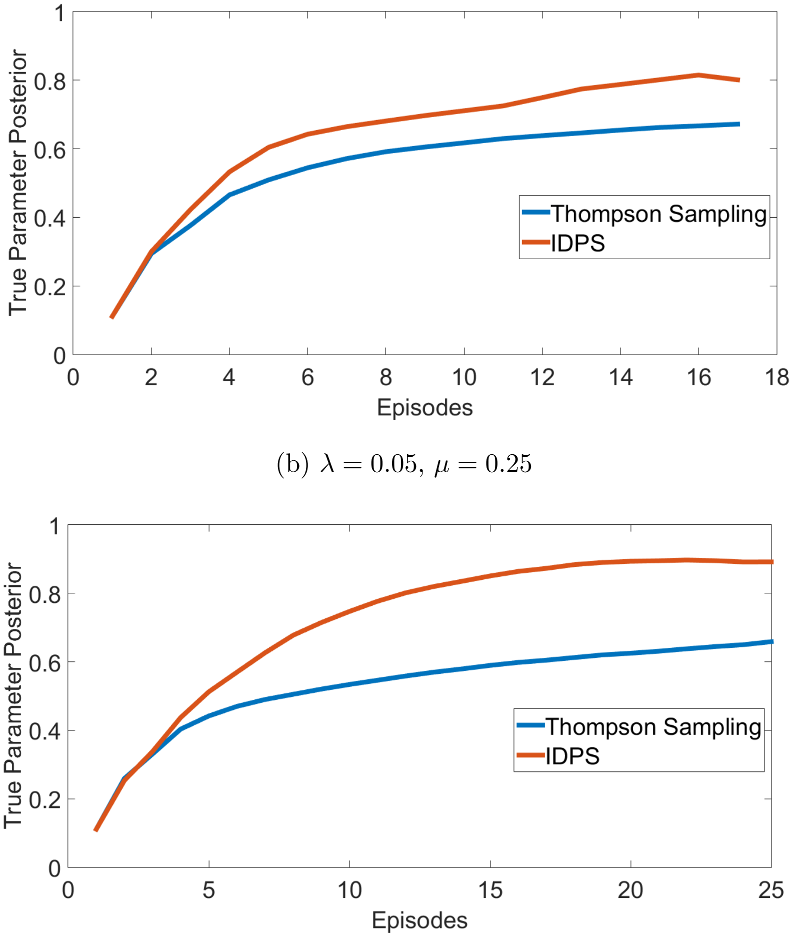
Now the planner is uncertain about the parameter value μ that characterizes the patient's disease and toxicity level measurement outcomes, as well as the transition distribution characterizing the patient's disease evolution and side effect evolution λ. The disease maker knows that $\mu \in \mathcal{Q} = \{0.05, 0.1, \ldots, 0.4, 0.45\}$, $\lambda \in \Omega = \{0.1, 0.2, 0.3, \ldots, 0.8, 0.9\}$. Figure 3.6 plots the cumulative reward averaged over 50 runs for IDPS and Thompson Sampling. Figure 3.5 plots the posterior of the true parameter values for IDPS and Thompson Sampling. We observe that for MDPs with partial observability IDPS learns faster than Thompson Sampling approaches which is consistent with the objective of this research.

3.5 Conclusion and future work

This chapter discussed learning in POMDPs with parametric uncertainty using information theoretic principles. I extended the IDPS approach developed in Chapter 1 to POMDPs. I analyzed three separate cases where the decision-maker is uncertain about the transition distribution or the measurement outcome distribution or both. This demonstrates the generalizability of information theoretic approaches, specifically IDPS. Future work can develop methods to make the IDPS more efficient in practice by rolling in the computation of information gain and expected regret in an online learning framework.

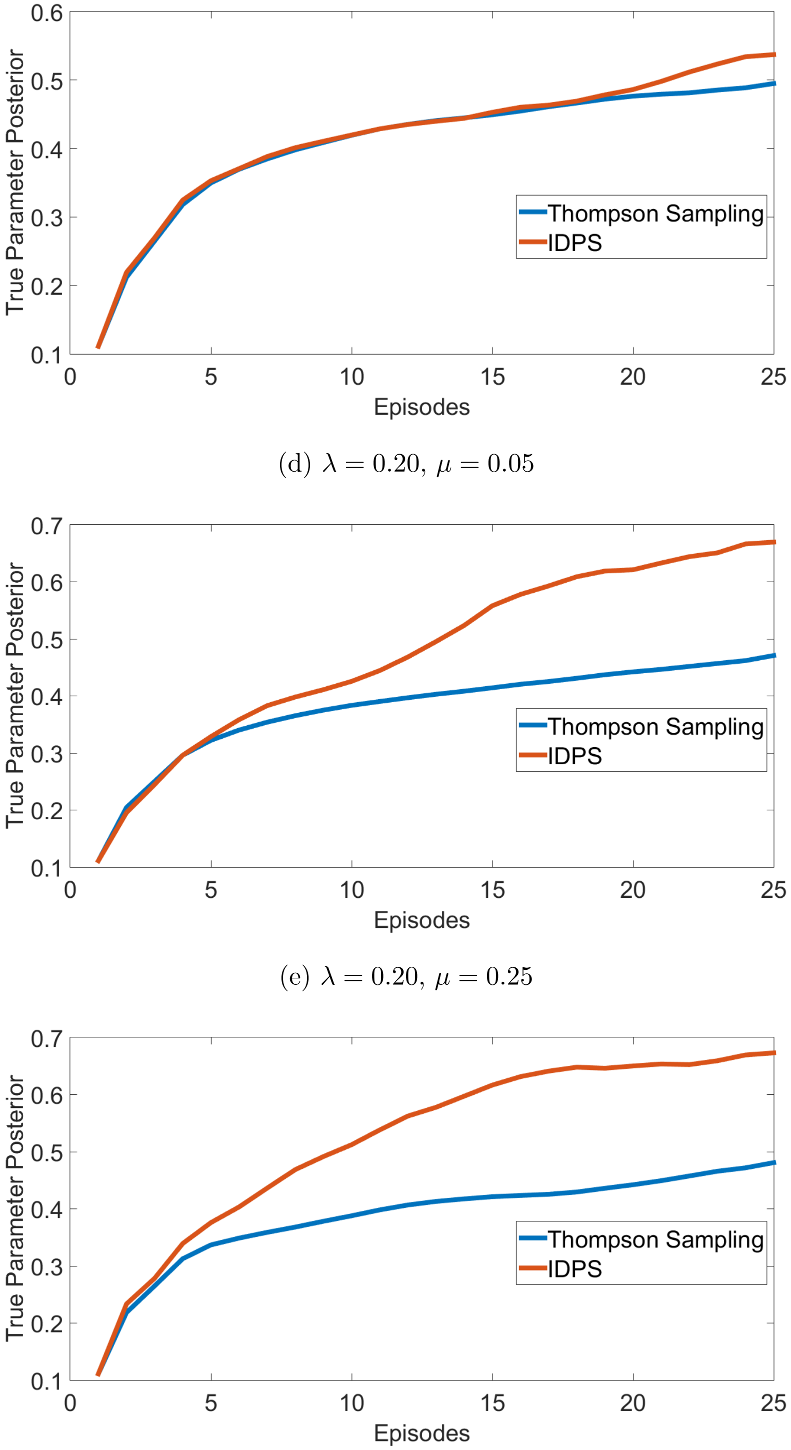
Figure 3.5: Unknown Transition and Measurement Outcome Distributions: Posterior for true parameter values.

(a) $\lambda = 0.05$, $\mu = 0.05$



(b) $\lambda = 0.05$, $\mu = 0.25$

Figure 3.5: Unknown Transition and Measurement Outcome Distributions: Posterior for true parameter values.

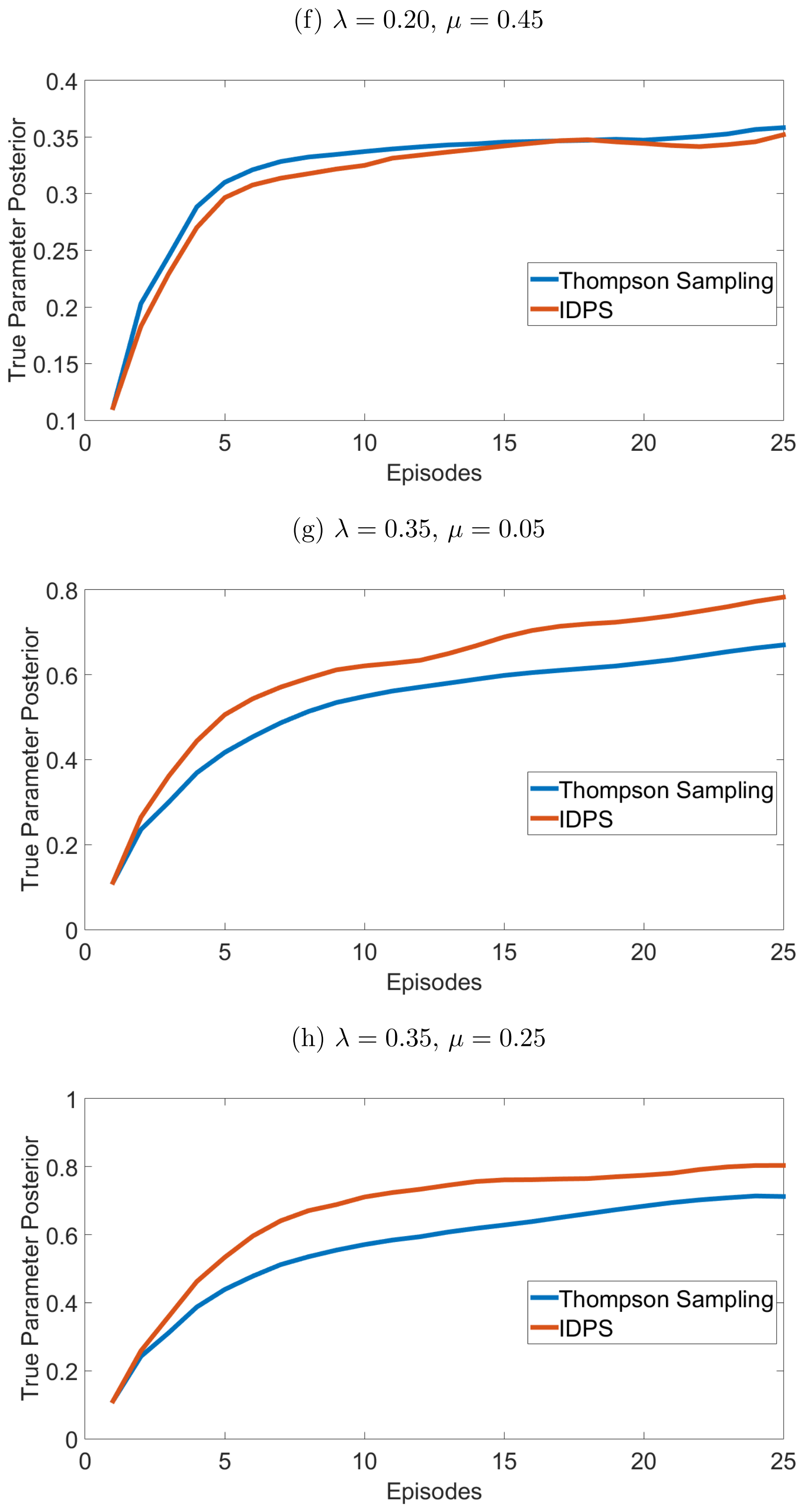


(c) $\lambda = 0.05$, $\mu = 0.45$

(d) $\lambda = 0.20$, $\mu = 0.05$

(e) $\lambda = 0.20$, $\mu = 0.25$

Figure 3.5: Unknown Transition and Measurement Outcome Distributions: Posterior for true parameter values.



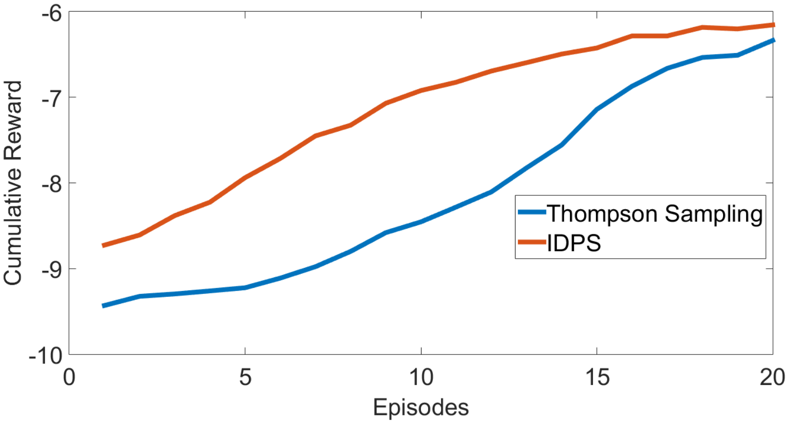
(f) $\lambda = 0.20$, $\mu = 0.45$

(g) $\lambda = 0.35$, $\mu = 0.05$

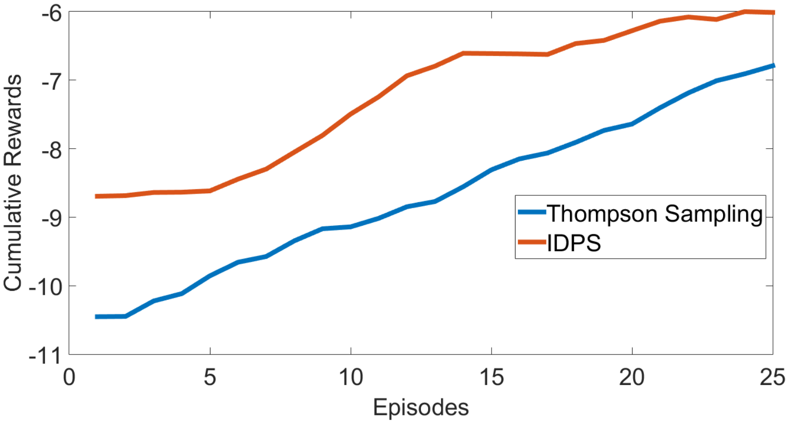
(h) $\lambda = 0.35$, $\mu = 0.25$

Figure 3.6: Unknown Transition and Measurement Outcome Distributions: Averaged Cumulative Reward

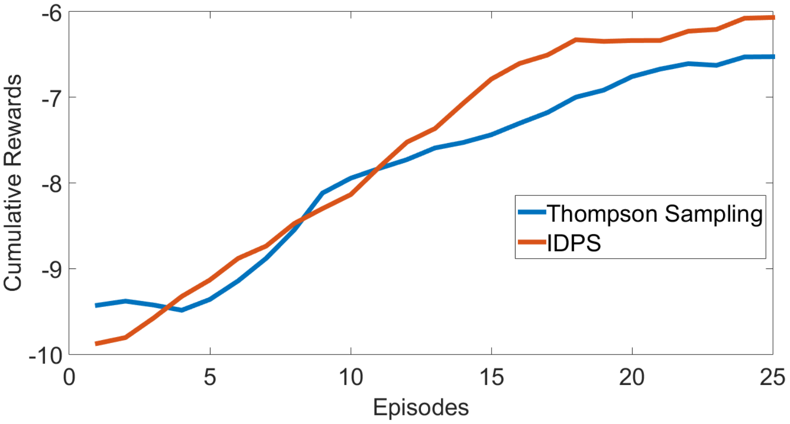
(a) $\lambda = 0.05$, $\mu = 0.05$



(b) $\lambda = 0.05$, $\mu = 0.25$



(c) $\lambda = 0.05$, $\mu = 0.45$



Chapter 4

STOCHASTIC THERMODYNAMICS-INSPIRED INFORMATION THEORETIC LEARNING IN MDPS

The last three chapters provided a strong evidence in the potential for using information theoretic approaches for learning in MDPs with parametric uncertainty. The key entity that includes the cost/benefit of information gain and the actual performance of the method is the *Information Ratio* Ψ. This quantity yields a convex optimization program that can be solved at each time step of the MDP to get an optimal policy which balances the trade-off between exploration and exploitation. The construction of *information ratio* comes from heuristic arguments and is not grounded in any fundamental principles of system/information dynamics, but I used this quantity to prove the benefits of using information theoretic approaches. In this chapter, I take inspiration from stochastic thermodynamics to derive a problem formulation for online learning in uncertain MDPs while grounded in system dynamics.

To this effect, I make an explicit link between the information entropy and the stochastic dynamics of a system coupled to an environment. I analyze various sources of entropy production: due to the decision-maker's uncertainty about the system-environment interaction characteristics; due to the stochastic nature of system dynamics; and the interaction of the decision maker's knowledge with system dynamics. This analysis provides a framework that can be formulated either as a maximum entropy program to derive efficient policies that balance the exploration and exploitation trade-off, or as a modified cost optimization program that includes informational costs and benefits. This work provides a more grounded reformulation of the IDPS ideas developed in Chapter 1, and bind the structural aspects of information developed in Chapter 2, and Chapter 3 into a generic framework.

The challenge in any Bayesian learning approach is that there is no clear consensus on the

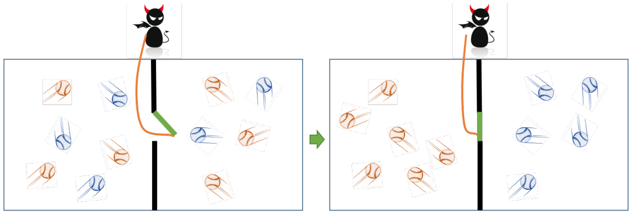


Figure 4.1: Maxwell's demon

actual problem that needs to be solved. Generally, we want to find a policy that *maximizes* cumulative reward while learning with uncertain or partial information. BAMDPs provide a classic formulation of this problem. But as discussed in Chapter 1 this formulation does not take into account the cost of information gain, and hence intuitively one can find better policies that leverage information gain while learning. The information theoretic methods developed so far rely on the heuristic idea of information ratio, which, I believe, is somewhat ad-hoc. In addition, this ratio does not give a strong insight into the global problem that is being solved. I find a relation between the optimization of reward and the cost of information that is embedded in the dynamics of the system and its interaction with the

4.1 Physical nature of information

In order to motivate the idea of the physical nature of information, I dive into the role of information in thermodynamics of gases. To guide the reader, a natural connection is to interpret particle configurations in gas systems as sample trajectories in stochastic system. Physicist Ludwig Boltzmann showed that with time, a system evolves towards lower states of energy, where the energy dispersed increases the entropy of the system due to the nature of statistics (Boltzmann, 1974). As Wolchover (2017) commented, " *There are many ways for energy to be spread among the particles in a system than concentrated in a few, so as particles move around and interact, they naturally tend toward states in which their energy is increasingly shared. This has been classically understood as the second law of thermodynamics. But Maxwell's letter (Maxwell, 1921) described a thought experiment in which an* *enlightened being, called Maxwell's demon (Figure 4.1), uses its knowledge to lower entropy and violate the second law. The demon knows the positions and velocities of every molecule in a container of gas. By partitioning the container and opening and closing a small door between the two chambers, the demon lets only fast-moving molecules enter one side, while allowing only slow molecules to go the other way. The demon's actions divide the gas into hot and cold, concentrating its energy and lowering its overall entropy. The once useless gas can now be put to work. This thought experiment lead to questions on how a law of nature could depend on one's knowledge of the positions and velocities of molecules. [This implies that second law of thermodynamics require a reinterpretation to include the subjective nature of information.] Charles Bennett (Bennett, 1987), building on work by Leo Szilard (Szilárd, 1976) and Rolf Landauer (Landauer, 1961), resolved the paradox by formally linking thermodynamics to the science of information. Bennett argued that the demon's knowledge is stored in its memory, and memory has to be erased, which takes work. (Landauer, 1961) calculated that at room temperature, it takes at least 2.9 zeptojoules of energy for a computer to erase one bit of stored information.) In other words, as the demon organizes the gas into hot and cold and lowers the gas's entropy, its brain burns energy and generates more than enough entropy to compensate. The overall entropy of the gas-demon system increases, satisfying the second law of thermodynamics. These findings revealed that, as Landauer put it, "Information is physical" (Landauer, 1991). More information implies that more work can be extracted. Maxwell's demon can wring work out of a single-temperature gas because it has far more information than the average user."*

This interaction of entropy and dynamics, capture by the second law, creates a strong foundation to analyze stochastic systems. There is a natural equivalence between stochastic thermodynamics and stochastic control theory. Any decision process can be modeled as a classic control problem. Generally, the quantities which are of interest are averaged over trajectories of the system rather than sample path behaviors. Thermodynamics has provided an intuitive framework and solution about averaged entities on stochastic systems. I study this equivalence and bridge gaps in the existing literature on learning in MDPs. I develop an equivalent thermodynamic system and apply an information theoretic framework to find a formulation of the learning problem to compute good policies.

There has been some work in the literature to bridge this gap between control theory and stochastic thermodynamics. Brockett and Willems (1979) studies second law of thermodynamics from the point of view stochastic control theory. They compute a criterion which, when satisfied, permits one to assign a temperature to a stochastic system in a way that Carnot cycles become the optimal trajectories of optimal control problems. Propp (1985) also studied the connection between thermodynamic and Markovian systems. There, an input-output framework for thermodynamics was proposed, which allowed to introduce the notion of states, controls and response, thus drawing a connection between the two fields. There has also been a recent surge in understanding the field of stochastic thermodynamics to study Markovian processes at the trajectory level using statistical quantities (Seifert et al., 2011; Aurell et al., 2012). Saridis (1988) proposed a formulation that gives a generalized energy interpretation to the optimal control problem. This framework provides compatibility between the control problem and the information theoretic methodology for the intelligent control system using entropy as the common measure. A reformulation of the optimal control problem is based on the idea of expressing the design of the desirable control by the uncertainty of selecting a control law that minimized a given performance index.

4.2 Clairvoyant MDP: an information theoretic perspective

Consider the Bellman's equation for MDP $M = \{S, A, T, R, N\}$.

$$V^*(s) = \min_a \sum_{s'} T(s'|s,a)[R(s'|s,a) + V^*(s')]. \tag{4.1}$$

I consider an alternate formulation to this classical MDP, with a small loss of generality. Todorov (2009) proposed a linear problem where actions that are considered symbolic in the above formulation are replaced through making decisions over transition distributions. Therefore, the decision maker specifies a control dynamics distribution $a(s'|s) = T(s'|s,a)$.

This allows us to write an equivalent reward form as

$$q(s,a) = \ell(s) + \underset{s'\sim a(\cdot|s)}{E} \ln\left(\frac{a(s'|s)}{p(s'|s)}\right),$$

where the state cost $\ell(s)$ is an arbitrary function encoding how undesirable different states are and $p(s'|s)$ is an arbitrary transition distribution. Using this construction the Bellman's equation can be rewritten as:

$$V^*(s) = \min_a \left(\ell(s) + \underset{s'\sim a(\cdot|s)}{E}\left[\ln\frac{a(s'|s)}{p(s'|s)} + V^*(s')\right]\right). \tag{4.2}$$

Now, I define the quantity $G(s) = \underset{s'\sim p(\cdot|s)}{E} exp(-V^*(s'))$. Therefore, through some algebraic manipulation, I get

$$\underset{s'\sim a(\cdot|s)}{E}\left[\ln\frac{a(s'|s)}{p(s'|s)} + V^*(s')\right] = -\ln(G(s)) + \mathbb{KL}\left(a(\cdot|s)||\frac{p(\cdot|s)\exp(-V^*(\cdot))}{G(s)}\right),$$

which gives

$$V^*(s) = \min_a \left[\ell(s) - \ln(G(s)) + \mathbb{KL}\left(a(\cdot|s)||\frac{p(\cdot|s)\exp(V^*(\cdot))}{G(s)}\right)\right]. \tag{4.3}$$

An interesting observation is that the right hand side of the above function is minimized when the KL divergence is 0, which gives the optimality condition as

$$a^*(s'|s) = \frac{p(s'|s)\exp(-V^*(s'))}{G(s)} \tag{4.4}$$

$$= \frac{p(s'|s)\exp(-V^*(s'))}{\sum_{s'} p(s'|s)exp(-V^*(s'))} \tag{4.5}$$

Now consider the following Lemma (Theodorou and Todorov, 2012; Theodorou, 2015).

Lemma 4.2.1. *Consider distributions $\mathbb{A}$ and $\mathbb{P}$ defined on the same probability space with sample set Ω, such that $\mathbb{A}$ is absolutely continuous with respect to $\mathbb{P}$, and $Q : \Omega \mapsto \mathbb{R}$ is a* *measurable function, then the following inequality holds*

$$\frac{1}{\rho}\ln\left(\underset{\mathbb{P}}{E}\left[e^{\rho Q(s)}\right]\right) \leq \underset{\mathbb{A}}{E}[Q(s) + |\rho|^{-1}\mathbb{KL}(\mathbb{A}||\mathbb{P})],$$

where $\rho \in \mathbb{R}^-$.

Proof. The proof is reproduced here for completeness. It is a straightforward derivation from Jensen's inequality. Consider,

$$\begin{aligned}
\ln\left(\underset{\mathbb{P}}{E}\left[e^{\rho Q(s)}\right]\right) &= \ln\sum_s p(s)e^{[\rho Q(s)]} \\
&= \ln\left[\sum_s a(s)\frac{p(s)}{a(s)}\exp\left(\rho Q(s)\right)\right] \\
&\overset{a}{\geq} \sum_s a(s)\ln\left[\frac{p(s)}{a(s)}\exp\left(\rho Q(s)\right)\right] \\
&= \rho\underset{\mathbb{A}}{E}[Q(s)] + \sum_s a(s)\ln\frac{p(s)}{a(s)} \\
&= \rho\left(\underset{\mathbb{A}}{E}[Q(s)] - \rho^{-1}\mathbb{KL}(\mathbb{A}||\mathbb{P})\right),
\end{aligned}$$

where inequality "a" follows from Jensen's inequality and concavity of the ln function. Now dividing both sides by $\rho \in \mathbb{R}^-$ gives the required inequality.

□

Next, consider Equation (4.2), where I substitute $Q(s') = \ell(s) + V^*(s')$. Now using Lemma 4.2.1 with $\rho = -1$, $\mathbb{P} = p(s'|s)$, $\mathbb{A} = a(s'|s)$, I get

$$-\ln\left(\underset{s'\sim p(\cdot|s)}{E}\left[e^{-\ell(s)-V^*(s')}\right]\right) \leq \sum_{s'} a(s'|s)\left[\ell(s) + V^*(s') + \ln\frac{a(s)}{p(s)}\right],$$

which implies

$$-\ln\left(\underset{s'\sim p(\cdot|s)}{E}\left[e^{-\ell(s)-V^*(s')}\right]\right) = \min_a \sum_{s'} a(s'|s)\left[\ell(s) + V^*(s') + \ln\frac{a(s)}{p(s)}\right].$$

The right hand side of the above equation is the right hand side of the Bellman equation in Equation (4.2). Therefore

$$V^*(s) = -\ln \left(\underset{s' \sim p(\cdot|s)}{E} \left[e^{-\ell(s) - V^*(s')} \right] \right).$$

This framework can also be used as an estimation framework where instead of minimizing the expected cumulative cost, the decision maker can maximize the KL divergence for a required performance. Therefore, the optimization problem in Equation (4.2) becomes

$$\min_{\mathbb{A}} \mathbb{KL}(\mathbb{A}||\mathbb{P}),$$

subject to

$$\sum_{s'} a(s'|s) = 1,$$

$$V(s) = K,$$

where K is the required performance. In the interest of providing interesting connection, I consider continuous optimization. Using the Lagrangian method, the optimization program reduces to

$$\begin{aligned} \mathcal{L} &= \mathbb{KL}(\mathbb{A}||\mathbb{P}) + \mu(V(s) - K) + \lambda(\int_{s'} a(s'|s) ds' - 1) \\ &= -\int_{s'} a(s'|s) \left(ln \frac{a(s'|s)}{p(s'|s)} + \mu V(s) + \lambda \right) ds' + \mu K + \lambda \end{aligned}$$

Now, maximizing with respect to $a(s'|s)$ gives

$$ln \frac{a^*(s'|s)}{p(s'|s)} + \mu V(s) + \lambda = 0,$$

which gives

$$a^*(s'|s) = \int exp(-muV(s) - \lambda) p(s'|s) ds'.$$

Substituting in the first constraint for $\int a(s'|s)ds = 1$, gives

$$\lambda = \ln \int p(s'|s) \exp(-\mu V(s))ds'.$$

Substituting λ back and discretizing gives the optimal solution for a^* for a given level of performance. In the case where $K = V^*(s)$, this solution gives the optimal a^* as in Equation 4.5. This result is very similar to the one derived using HJB principle in the classic paper by Saridis (1988). For the reader's convenience, I recall that result in Appendix 4.6.

4.3 Thermodynamics of information

This section provides a brief introduction on the relationship between information and thermodynamics. We consider a system M (such as a gas in a container) that is connected to external reservoirs and other systems. Suppose the microstate of the system (for example, the coordinates and momentum of particles of the gas) is given by x, and suppose that the information gained as a result of measurement is denoted by m. This measurement is what helps to prepare the state of the system. Let us denote a generic statistical state of the system with $\rho(x)$ (for example, the distribution of coordinate states and momentum of the gas molecules). I assume that in state $\rho(x)$ the system is in statistical equilibrium. Now after making the measurement, the new state of the system in $\rho(x|m)$, which in general is out of equilibrium. For example, in the context of the Sczilard's engine described in Section 4.1, after measurement the statistical state is confined to either the left or right half of the box. Information drives the system away from equilibrium. The thermodynamics of information allows us to reason about this scenario by associating an equivalent energy cost, thus justifying this movement from equilibrium to a non-equilibrium state.

The most obvious entity that relates statistical entities to distributions is the entropy of the system. In this case, the non equilibrium entropy is defined using a scaled version of the Shannon Entropy as

$$S(\rho) = -\sum_x \rho(x) \ln \rho(x) = H(X),$$

where $H(X)$ is the Shannon entropy. At equilibrium this entropy coincides with the canonical entropy

$$\rho(x) = \exp^{-\beta E(x)} / Z,$$

where $E(x)$ is the Hamiltonian of the system, and Z is the partition function, and β is the inverse temperature. Using this we recover the thermodynamic relationship between Free energy $\mathcal{F}(\rho) = -\beta^{-1} \ln Z$, and internal Energy $E = E[H]$ and Entropy: $\mathcal{F} = E - \beta^{-1} S$. The free energy is interpreted as the amount of useful energy that can be used to extract work, taking in account all entropy related costs. The classical second law of thermodynamics for non equilibrium system, therefore, can be written as

$$\Delta S \geq 0 \implies W - \Delta \mathcal{F} \geq 0, \tag{4.6}$$

where W is the average work done on the system.

The rest of the section evaluates the change in non-equilibrium free energy due to a measurement M. For this purpose the corresponding information gain is defined as

$$I(X; M) = H(X) - H(X|M).$$

Now, in the event that an external system changes the system parameter after an observation is made, results in work extracted from the system. The refined second law of thermodynamics then becomes

$$W - \Delta \mathcal{F} \geq -\beta^{-1} I(X; M) \tag{4.7}$$

An interesting observation is that ultimately, the information used to extract work during feedback was supplied as work by the external system during the measurement process.

4.4 Markovian systems and second law of thermodynamics

Now let's consider the second law of thermodynamics $W \geq \Delta\mathcal{F}$ without feedback (Equation 4.6), and compare it with the Lemma 4.2.1

$$\frac{1}{\rho}\ln\left(\underset{\mathbb{P}}{E}\left[e^{\rho Q(s)}\right]\right) \leq \underset{\mathbb{A}}{E}[Q(s) + |\rho|^{-1}\mathbb{KL}(\mathbb{A}||\mathbb{P})].$$

The quantity on the left hand side is the Free Energy change $\Delta\mathcal{F}$ and the work done on the system is the expected cumulative cost given by the right hand side of the equation. Substituting the relevant entities for the MDP defined in Section 4.2 provides a bridge between MDP and the respective thermodynamic interpretation. Therefore, using the mathematical equivalence, the policy that minimizes work done (or maximum work extracted from the system) gives the optimal solution for the MDP.

The above results give sufficient evidence to explore equivalence between thermodynamic entities and Markov Decision Processes. In order to develop a learning framework in uncertain MDPs using information theoretic arguments, I develop the definition of thermodynamic quantities at the level of sample trajectories for Markovian system in the next section.

4.4.1 Second law of thermodynamics for a Markovian system in a heat bath

This section reviews the stochastic thermodynamics for Markovian Systems (Ito and Sagawa, 2016). Stochastic Thermodynamics is a theoretical framework to define quantities such as work and heat at the level of sample trajectories.

Consider a system M that evolves stochastically. We assume a physical situation where system M is connected to a single heat bath at inverse temperature β. Also assume that the system M is driven by an external parameter π and the system is not subject to non-conservative forces. For simplicity, we will assume discrete time $t_k, k = \{1, 2, \cdots, N\}$, although, the mathematical setup does not force any assumption regarding the continuity of time. Let x_k be the state of the system at time t_k, and π_k be the external parameter of the system at time t_k. Let $p(x_k|x_{k-1}, \pi_k)$ be the conditional probability of state x_k under the past trajectory and external parameter π_k.

Now building on the thermodynamic principles, we define the Hamiltonian of the system as $E(x_k, \pi_k)$. The Hamiltonian change in the system is decomposed into 2 parts heat Q_k and work W_k. The heat absorbed by the system from heat bath at time t_k is defined as

$$Q_k = E(x_{k+1}, \pi_k) - E(x_k, \pi_k),$$

and the work done on the system M is defined as

$$W_k = E(x_k, \pi_k) - E(x_k, \pi_{k-1}).$$

For a given trajectory $\{x_1, x_2, \cdots, x_n\}$ the total heat is $Q = \sum_{i=1}^{N-1} Q_k$ and total work is $W = \sum_{i=1}^{N-1} W_k$, where x_0 is defined as a buffer state such that $p(x_1|x_0, \pi) = 1$ for any π. This is done to impose consistency as it will become apparent later on.

Using the above definitions, one can easily show that $\Delta E_k = Q_k + W_k$, which is the first law of thermodynamics.

Now let us define the quantity $p_B(x_k|x_{k+1}, \pi_k)$ as the backward transition probability. In the absence of any non conservative the detailed balance (Seifert, 2005) is satisfied which gives

$$\frac{p(x_{k+1}|x_k, \pi_k)}{p_B(x_k|x_{k+1}, \pi_k)} = e^{-\beta Q_k}.$$

Now, I define the stochastic entropy of the system as $h(x_k) = -ln(x_k)$. Therefore the *entropy production* is defined as the sum of stochastic entropy change in the system and the bath. The stochastic entropy change in the system is given by

$$\Delta h_k^M = h(x_{k+1}) - h(x_k).$$

The total stochastic entropy change therefore is given by

$$\Delta h^M = ln\frac{p(x_1)}{p(x_N)}. \tag{4.8}$$

The stochastic entropy change in the heat bath is given by the heat dissipation into the bath

$$\Delta h_k^{bath} = -\beta Q_k.$$

The total entropy change in the bath is given by

$$\Delta h^{bath} = ln\frac{p(x_N|x_{N-1},\pi_{N-1})p(x_{N-1}|x_{N-2},\pi_{N-2})\dots p(x_2|x_1,\pi_1)}{p_B(x_1|x_2,\pi_1)p_B(x_2|x_3,\pi_2)\dots p(x_{N-1}|x_N,\pi_{N-1})}. \tag{4.9}$$

Therefore the entropy production σ is

$$\sigma = ln\frac{p(x_N|x_{N-1},\pi_{N-1})p(x_{N-1}|x_{N-2},\pi_{N-2})\dots p(x_2|x_1,\pi_1))p(x_1)}{p_B(x_1|x_2,\pi_1)p_B(x_2|x_3,\pi_2)\dots p(x_{N-1}|x_N,\pi_{N-1})p(x_N)}.$$

For brevity I define the trajectory of the system as $O = \{x_1, x_2, \dots, x_N\}$. Therefore the total entropy production becomes

$$\sigma = ln\frac{p(O)}{p_B(O)}.$$

Therefore, the entropy production is determined by the ratio of the probabilities of a trajectory and its time-reversal.

Simple algebraic calculation on this definition yields the second law of thermodynamics which states that

$$E[\sigma] \geq 0.$$

The equivalent stochastic energetics definition gives the form as in Equation (4.6)

$$W \geq \Delta\mathcal{F},$$

where $\mathcal{F}(\lambda_k) = -\beta^{-1} \ln \sum_X \exp(-\beta E(x, \lambda_k))$. This result can be derived using the integral fluctuation theorem and the arguments presented in Seifert (2005).

4.4.2 Second law of thermodynamics for a Markovian system in connection with an external entity

Here I consider the Markovian System M in contact with an external system D in addition to the heat bath. This external system, for instance can be the decision maker in the context of the MDP (More on this in the later sections). In particular, I state the generalized second law of thermodynamics, which states that the entropy production is bounded by the initial and final mutual information between M and D, and the transfer entropy from M to D.

Let's consider the states of the system D at time t_k be d_k. Therefore, the joint time evolution of system $M \cup D$ is defined as $\{(x_1, d_0), (x_2, d_1), \cdots, (x_N, d_{N-1})\}$. For brevity, I define $pa(x_{k+1})$ as the parent of state x_{k+1} which is the set of all states which has a non zero transition probability to x_{k+1}, therefore $pa(x_{k+1}) = \{x_k, d_{k-1}\}$, such that $p(x_{k+1}|x_k, d_{k-1}) > 0$.

At the initial state I assume that $pa(x_1) \subseteq D$. The initial correlation between system S and D is then characterized by the mutual information between x_1 and $pa(s1)$. The corresponding stochastic mutual information is given by

$$I_{ini} = I(x_1; pa(x_1)).$$

Now, let's define $an(x_{k+1})$ as the ancestors of x_k in the order that they were observed. Therefore $an(x_{k+1}) = \{(x_1, d_0), (x_2, d_1), \cdots, (x_k, d_{k-1})\}$. The final correlation between system S and D is then characterized by the mutual information between x_N and $an(x_N) \cap D$.

$$I_{fin} = I(x_N; \{d_0, d_1, \cdots, d_N\}).$$

Let's define another quantity $pa(d_k)$as the parent of d_k that corresponds to $pa(d_k) = \{x_{k-1}, d_{k-1}\}$Finally, I define the transfer entropy from M to D as

$$I_{tr}^k = I(d_k; pa(d_k) \cap M | d_1, d_2, \cdots, d_{k-1}).$$

The total transfer entropy for the entire dynamics is therefore given by

$$\sum_{k=1}^{N} I_{tr}^k = I_{tr}.$$

By combining all the above informational content in the combined system, I define the total informational exchange as

$$\Theta = I_{fin} - I_{tr} - I_{ini}.$$

Now, as in the simple case in Section 4.4.1, I define the entropy production in system M and the heat bath, while in the presence of system D.

Let $\mathcal{B}_{k+1} \subseteq D$ define the set of states in D that effect x_{k+1}, therefore $\mathcal{B}_{k+1} = \{d_{k-1}\}$. Now $p(x_{k+1}|x_k, \mathcal{B}_{k+1})$ describes the transition probability from x_k to x_{k+1} under the condition that the states of D that affect M are given by $\mathcal{B}_{k+1}$. We then define the backward transition probability as $P_B(x_k|x_{k+1}, \mathcal{B}_{k+1})$. Following the definition of entropy change in the heat bath from time k to $k+1$ as in Equation (4.9) is given by:

$$\Delta s^{bath} = \sum_k x_k^{bath}$$

$$= \sum_k \ln \frac{p(x_{k+1}|x_k, B_{k+1})}{p_B(x_k|x_{k+1}, B_{k+1})}.$$

The total entropy change in the system M is similar to Equation (4.8)

$$\Delta s^{sys} = \ln \frac{p(x_1)}{p(x_N)}.$$

The total entropy production is therefore,

$$\sigma = \ln \frac{p(x_1)}{p(x_N)} \Pi_k \frac{p(x_{k+1}|x_k, B_{k+1})}{p_B(x_k|x_{k+1}, B_{k+1})}.$$

Now, we can write the refined second law of thermodynamics, through some algebraic manipulation it can be shown that

$$E[\sigma] \geq \Theta.$$

Using the integral fluctuation theorem and theory of stochastic energetics, this result can be restated as in Equation (4.7)

$$W - \Delta\mathcal{F} \geq -\beta^{-1}\Theta \tag{4.10}$$

4.5 MDP with uncertainty: a stochastic thermodynamics perspective

The framework in the previous section provides a way to model the effect of information gain in MDPs with uncertainty with the objective of maximizing the work that can be extracted out of the system. The system M considered in the previous section is the system that is acting in the real environment, the system D is the decision maker, who changes some parameter of the system M in order to achieve the required objective. Both these systems are suspended in a "heat bath" to account for the part of the work that is dissipated and cannot be used for any useful work. The thermodynamic framework allows us to define the objective of the optimization program when the MDPs have model uncertainty. To be consistent, the uncertainty in the MDPs is assumed to be completely reflected through the uncertainty in the transition probabilities. In this section, I propose 2 different perspectives of how the system D interacts with system M: a) the first perspective is where system D directly maintains a distribution over the policies and changes this distribution based on feedback; b) the second perspective is where the system D maintains a distribution over a parameter of the transition distribution and adapts this based on feedback in order to find a good policy. The case "b" is consistent with the frameworks in previous chapter. Although, as we will see, both cases "a" and "b" generate the same results and therefore, are interchangeable from a modeling perspective.

4.5.1 MDP with distribution over policies

Consider an MDP $M = \{S, A, T, R, N\}$[1] and the decision maker $D = \{\pi\}$. The decision maker maintains a probability distribution $\nu_k(\pi|s_k)$ over policies π at every time step t_k in state s_k. The probability distribution is updated based on feedback. This setup is analogous to the thermodynamic setup described in Section 4.4.2. In a standard MDP, the objective is to minimize the expected cumulative cost

$$V^{\pi}(s_t) = \sum_{k=t}^{N-1} E[c(s_k, \pi(s_k))],$$

where the expectation is taken over $\{s_k, \pi(s_k)\}$, in terms of the classical discrete MDP $c(s_k, \pi(s_k)) = E_{s_{k+1} \sim T(\cdot|s_k, \pi(s_k))}[R(s_{k+1}|s_k, \pi(s_k))]$.

As in Section 4.2, the MDP problem for finding a policy to achieve the maximum *performance* can be formulated as either a maximum entropy optimization program or the classical expected cost optimization. In will start by formulating an expected cost optimization pro-

[1] Please note that for the purpose of this discussion I will consider R as the cost function (rather than the reward function)

gram using the Second Law of Thermodynamics. Equation 4.10 can be written as

$$W + \beta^{-1}\Theta \geq \mathcal{F}.$$

Note that the free energy $\mathcal{F}$ is the amount of useful energy, and the infimum of the left hand side will give the most amount of net work that can be extracted out of the system. Therefore, the optimization program becomes

$$\min_{\nu_t;t=1:N} W + \beta^{-1}\Theta,$$

where $W = \sum_{k=1}^{N-1} E[c(s_k, \pi(s_k))]$, $\Theta = I_{fin} - I_{tr} - I_{ini}$, and $\nu_t = p(\pi_t|s_t, \pi_{t-1})$. From previous section $x_k = \{s_k\}$, and $d_k = \pi_k$. For the classical MDP $p(s_1) = \delta(s_1 - s_{init})$, and $pa(s_1) = \emptyset$. Therefore,

$$I_{ini} = I(s_1; pa(s_1)) = 0.$$

The final information correlation is given by

$$I_{fin} = I(s_N; \pi_1, \cdots, \pi_{N-1}),$$

note that $p(\pi_1, \cdots, \pi_{N-1}) = \Pi_{i=2}^{N-1} p(\pi_i|\pi_{i-1})$.

$$I_{tr}^k = I(\pi_k; s_{k-1}|\pi_1, \cdots, \pi_{k-1}) = I(\pi_k; s_{k-1}|\pi_{k-1})$$

Therefore the optimization program becomes

$$\min_{\nu_t:t=1,\cdots,N} \left(\sum_{k=1}^{N-1} \big(E[c(s_k, \pi(s_k))] - \beta^{-1} I(\pi_{k+1}; s_k|\pi_k) \big) + \beta^{-1} I(s_N; \pi_1, \cdots, \pi_{N-1}) \right).$$

For the case, $I_{fin} = 0$, the solution to the resulting optimization program is discussed in Tanaka et al. (2017).

The above problem can reformulated as a maximum entropy principle, which translates to

$$\max_{\nu_t:t=1,\cdots,N} \sum_{k=1}^{N-1} I(\pi_{k+1};s_k|\pi_k) - I(s_N;\pi_1,\cdots,\pi_{N-1})$$

subject to

$$\sum_k \nu_t(\pi_k) = 1 \ \forall k,$$

$$\sum_{k=1}^{N-1} E[c(s_k,\pi_k)] = K,$$

where K is the required performance. When $K = V^*$, the resulting policy is the optimal policy with respect to the cost based optimization program.

4.5.2 MDP with parametric uncertainty

In this case the decision maker D maintains a distribution over the parameter of the system. This case is similar to the ones studied in previous chapters. Therefore the state of the system D is denoted by λ_k at time t_k. Again, the specific informational correlations are given by

$$I_{ini} = I(s_1; pa(s_1)) = 0,$$

$$I_{fin} = I(s_N;\lambda_1,\cdots,\lambda_{N-1}),$$

and

$$I_{tr}^k = I(\lambda_k;s_{k-1}|\lambda_1,\cdots,\lambda_{k-1}) = I_{tr}^k = I(\lambda_k;s_{k-1}|\lambda_{k-1}).$$

The optimization program becomes

$$\min_{\nu_t:t=1,\cdots,N} \left(\sum_{k=1}^{N-1} \big(E[c(s_k,\pi(s_k))] - \beta^{-1} I(\lambda_{k+1};s_k|\lambda_k)\big) + \beta^{-1} I(s_N;\lambda_1,\cdots,\lambda_{N-1}) \right),$$

where $\nu_t = p(\pi_t|s_t)$, and the distribution over λ is updated using Bayesian learning. As in the previous section this can also be formulated as a maximum entropy framework.

4.6 Discussion and future work

This chapter provides a framework for formulating an optimization program for solving uncertain MDPs built from fundamental principles of system dynamics and information theory. The exact formulation of the optimization program depends on the specific nature of interaction between the decision maker and the system to be controlled. Sections 4.5.1 and 4.5.2 provide optimization program for 2 different scenarios. Given these formulation, we can use many of the techniques for optimization (including the Bellman's principle) to solve for a solution. This will be a future work. An important discussion point is the entity β in the above equations. Thermodynamically, β capture the inverse temperature (with a scaling constant). The temperature is a property of the heat bath and assumed to be constant throughout the dynamic process. In the context of a decision process, the temperature is a property of the decision process and can be estimated. A good way to estimate temperature will be to find an *equilibrium* solution and solve it inversely to get the temperature. For instance, given a MDP $M = \{S, A, T, R, N\}$ one can chose the starting state for which there a solution is known apriori and that can be used to estimate the temperature of the decision process. In the event we do not have access to this knowledge, the temperature can be considered a pseudo state and a new MDP can be defined $M' = \{\{S, \beta\}, A, T', R', N\}$.

Another important point is that the above framework works when certain conditions on the underlying Markovian process is satisfied. One sufficient condition, as discussed in Section 4.4, is the detailed balance equation, which implies reversibility of the Markovian system. We know that is not a necessary condition, in fact, it can be shown that the results still hold for non-reversible Langevin dynamics. Additional research is required to state and prove the necessary and sufficient conditions for this framework to hold.

In conclusion, this work opens up avenues for further research in employing information theoretic arguments to learning in MDPs with model uncertainty. This work is the first comprehensive work, to the best of my knowledge, to explicitly model information content and system dynamics for MDPs. I provide a framework to formulate the optimality criterion for MDPs with model uncertainty. Hopefully, future work can extend the rich theory of MDPs to learn and make good decisions in the situations of information uncertainty.

Appendices

Appendix 1

MAXIMUM ENTROPY PRINCIPLE: A CONTROL THEORETIC APPROACH

A classic paper by Saridis (1988) derives a maximum entropy framework for Control systems. I present this here as it is interesting to see connections without using the definitions from Thermodynamics.

Consider a generic decision system formulated in a classic control theoretic framework. Assume that the dynamics of the system are deterministic for simplicity. Then the dynamics are given by

$$\dot{x}(t) = f(x, u, t), \quad x(t_0) = x_0$$

and the associated cost function is

$$V^*(x_0, u, t_0) = \int_{t_0}^{T} L(x, u, t)dt; \quad L(x, u, t) > 0.$$

Here, $x(t) \in X$ is a n-dimensional state vector and $u(t) : X \times T$ is the m dimensional feedback control law. The solution is to find a control law $u_k(x, t)$ such that the value function V will take a value K such that $V_{min} \leq K < \infty$.

$$V^*(x_0, t_0 | u_K(x(t), t) = K$$

This satisfies the Hamilton-Jacobi-Bellman equation

$$\frac{\partial V}{\partial t} + \frac{\partial V^T}{\partial x} f(x, u_k, t) + L(x, u_k, t) = 0.$$

In order to formulate the problem in entropy terms we consider the decision-maker's uncertainty of selecting the proper control from the set of admissible controls to satisfy the value function requirement to equal K (V_{min} in the case of optimal control). This may be expressed as a condition that the expected value of V equals K:

$$E_{u \sim p(u)}[V^*(x_0, t_0; u(x,t)] = K.$$

The expected value of V is taken over the set of admissible controls U, over which a probability density $p(u)$ is assumed to express the uncertainty of selecting the proper control. The corresponding entropy can then be expressed as

$$H(u,p) = -\int_U p(u) \ln p(u) du.$$

According to Jaynes principle (Jaynes, 1957), the least biased estimate possible on the given information is given by the probability distribution $p(u)$ that maximizes the above entropy $H(u,p)$. Following the method of Lagrange, define

$$I = H(u) - \mu(E[V] - K) - \lambda(\int p(u)du - 1).$$

Using calculus of variation to maximize I with respect to the distribution $p(u)$ yields

$$\ln(p) + 1 + \mu V + \lambda = 0.$$

Therefore,

$$p(u) = \exp^{-1-\lambda-\mu V^*(x_0,u(t),t)},$$

and the entropy with maximum information is given by

$$H(u) = 1 + \lambda + \mu E[V^*(x_0, u(x), t)].$$

For optimality, a control policy u is computed that minimizes the above entropy. This is, therefore, a max-min problem.

Saridis (1988) generalizes this analysis in the presence of dynamical uncertainty. Consider that $y \in Y$ is the observation on the state x. It is essentially shown that the entropy $H(u)$ can be decomposed in three parts as

$$H(u) = H(u|y) + H(y) - H(y|u),$$

where the associated probabilities are given by

$$p(u|y) = e^{-1-\lambda-\mu W(u(y),t)} \tag{A.1}$$

$$p(y) = e^{-\rho-\nu \int_0^T ||y-x||^2 dt} \tag{A.2}$$

$$p(y|u) = p(u|y)p(y)/p(u). \tag{A.3}$$

Here, $W(u(y),t) = E_{x_0,w(t)}\{V^*(x_0, u(x,t), w, \nu, t_0\}$, and ρ, ν are appropriate constants for the entropy estimation of $H(y)$ based on Jayne's principle.

In case of parametric uncertainty, when

$$\dot{x} = f(x, u, \lambda, w, t)$$

when y are the observations

$$H(y) = H(u|y, \lambda) + H(y|\lambda) + H(\lambda) - H(y, \lambda|u).$$

An interesting observation is that entropy in a stochastic control system is decoupled into 4 different parts which can be individually computed.

BIBLIOGRAPHY

Allenby, G. M., Rossi, P. E., and McCulloch, R. E. (2005). Hierarchical bayes models: a practitioners guide. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=655541.

Araya, M., Buffet, O., and Thomas, V. (2012). Near-optimal BRL using optimistic local transitions. In Langford, J. and Pineau, J., editors, *Proceedings of the 29th International Conference on Machine Learning*, pages 97–104, New York, NY, USA.

Asmuth, J., Li, L., Littman, M. L., Nouri, A., and Wingate, D. (2009). A bayesian sampling approach to exploration in reinforcement learning. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 19–26, Montreal, Quebec, Canada.

Aurell, E., Gawędzki, K., Mejía-Monasterio, C., Mohayaee, R., and Muratore-Ginanneschi, P. (2012). Refined second law of thermodynamics for fast random processes. *Journal of statistical physics*, 147(3):487–505.

Barto, A. G. and Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13(4):341–379.

Bennett, C. H. (1987). Demons, engines and the second law. *Scientific American*, 257(5):108–116.

Bertsekas, D. (2005). *Dynamic Programming and Optimal Control*. Athena Scientific, Nashua, NH, USA, 3rd edition.

Bertsimas, D. and Perakis, G. (2006). Dynamic pricing: A learning approach. In Lawphongpanich, S., Hearn, D. W., and Smith, M. J., editors, *Mathematical and Computational Models for Congestion Charging*, volume 101 of *Applied Optimization*, pages 45–79. Springer.

Boltzmann, L. (1974). The second law of thermodynamics. In *Theoretical physics and philosophical problems*, pages 13–32. Springer.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge, UK.

Brockett, R. and Willems, J. (1979). Stochastic control and the second law of thermodynamics. In *Decision and Control including the 17th Symposium on Adaptive Processes, 1978 IEEE Conference on*, volume 17, pages 1007–1011. IEEE.

Broët, P., Richardson, S., and Radvanyi, F. (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology*, 9(4):671–683.

Cahill, N., Kemp, A. C., Horton, B. P., and Parnell, A. C. (2015). A bayesian hierarchical model for reconstructing sea levels: From raw data to rates of change. *arXiv preprint arXiv:1508.02010*.

Cao, F. and Ray, S. (2012). Bayesian hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 73–81.

Castro, P. S. and Precup, D. (2007). Using linear programming for Bayesian exploration in Markov decision processes. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2437–2442, Hyderabad, India.

Chipman, H. and McCulloch, R. E. (2000). Hierarchical priors for bayesian cart shrinkage. *Statistics and Computing*, 10(1):17–24.

Dearden, R., Friedman, N., and Russell, S. (1998). Bayesian Q-learning. In *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, pages 761–768, Madison, Wisconsin, USA.

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the maxq value function decomposition. *J. Artif. Intell. Res.(JAIR)*, 13:227–303.

Dreyfus, S. E. and Law, A. M. (1977). *The Art and Theory of Dynamic Programming.* Academic Press, New York, NY, USA, 1st edition.

Duff, M. O. (2001). Monte-carlo algorithms for the improvement of finite-state stochastic controllers: Application to bayes-adaptive markov decision processes. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, AISTATS*, Key West, FL, USA.

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

Fonteneau, R., Busoniu, L., and Munos, R. (2013). Optimistic planning for belief-augmented Markov decision processes. In *Adaptive Dynamic Programming And Reinforcement Learning (ADPRL), 2013 IEEE Symposium on*, pages 77–84.

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

Gelly, S., Kocsis, L., Schoenauer, M., Sebag, M., Silver, D., Szepesvári, C., and Teytaud, O. (2012). The grand challenge of computer go: Monte carlo tree search and extensions. *Communications of the ACM*, 55(3):106–113.

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

Glynn, P. W. and Ormoneit, D. (2002). Hoeffding's inequality for uniformly ergodic markov chains. *Statistics & Probability Letters*, 56(2):143–146.

Gopalan, A. and Mannor, S. (2015). Thompson sampling for learning parameterized markov decision processes. In *COLT*, pages 861–898.

Gray, R. M. (2011). *Entropy and information theory.* Springer Science & Business Media, Berlin/Heidelberg, Germany, 2nd edition.

Guez, A., Silver, D., and Dayan, P. (2012). Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Advances in Neural Information Processing Systems*, pages 1025–1033.

Guez, A., Silver, D., and Dayan, P. (2014). Better optimism by Bayes: adaptive planning with rich models. http://arxiv.org/abs/1402.1958.

Hauskrecht, M., Meuleau, N., Kaelbling, L. P., Dean, T., and Boutilier, C. (1998). Hierarchical solution of markov decision processes using macro-actions. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 220–229. Morgan Kaufmann Publishers Inc.

He, R., Brunskill, E., and Roy, N. (2011). Efficient planning under uncertainty with macro-actions. *Journal of Artificial Intelligence Research*, 40:523–570.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Ito, S. and Sagawa, T. (2016). Information flow and entropy production on bayesian networks. *Mathematical Foundations and Applications of Graph Entropy*, 3:2.

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

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620.

Kawaguchi, K. and Araya, M. (2013). A greedy approximation of bayesian reinforcement learning with probably optimistic transition model. arXivpreprintarXiv:1303.3163.

Kim, M., Ghate, A., and Phillips, M. (2009). A Markov decision process approach to temporal modulation of dose fractions in radiation therapy planning. *Physics in Medicine and Biology*, 54(14):4455–4476.

Kocsis, L. and Szepesvári, C. (2006). Bandit based Monte Carlo planning. In *European Conference on Machine Learning*, pages 282–293, Berlin, Heidelberg.

Kolter, J. Z. and Ng, A. Y. (2009). Near-bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 513–520. ACM.

Kotas, J. and Ghate, A. (2015). Optimal bayesian learning of dose-response parameters from a cohort. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2630392.

Kotas, J. and Ghate, A. (2016). Response-guided dosing for rheumatoid arthritis. *IIE Transactions on Healthcare Systems Engineering*, 6(1):1–21.

Krishnamurthy, V. (2016). *Partially observed Markov decision processes*. Cambridge University Press, Cambridge, United Kingdom, 1st edition.

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In *The Oxford Handbook of Computational and Mathematical Psychology*, page 279. Oxford University Press, USA.

Kumar, P. R. (1985). A survey of some results in stochastic adaptive control. *SIAM Journal on Control and Optimization*, 23(3):329–380.

Kumar, P. R. and Varaiya, P. P. (2016). *Stochastic Systems: Estimation, Identification, and Adaptive Control*. SIAM, Philadelphia, PA, USA.

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM journal of research and development*, 5(3):183–191.

Landauer, R. (1991). Information is physical. *Physics Today*, 44(5):23–29.

Lim, Z., Sun, L., and Hsu, D. (2011). Monte carlo value iteration with macro-actions. In *Advances in Neural Information Processing Systems*, pages 1287–1295.

Maass, K. and Kim, M. (2017). A markov decision process approach to optimizing cancer therapy using multiple modalities. *arXiv preprint arXiv:1706.09481*.

Maxwell, J. C. (1921). *Theory of heat*. Longmans.

Osband, I., Russo, D., and Van Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011.

Osband, I. and Van Roy, B. (2014). Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474.

Pineau, J. and Thrun, S. (2002). An integrated approach to hierarchy and abstraction for pomdps. Technical Report CMU-RI-TR-02-21, Carnegie Mellon University, Pittsburgh, PA, USA.

Poupart, P., Vlassis, N., Hoey, J., and Regan, K. (2006). An analytic solution to discrete bayesian reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 697–704, New York, NY, USA. ACM.

Powell, W. B. and Ryzhov, I. O. (2012). *Optimal Learning*. Wiley InterScience, Hoboken, New Jersey, USA.

Propp, M. B. (1985). *The thermodynamic properties of Markov processes*. PhD thesis, Massachusetts Institute of Technology.

Puterman, M. L. (1994). *Markov decision processes : Discrete stochastic dynamic programming*. John Wiley and Sons, New York, NY, USA.

Qian, P. Z. and Wu, C. J. (2008). Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics*, 50(2):192–204.

Ross, S., Chaib-draa, B., and Pineau, J. (2008). Bayes-adaptive POMDPs. In *Advances in Neural Information Processing Systems*, pages 1225–1232.

Russo, D. and Van Roy, B. (2014). Learning to optimize via information directed sampling. In *Advances in Neural Information Processing Systems*.

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

Russo, D. and Van Roy, B. (2017). Learning to optimize via information directed sampling. *Operations Research*, 66(1):230–252.

Ryzhov, I. O., Powell, W. B., and Frazier, P. I. (2012). The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195.

Salemi-Parizi, M. and Ghate, A. (2016). Lot-sizing in sequential auctions while learning demand and bid distributions. In *Proceedings of the Winter Simulation Conference*, pages 895–906, Washington, D. C., USA.

Saridis, G. N. (1988). Entropy formulation of optimal and adaptive control. *IEEE Transactions on Automatic Control*, 33(8):713–721.

Schmid, C. H. and Brown, E. N. (2000). Bayesian hierarchical models. *Methods in enzymology*, 321:305–330.

Seifert, U. (2005). Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Physical review letters*, 95(4):040602.

Seifert, U., Garrido, P. L., Marro, J., and de los Santos, F. (2011). Stochastic thermodynamics: An introduction. In *AIP Conference Proceedings*, volume 1332, pages 56–76.

Sorg, J., Singh, S., and Lewis, R. L. (2012). Variance-based rewards for approximate bayesian reinforcement learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*.

Strens, M. A. (2000). A bayesian framework for reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 943–950, San Francisco, CA, USA.

Stryhn, H. and Christensen, J. (2014). The analysis—hierarchical models: past, present and future. *Preventive veterinary medicine*, 113(3):304–312.

Sutton, R. S., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211.

Szilárd, L. (1976). On entropy reduction in a thermodynamic system by interference by intelligent subjects. *Zhurnal Physik*, 53.

Tanaka, T., Sandberg, H., and Skoglund, M. (2017). Finite state markov decision processes with transfer entropy costs. *arXiv preprint arXiv:1708.09096*.

Theocharous, G. and Kaelbling, L. P. (2003). Approximate planning in pomdps with macro-actions. In *NIPS*, pages 775–782.

Theodorou, E. A. (2015). Nonlinear stochastic control and information theoretic dualities: Connections, interdependencies and thermodynamic interpretations. *Entropy*, 17(5):3352–3375.

Theodorou, E. A. and Todorov, E. (2012). Relative entropy and free energy dualities: Connections to path integral and kl control. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pages 1466–1473. IEEE.

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

Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, 106(28):11478–11483.

Vien, N. A., Lee, S., and Chung, T. (2016). Bayes-adaptive hierarchical mdps. *Applied Intelligence*, 45(1):112–126.

Wang, T., Lizotte, D., Bowling, M., and Schuurmans, D. (2005). Bayesian sparse sampling for on-line reward optimization. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 956–963.

Wilson, A., Fern, A., Ray, S., and Tadepalli, P. (2007). Multi-task reinforcement learning: a hierarchical bayesian approach. In *Proceedings of the 24th international conference on Machine learning*, pages 1015–1022. ACM.

Wolchover, N. (2017). The quantum thermodynamics revolution. Accessed: 2017-05-05.

Yeh, W. W.-G. (1985). Reservoir management and operations models: A state-of-the-art review. *Water resources research*, 21(12):1797–1818.