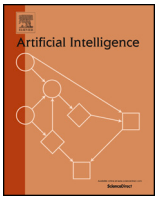
Artificial Intelligence 330 (2024) 104096

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



Finding the optimal exploration-exploitation trade-off online through Bayesian risk estimation and minimization ☆

Stewart Jamieson [a,b,c], Jonathan P. How [b], Yogesh Girdhar [c,*]

[a] MIT-WHOI Joint Program in Oceanography/Applied Ocean Science and Engineering, Cambridge and Woods Hole, 02139, MA, USA
[b] Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, 02139, MA, USA
[c] Department of Applied Ocean Physics and Engineering, Woods Hole Oceanographic Institution, 266 Woods Hole Rd, Woods Hole, 02543, MA, USA

ARTICLE INFO

Keywords:
Bayesian risk
Stochastic online learning
Multi-armed bandits
Partial monitoring

ABSTRACT

We propose *endogenous Bayesian risk minimization* (EBRM) over policy sets as an approach to online learning across a wide range of settings. Many real-world online learning problems have complexities such as action- and belief-dependent rewards, time-discounting of reward, and heterogeneous costs for actions and feedback; we find that existing online learning heuristics cannot leverage most problem-specific information, to the detriment of their performance. We introduce a belief-space Markov decision process (BMDP) model that can capture these complexities, and further apply the concepts of *aleatoric*, *epistemic*, and *process* risks to online learning. These risk functions describe the risk inherent to the learning problem, the risk due to the agent's lack of knowledge, and the relative quality of its policy, respectively. We demonstrate how computing and minimizing these risk functions guides the online learning agent towards the optimal exploration-exploitation trade-off in any stochastic online learning problem, constituting the basis of the EBRM approach. We also show how Bayes' risk, the minimization objective in stochastic online learning problems, can be decomposed into the aforementioned aleatoric, epistemic, and process risks.

In simulation experiments, EBRM algorithms achieve state-of-the-art performance across various classical online learning problems, including Gaussian and Bernoulli multi-armed bandits, best-arm identification, mixed objectives with action- and belief-dependent rewards, and dynamic pricing, a finite partial monitoring problem. To our knowledge, it is also the first computationally efficient online learning approach that can provide online bounds on an algorithm's Bayes' risk. Finally, because the EBRM approach is parameterized by a set of policy algorithms, it can be extended to incorporate new developments in online learning algorithms, and is thus well-suited as the foundation for developing real-world learning agents.

1. Introduction

There is a trend towards autonomous decision-making in increasingly unstructured and complex tasks and environments, as autonomous decision-making agents become increasingly pervasive in many societies. Fully self-driving vehicles move passengers throughout cities, algorithms help diagnose and prescribe treatments to ill patients, and autonomous robots operate in environments

☆ This article belongs to Special Issue: Risk-aware Autonomous Systems: Theory and Practice.

* Corresponding author.
E-mail addresses: sjamieson@mit.edu, sjamieson@whoi.edu (S. Jamieson), jhow@mit.edu (J.P. How), ygirdhar@whoi.edu (Y. Girdhar).

https://doi.org/10.1016/j.artint.2024.104096
Received 15 May 2022; Received in revised form 29 September 2023; Accepted 2 February 2024

Available online 21 February 2024
0004-3702/

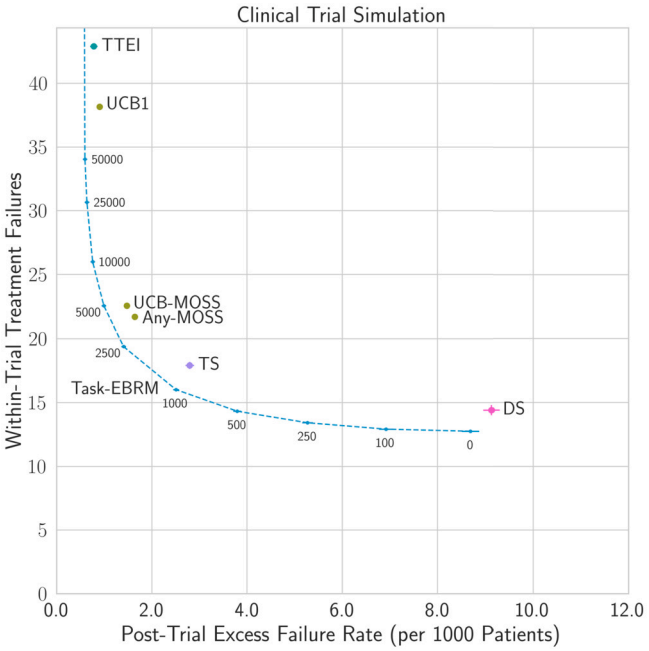


Fig. 1. Consider a simulation where algorithms sequentially choose which of 4 candidate treatments is administered to each of 423 patients in a clinical trial (see "Multi-Arm Trial Setting" in [2]). The treatment with the highest empirical success rate in the trial is administered to a global patient population. The y-axis indicates the number of failed treatments among the patients in the trial, while the x-axis shows the expected excess failure rate of the treatment administered to the global population compared to if the *true* best treatment were identified. Algorithms further left *explore* more, ensuring that the best performing treatment is confidently identified, while algorithms further down *exploit* more, finding greater success among the 423 trial patients. The EBRM algorithm (blue curve) optimizes this trade-off by taking the global patient population size (annotations) into account in its parametric reward objective. Algorithm abbreviations are defined in Subsection 6.1. (For interpretation of the colours in figures, the reader is referred to the web version of this article.)

ranging from homes and assisted-living facilities to Mars and the deep sea. While these agents attempt to make the best decisions possible despite limited information, decision-making under uncertainty always carries *risk*, and taking risks results in an accumulation of *regret* over past decisions. Maximizing long-term performance in some task is equivalent to minimizing the accumulation of this regret, and to do so autonomous agents employ algorithms for *online learning*, which describe techniques for "learning while doing".

Practitioners often use heuristics designed to optimally solve simple, archetypal online learning problems (OLPs) to instead solve all kinds of complex, real-world OLPs. These heuristics generally work well when compared to naïve strategies that do not leverage insights from online learning research. However, the complexities of real-world online learning problems cannot be accounted for by simple heuristics, despite the often important role of such complexities in determining the performance of a particular strategy. Due to their ease of use, practitioners faced with complex and novel online learning problems often choose to apply popular heuristics regardless, or to develop a new heuristic. This has led to a multitude of heuristics, many of which must be tuned for each new problem before they can achieve good performance. Furthermore, thorough analysis is required to determine whether the design of a candidate heuristic conflicts with desirable behaviour for the learning agent.

The most challenging part of designing an online learning algorithm is often deciding how it will navigate the *exploration-exploitation trade-off* [1]. Exploitation refers to an agent taking an action consistent with a plan that maximizes the agent's expected long-term task performance, where that plan is based on a model of the likelihood of possible outcomes for each action. Finding such a plan, or a close approximation, can be achieved by a variety of planning algorithms, however such a plan is often poorly suited to discover inaccuracies in the model it is optimized for. Exploration is, conversely, the act of taking one or more actions expected to reveal missing information in the model, which may thereby enable a better (exploitative) plan to be developed for later use.

Finding and applying the correct balance between exploration and exploitation in a given task is a key online learning challenge for both researchers and practitioners; the optimal balance can shift for even slight changes in task, or in how task performance is defined. For example, Fig. 1 describes a clinical trial OLP based on the work of [2]. This problem was complicated by dual objectives; the main goal of such clinical trials is to identify the best treatment among a set of candidates, but the well-being of the trial participants is also highly valued and thus it is preferable to assign as many of them to the best candidate treatments as possible [2]. Some exploration is required to identify the leading treatments, while exploration beyond this point (i.e. continuing to assign participants to less-tested, but seemingly inferior, treatments) may be beneficial for confirming the identity of the best treatment but harmful to the study participants. Each online learning heuristic balances the trade-off differently. For this particular challenge, the researchers needed to develop a new heuristic to appropriately balance between these objectives [2].

As a result of such complexities, the field has seen the development of a large number of online learning heuristics [1,10,13,15–17,20–25,27,29,42,64,65], many of which require extensive work to be tuned to a specific problem [67], and most offering

performance guarantees in only archetypal OLPs, if any. A priority for the online learning community is to distill previous insights into an approach that is easy to use, computationally efficient, provides performance guarantees *and* can model various goals and common complexities. This work is a step in that direction, in which we present *endogenous Bayesian risk minimization* (EBRM) as an approach to solve a wide range of stochastic online learning problems efficiently, with estimated risk bounds, and taking common complexities into account. In our tests, the EBRM-based algorithms demonstrate leading performance even in the well-studied archetypal OLPs that previous state-of-the-art algorithms compared against were designed for, while surpassing them to a greater degree in more complex and realistic problems.

An agent using an EBRM algorithm begins with a base "open-loop policy", which is the agent's best guess of which fixed sequence of future actions would maximize its overall task performance as measured by some reward function. The proposed Greedy-EBRM approach reasons about how much immediately useful information each available action may provide about the hidden parameters of the online learning problem. If the total risk of taking some action and then following an improved (posterior) policy is expected to be less than the risk of the base (prior) policy, a Greedy-EBRM algorithm takes that action. AsympGreedy-EBRM, the main focus of this work, is an improvement on Greedy-EBRM which ensures asymptotic convergence to an optimal policy by taking into account the long-term value of the information provided by each action. EBRM approaches produce agent policies tailored to particular OLP specifications; thus, unlike when using heuristic approaches, the agent's behaviour is always aligned with the OLP goals (i.e., the reward function) without any hyperparameter tuning. We describe the EBRM approach in more detail in Section 5.

1.1. Summary of contributions

We present a mix of theoretical analysis and practical algorithms that can serve as the foundation for further work in applying online learning principles and insights towards being able to efficiently solve complex real-world online learning problems. In particular, our contributions are applicable to a wide range of stochastic online learning problems, including problems augmented with belief-dependent rewards, time-discounted rewards, and action feasibility criteria. Specifically, we contribute:

- The decomposition of Bayes' risk into aleatoric, epistemic and *process* risks, which help bound the expected regret and provide insights into the optimal exploration-exploitation trade-off.
- The AsympGreedy-EBRM approach to online learning, which enables risk-bounded, high performance online learning in complex OLPs while remaining computationally tractable, and with guaranteed asymptotic convergence in problems with identifiable hidden parameters.
- Empirical results demonstrating the superior performance of EBRM algorithms against state-of-the-art baselines in OLPs representative of real-world problems, including:
  - multi-armed bandit optimization,
  - best-arm identification,
  - dynamic pricing, and
  - a problem with mixed objectives, specifically rewards derived from a combination of the agent's actions *and* its posterior beliefs of the unknown problem parameters.
- Methods to efficiently compute, across a wide range of stochastic online learning problems, online regret bounds and the expected risk and value-of-information for various actions.

2. Related works

The field of online learning has been strongly motivated by practical applications since the seminal work of Thompson in the 1930s [3]. The original motivating idea of exploring the efficacy of a discrete set of medical treatments while simultaneously exploiting the leading treatments to save lives continues to be a research interest [2], and is an exemplar of the ubiquitous stochastic multi-armed bandit (MAB) problem [1]. In this terminology, the "arms" of the bandit represent, for example, different treatments. In stochastic multi-armed bandit problems, an observation is generated each round from an unknown distribution specific to the arm played that round, and the reward is the sum of the observations. Multi-armed bandits are a good structure with which to approximate many real-world problems, and there are a variety of successful online learning algorithms designed for them. Variations on this structure include infinite-time and time-discounted bandits, as well as best-arm identification ("pure exploration") [1].

Partial monitoring is a generalization of multi-armed bandits that allows for more general relationships between observations and rewards [4]. Partial monitoring problems are characterized by different levels of observability, which bound how well any algorithm can perform [5,6]. For example, the dynamic pricing problem models how adjusting a product's price changes profits when different customers are willing to pay different prices [7]. The lack of local observability in dynamic pricing makes some instances of this problem fundamentally harder than bandit problems [8].

Unfortunately, most real-world problems do not fit perfectly into common OLP structures or the assumptions of common online learning algorithms. As discussed in Section 1, the objective of identifying the best treatment from a range of options is well modelled as best-arm identification, but this formulation neglects the objective of exploiting effective treatments for the sample population [9]. Similarly, algorithms which had the most patient recoveries among a sample population tended to have insufficient statistical power to achieve the clinical trial's goal of supporting the superiority of a specific treatment [2] (see [10] for more general discussion on this issue). Precisely controlling the balance between competing objectives is not a feature of previous online learning heuristics.

2.1. Online learning algorithms and design principles

Decades of research into online learning problems have generated a variety of widely adopted algorithm design principles. For example, "optimism in the face of uncertainty" [11,12] is the driving principle behind the popular and well-studied family of Upper Confidence Bound algorithms, which achieve optimal asymptotic performance bounds in many OLPs [13–18]. Other design principles, such as "maximize the expected improvement" [19,20] and "follow the knowledge gradient" [21,22] have similarly led to the development of online learning algorithms with strong guarantees and empirical results. In particular, many of these algorithms are "no-regret" for some OLPs, in that the regret of the algorithm grows sublinearly over time so the average regret decays asymptotically to zero. Information-directed sampling (IDS) is perhaps the best of the state-of-the-art strategies for long-/infinite-horizon multi-armed bandit and partial monitoring problems, and introduces novel information-theoretic techniques with strong finite-time regret bounds [23–25]. While the AsympGreedy-EBRM algorithm presented in this work is more flexible and generally outperforms IDS even on bandit problems, there may be an opportunity for future work to develop even more effective "information-directed" EBRM algorithms.

The most similar online learning algorithm to the proposed EBRM strategy is likely the Knowledge Gradient (KG) algorithm, which greedily chooses informative actions in order to increase the expected performance of a simple "stop-learning" policy [21]. The Greedy-EBRM algorithm presented in this work can be viewed as a generalization of the KG approach to incorporate additional problem complexities and constraints. We extend KG principles beyond multi-armed bandits, and produce similar results to that of knowledge-gradient optimality for monotone submodular value-of-information functions [26], but for broader classes of OLPs.

Strategies for best-arm identification include Top-Two Thompson Sampling [27] and the original KG exploration algorithm [28, 29]. The Top-Two Expected Improvement algorithm [20] is a more recent approach that builds upon the ideas in these strategies and provides improved performance and regret bounds.

2.2. Markov decision processes and partial observability

Markov Decision Processes (MDPs) [30] are a highly flexible framework that can model a much broader range of decision problems than online learning problems. In particular, *partially observable* Markov decision processes (POMDPs) [31] can describe any problem in which there is some hidden *state* (set of parameters), which may change, and the consequences of actions taken by an agent depend on the value of that state. The agent's goal is to collect *reward* (a performance metric), but the amount of reward collected typically depends on the state; like in OLPs, the agent must generally trade-off between taking actions that help to reveal the hidden state (explore), or actions that collect reward (exploit) [32,33]. The solution of a (PO)MDP is a *policy* (strategy for choosing actions), often found with *reinforcement learning* [34,35].

The "Markovian" property of POMDPs requires that the effects of an action depend only on the current state, but the model is otherwise able to capture many kinds of complexities, such as competing reward objectives, action costs or constraints, and complex state transition dynamics. POMDP models have been used to develop solutions for problems as diverse as railroad maintenance planning [36], unmanned aerial vehicle contingency management [37], recommendation systems [38], and treatment planning for sepsis patients [39]. However, many practitioners instead use online learning heuristic algorithms that ignore these complexities. A major factor in this decision is that specifying the components of a POMDP and computing its optimal solution tends to be tedious and computationally expensive, if not entirely infeasible [1].

Bayes-Adaptive MDPs (BAMDPs) [40] are a subclass of POMDPs which describe problems in which a fully observable state is separate from a stationary (fixed) hidden state. In this work, we consider all problems which can be represented by BAMDPs, with the additional constraints that the state transition following an action is deterministic given only the *observable* state, while the (possibly stochastic) reward of an action depends only on the *hidden* state. We formulate stochastic online learning as a fully-observable belief-space MDP (BMDP) [31], where "belief states" represent the BAMDP observable state combined with a probability distribution over possible hidden states. As the hidden state is stationary, an agent navigating a BMDP begins with a belief state which broadly distributes probability across many possible values of the hidden parameters, but moves towards one which concentrates probability on the most likely values.

2.3. Risk quantification, decomposition, and minimization

The *risk* of a policy describes the degree to which the performance of that policy may be exceeded by some other policy; even if a policy may produce good results *in expectation*, it may still carry substantial risk. Most autonomous agents use risk-neutral decision making, which only considers the expected performance of a policy. However, this can lead to undesirable behaviours, particularly in social or safety-critical contexts. Risk quantification is essential to developing risk-*aware* and risk-*averse* (or risk-sensitive) agents [41–44]. A risk-aware agent is capable of, for example, reporting to the user when it is in a high-risk state (i.e., a situation for which there is no low risk policy) [45]. A risk-averse agent goes further by actively avoiding high risk states even if they would, in expectation, lead to better outcomes [43,44].

It is often useful to separate risk into *aleatoric* and *epistemic* components [46–48]. The aleatoric component describes the degree to which an algorithms's results depend on random processes in the environment, while the epistemic component describes how much risk could be eliminated through better knowledge of the environment [48,49].[1] Actions which reveal information on the hidden

[1] Intuitively, aleatoric risk is due to the inherent unpredictability of "dice rolls", while epistemic risk is due to uncertainty in whether the dice is loaded (and how).

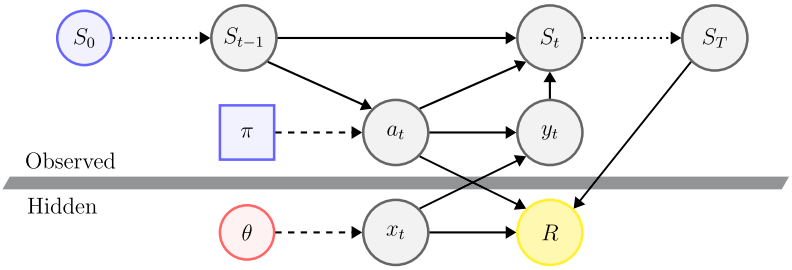


Fig. 2. The online learning problem formulation. The Environment (red) picks a set of hidden parameters θ. The User (blue) specifies the learning agent's initial state S_0 and a policy function π that defines the distribution from which to draw each action, such that $a_t \sim \pi(S_{t-1})$. The goal is to find a policy that maximizes the amount of reward (yellow) collected, R, which depends on the action-outcome pairs (x_t, a_t) for $t = 1, \ldots, T$, and on the agent's final state S_T. Dashed arrows represent random sampling from a probability distribution, and dotted arrows represent omitted parts of the graph. All variables and their relationships are defined in Table 1.

Table 1
Components of a Stochastic OLP. Vec: Vector. Dist: Distribution. Fn: Function. Comp: Compound.

OLP Parameter	Observability	Type	Symbol	Determined by	Space
Hidden Parameters	Inferred	Vec.	θ	Environment	Θ
Hidden Outcomes	Inferred	Vec.	x	$x_t \sim g_\theta$	$\mathcal{X}$
Actions	Observed	Vec.	a	$a_t \sim \pi\left(S_{t-1}\right)$	$\mathcal{A}$
Observations	Observed	Vec.	y	$y_t = \Phi\left(x_t, a_t\right)$	$\mathcal{Y}$
Auxiliary State	Observed	Vec.	ξ	$\xi_t = \delta\left(\xi_{t-1}, a_t\right)$	Ξ
Belief Distribution	Observed	Dist.	b	Eq. (3.1)	$\mathcal{P}(\Theta)$
Belief State	Observed	Comp.	S	$S_t = \{b_t, \xi_t\}$	$S = \mathcal{P}(\Theta) \times \Xi$
Policy	Known	Fn.	π	User	$S \to \mathcal{P}(\mathcal{A})$
State Update Fn.	Known	Fn.	δ	OLP	$\Xi \times \mathcal{A} \to \Xi$
Action Reward Fn.	Known	Fn.	$R(x, a)$	OLP	$\mathcal{X} \times \mathcal{A} \to \mathbb{R}$
Belief Reward Fn.	Known	Fn.	$R(b)$	OLP	$\mathcal{P}(\Theta) \to \mathbb{R}$
Observation Fn.	Known	Fn.	Φ	OLP	$\mathcal{X} \times \mathcal{A} \to \mathcal{Y}$
Process Model	Known	Fn.	g_θ	OLP	$\Theta \to \mathcal{P}(\mathcal{X})$
Feasibility Criterion	Known	Fn.	Ω	OLP	$\Xi \to \mathfrak{P}(\mathcal{A})$
Discount Factor	Known	Scalar	γ	OLP	$\gamma \in (0, 1]$

state of the POMDP reduce epistemic risk. Conversely, aleatoric risk cannot be avoided in OLPs, as it is fixed for a given hidden state.

In this work, we consider the Bayesian setting where the hidden state is modelled as being drawn from a "prior" distribution; the *Bayes' risk* of a policy is the expected risk for a random hidden state distributed according to this prior. This setting is explored in Bayesian reinforcement learning [34,50,51], and techniques have been developed to learn risk-aware and risk-averse policies for general BAMDPs [33,52]. However, such approaches require training a policy by running many episodic simulations of the problem. We propose EBRM as a risk-aware solution which, like the online learning heuristics discussed in Subsection 2.1, does not require any training, thus making it far more convenient to use.

3. Online learning as a belief-space Markov decision process

We begin by formulating stochastic online learning problems, depicted in Fig. 2, as belief-space Markov decision processes with particular structure. We will then discuss how the performance of a policy is measured in an OLP, and how this relates to defining the risk of that policy.

Notation We index variables related to sequential decisions (BMDP states) using the time $t \in \mathbb{N}_{>0}$. We define $[K] := \{1, \ldots, K\}$ for $K \in \mathbb{N}_{>0}$. We denote the space of probability measures over a Borel measurable set $\mathcal{X}$ as $\mathcal{P}(\mathcal{X})$, and the power set of a set $\mathcal{X}$ as $\mathfrak{P}(\mathcal{X})$. The indicator function $\mathbb{1}$ is defined as $\mathbb{1}(\text{True}) = 1$ and $\mathbb{1}(\text{False}) = 0$.

3.1. Bayesian BMDP formulation of stochastic OLPs

Let $X = \{x_t \in \mathcal{X}\}_{t \in \mathbb{N}_{>0}}$ denote the hidden *outcomes* of the OLP. When an agent performs action $a_t \in \mathcal{A}$ at time t, it produces an observation $y_t = \Phi\left(x_t, a_t\right)$ and a reward $R\left(x_t, a_t\right)$, according to known functions Φ and R. A *stochastic* OLP is one in which each hidden outcome is assumed to be independently drawn from $g_\theta \in G_\Theta$, where G_Θ is a family of probability distributions over $\mathcal{X}$ parameterized by $\theta \in \Theta$; thus, the hidden outcomes are i.i.d. given θ.

The agent's actions are chosen by a policy π that takes into account the agent's current *belief* of the hidden parameters $b_t \in \mathcal{P}(\Theta)$ and an observable *auxiliary state* $\xi_t \in \Xi$. The belief describes the likelihood of hidden parameter values and is used to infer the

expected reward of each action. The auxiliary state indicates the time and resources available to the agent and determines which actions are *feasible*. Together, these compose the agent's *belief state*, $S_t = \{b_t, \xi_t\}$. The policy generates actions according to $a_{t+1} \sim \pi(S_t)$.

The auxiliary state ξ_t contains all information relevant to the agent's decision making, other than the hidden parameters. For example, if actions have time or resource costs, it indicates the agent's remaining time or spending budget. The auxiliary state changes according to $\xi_t = \delta(\xi_{t-1}, a_t)$ for some known and deterministic function δ. The feasibility criterion $\Omega(\xi_t)$ indicates which actions are available in each state, so a "feasible" policy π must satisfy $\Pr\left(a_{t+1} \in \Omega(\xi_t)\right) = 1$ for $a_{t+1} \sim \pi(S_t)$. The agent ceases from taking further actions upon reaching a *terminal state*, indicated by $\Omega(\xi_t) = \varnothing$.

The belief distribution b_t characterizes the posterior likelihood of the hidden parameters θ based on past action-observation pairs $\left\{\left(a_1, y_1\right), \ldots, \left(a_t, y_t\right)\right\}$. For convenience, we denote arbitrary belief distributions as b or b'. The initial belief distribution $b_0(\theta)$ is specified by the user as a prior over the hidden parameters. The observation function Φ and process model g_θ, together, implicitly define the likelihood $\Pr\left(y_t \mid \theta, a_t\right)$, and so b_t is given by Bayes' law,

$$b_t(\theta) := \Pr\left(\theta \mid a_{1:t}, y_{1:t}\right) = \frac{\Pr\left(y_t \mid \theta, a_t\right) b_{t-1}(\theta)}{\Pr\left(y_t \mid a_t\right)}. \tag{3.1}$$

We are often interested in *counterfactual* belief states, which may arise when considering a possible next outcome-action pair (x, a). Such a belief state is denoted $S_{t+1}^{x,a} = \{b_{t+1}^{x,a}, \xi_{t+1}^{a}\}$, with

$$b_{t+1}^{x,a}(\theta) = \frac{\Pr\left(\Phi(x,a) \mid \theta, a\right) b_t(\theta)}{\Pr\left(\Phi(x,a) \mid a\right)}, \quad \xi_{t+1}^{a} = \delta(\xi_t, a). \tag{3.2}$$

The components discussed thus far specify the BMDP context; the remaining part of the specification is the BMDP goal. A variety of useful goals can be expressed through action-based rewards accumulated as actions are taken and an information-based reward based on the terminal belief state. Accordingly, we define the *action reward* $R(x,a) \in \mathbb{R}$ and *belief reward* $R(b) \in \mathbb{R}$, such that the goal is to maximize the sum,

$$\gamma^T R(b_T) + \sum_{\tau=1}^{T} \gamma^\tau R(x_\tau, a_\tau), \tag{3.3}$$

evaluated upon reaching a terminal state $S_T = \{b_T, \xi_T\}$ where $\Omega(\xi_T) = \varnothing$. The discount factor $\gamma \in (0, 1]$ specifies the degree to which earlier rewards are preferred to later rewards. The BMDP specification requires a *Markovian* reward model ρ, which depends only on the current state and action; it suffices to define,

$$\rho\left(S_t, a\right) := \mathbb{E}_{x|b_t}\left[R(x,a) + R(b_{t+1}^{x,a}) \mathbb{1}\left(\Omega(\xi_{t+1}^{a}) = \varnothing\right)\right], \tag{3.4}$$

where $b_{t+1}^{x,a}$ and ξ_{t+1}^{a} are defined as in Eq. (3.2). Together, the reward model ρ and discount factor γ formally specify the BMDP goal.

3.2. Optimal online learning

Every policy π has an associated *value function*, $V^\pi(S)$, defined recursively by the Bellman equation [30],

$$V^\pi(S_t) := \mathbb{E}_{a\sim\pi(S_t)}\left[\rho\left(S_t, a\right) + \mathbb{E}_{x|b_t}\left[\gamma V^\pi(S_{t+1}^{x,a})\right]\right], \tag{3.5}$$

$$\Pr\left(x \mid b_t\right) = \int_\Theta g_\theta(x) b_t(\theta) d\theta. \tag{3.6}$$

By construction, $V^\pi(S)$ is equal to the terminal reward expected to be received by following the policy π until reaching a terminal state, as computed in Eq. (3.3).

The calculation in Eq. (3.5) assumes the distribution of each observation follows from the conditional density $\Pr(x \mid b)$ generating the hidden outcome x based on the previous belief state. We also consider *conditional* value functions $V^\pi(S_t; \theta)$ and $V^\pi(S_t; X)$, which represent the reward achieved by the policy π for a specific θ or sequence of hidden states X respectively,

$$V^\pi(S_t; \theta) := \mathbb{E}_{a\sim\pi(S_t)}\left[\mathbb{E}_{x|\theta}\left[R(x,a) + R(b_{t+1}^{x,a}) \mathbb{1}\left(\Omega(\xi_{t+1}^{a}) = \varnothing\right) + \gamma V^\pi(S_{t+1}^{x,a}; \theta)\right]\right], \tag{3.7}$$

$$V^\pi(S_t; X) := \mathbb{E}_{a\sim\pi(S_t)}\left[R(x_{t+1}, a) + R(b_{t+1}^{x_{t+1},a}) \mathbb{1}\left(\Omega(\xi_{t+1}^{a}) = \varnothing\right) + \gamma V^\pi(S_{t+1}^{x_{t+1},a}; X)\right]. \tag{3.8}$$

From these definitions, it follows that $V^\pi(S_t) \equiv \mathbb{E}_{\theta|b_t}\left[V^\pi(S_t; \theta)\right] \equiv \mathbb{E}_{X|b_t}\left[V^\pi(S_t; X)\right]$.

3.2.1. Types of policies

A *deterministic* policy π is one that always picks the same action for a given belief state; that is, if $a \sim \pi(S)$ then $\exists a' : \Pr(a = a') = 1$. Any policy which is not deterministic is called a *stochastic* policy. An *open-loop* policy is one for which the action distribution is a function of only the auxiliary state ξ_t; that is, one for which $\pi(S) = \pi(S')$ holds $\forall S, S'$ such that $\xi = \xi'$. Otherwise, the policy is a

Table 2
Risk functions; the risk of policy π_1 is defined with respect to policy π_2.

Name	Definition
Instance Risk	$r\left(\pi_1 \| \pi_2; S, X\right) := V^{\pi_2}(S;X) - V^{\pi_1}(S;X)$
Expected Risk	$\overline{r}\left(\pi_1 \| \pi_2; S, \theta\right) := V^{\pi_2}(S;\theta) - V^{\pi_1}(S;\theta)$
Bayesian Risk	$\overline{\overline{r}}\left(\pi_1 \| \pi_2; S\right) := V^{\pi_2}(S) - V^{\pi_1}(S)$

closed-loop policy. A policy which is both deterministic *and* open-loop can be completely characterized, for a given initial belief state, by the fixed sequence of actions that it would take from that state. We denote a deterministic open-loop policy and its corresponding action sequence as $\hat{\pi}$ and $A^{\hat{\pi}}$, respectively.

3.2.2. Policy optimality

A b-optimal policy $\pi_b^\star$ is a solution to, where Π is the set of all policies,

$$\pi_b^\star \in \arg\max_{\pi \in \Pi} V^\pi(S_0). \tag{3.9}$$

Such a policy achieves the most terminal reward possible, in expectation, from the initial state S_0, assuming that $\theta \sim b_0$. Similarly, a θ-optimal policy $\pi_\theta^\star$ and X-optimal policy $\pi_X^\star$ satisfy, from the initial state S_0,

$$\pi_\theta^\star \in \arg\max_{\pi \in \Pi} V^\pi(S_0;\theta), \tag{3.10}$$

$$\pi_X^\star \in \arg\max_{\pi \in \Pi} V^\pi(S_0;X). \tag{3.11}$$

As we assume the hidden outcomes are independently and identically distributed, it suffices to search the set of *deterministic* policies in order to find a b-, θ-, or X-optimal policy for an OLP.[2]

Most optimal policies are closed-loop, even if they are deterministic. However, deterministic open-loop policies are useful to consider as they are straightforward to describe and analyze. In particular, later sections will often refer to "optimal deterministic open-loop policies", defined below.

Definition 1. An **optimal deterministic open-loop (ODOL) policy** $\hat{\pi}_{S_t}^\star$ is characterized by, where $\hat{\Pi} \subset \Pi$ denotes the set of deterministic open-loop policies,

$$\hat{\pi}_{S_t}^\star \in \arg\max_{\hat{\pi} \in \hat{\Pi}} V^{\hat{\pi}}(S_t). \tag{3.12}$$

3.3. Measuring policy risk

A *risk function* measures how much less reward is expected to be attained by one policy π_1 compared to another policy π_2. We define three fundamental risk functions in Table 2.

Risk is often defined relative to an optimal policy so that it represents a shortfall relative to the maximum achievable reward.[3] We define some Bayesian risks relative to the X- and θ-optimal policy "families",

$$\overline{\overline{r}}(\pi \| \pi_X^\star; S_t) := \mathbb{E}_{X|b_t}\left[r(\pi \| \pi_X^\star; S_t, X)\right], \tag{3.13}$$

$$\overline{\overline{r}}(\pi \| \pi_\theta^\star; S_t) := \mathbb{E}_{\theta|b_t}\left[\overline{r}(\pi \| \pi_\theta^\star; S_t, \theta)\right]. \tag{3.14}$$

It is important to note the abuse of notation here, where the terms $\pi_\theta^\star$ and $\pi_X^\star$ on the left side of each equation are not specific policies, but rather represent conceptual policies optimal for *any* particular realization of θ and X, respectively. We can think of $\pi_\theta^\star$ and $\pi_X^\star$ in such contexts as "cheating" policies, which suggest actions based on information unavailable in the belief state.

We demonstrate in Section 4 that insights can be made by considering the risk of a policy with respect to the various optimal policies. Lemma 1 shows how a "hierarchy" of risk arises from comparing a policy π against these different optimal policies.

Lemma 1. *For any policy $\pi \in \Pi$ and any belief state $S_t \in \mathcal{S}$, we have that*

$$\overline{\overline{r}}(\pi \| \pi_X^\star; S_t) \geq \overline{\overline{r}}(\pi \| \pi_\theta^\star; S_t) \geq \overline{\overline{r}}(\pi \| \pi_b^\star; S_t) \geq 0.$$

[2] Prior works have explored optimal online learning policies when the hidden outcomes are generated by an *adversarial* process, for which stochastic policies can greatly outperform deterministic ones.

[3] In this case it also represents the (expected) future *regret* of a policy; regret is the evaluation metric for most online learning problems. Appendix A presents further discussion on the relationship between risk and regret.

Proof. Expand the risk functions in terms of the (conditional) value functions, express the values of optimal policies as maximization problems based on Eqs. (3.9)–(3.11), and apply Jensen's inequality. □

3.4. Exogenous and endogenous metrics

We distinguish between *exogenous* and *endogenous* measures of a learning agent's performance, which characterize whether a metric is computed from only the belief state S_t, or from external information.

An *exogenous* metric is a function that requires inputs which are not part of S_t. This makes it a means of external evaluation of the performance of a policy, as it requires knowing values of variables not provided to the learning agent. This may include the true values of X and θ, which could be known under experimental conditions by the experimentalist or in other settings by an "oracle". Such a metric can be an effective tool for evaluating different policies, but a real learning agent is unable to compute it online.

Conversely, an *endogenous* metric is any function of only S_t; as such, when an agent estimates the risk of its own policy we refer to it as an endogenous risk estimate. These are endogenous in that they can be computed using only S_t, the information available to the agent at time t. The algorithms contributed by this work are based on minimizing endogenous Bayesian risk *online*.

4. Online computation of endogenous risks

Quantifying the risk of policies can be useful for many tasks, such as identifying the best policy from some set or providing safety and performance guarantees. This section will explain how endogenous risk measures provide insights into the nature of an OLP and into the behaviour of a specific policy, and will discuss the feasibility of computing the risk and value of various classes of policies. Note that all measures of risk in this section will be endogenous, in that X and θ are taken to be random variables distributed according to $\Pr\left(X, \theta \mid b_t\right)$.

4.1. Endogenous Bayesian risk functions

The risk of a policy, as defined in Subsection 3.3, is always given relative to some "reference" policy; the choice of this reference policy offsets the risk estimate, as $\forall \pi_1, \pi_2, \pi_3$,

$$\overline{\overline{r}}(\pi_1 \| \pi_3; S) = \overline{\overline{r}}(\pi_1 \| \pi_2; S) + \overline{\overline{r}}(\pi_2 \| \pi_3; S). \tag{4.1}$$

By choosing reference policies that are "optimal", the risk computed can give insight into the OLP and into policies of interest. The first, and perhaps most important, risk function is the *total Bayes' risk*.

Definition 2. The **total Bayes' risk** of a policy π from some belief state S is

$$\text{TotalRisk}(\pi; S) := \overline{\overline{r}}(\pi \| \pi^\star_X; S).$$

We refer to this quantity as the "total risk" because $\pi^\star_X$ is the reference policy with the highest possible value, so the total Bayes' risk is the largest possible risk of π. The total risk has an intuitive interpretation: it indicates how much better an agent following some policy π could perform if it instead made perfect use of complete information regarding all of the hidden outcomes.

Total risk is a well-studied metric in online learning but is often challenging to compute, and considering it in isolation conceals insights into the OLP structure, the belief state, and the policy. The remainder of this subsection will explore a useful decomposition of the total risk into three distinct and edifying components: the *aleatoric* and *epistemic* risks of a belief state and the *process* risk of the policy.

4.1.1. Aleatoric risk

Due to the stochastic nature of the hidden outcomes, even complete information about the hidden parameters is generally not enough for the agent to be able to achieve the maximum possible reward. This limitation is inherent to the OLP itself, and is captured by the *aleatoric* risk. The aleatoric risk thus provides insight into the difficulty of an OLP; if it is by far the largest component of the Bayes' risk of a policy π, then most of the risk of the policy is due to random chance and cannot be eliminated. If this risk is unacceptably high, it may indicate that the OLP cannot be satisfactorily solved.

Definition 3. The **aleatoric risk** of an OLP with parameters θ is

$$\text{AleatoricRisk}(\theta) := \overline{r}(\pi^\star_\theta \| \pi^\star_X; \theta).$$

The **aleatoric Bayes' risk** of a belief state S is accordingly defined as

$$\text{AleatoricBayesRisk}(S) := \overline{\overline{r}}(\pi^\star_\theta \| \pi^\star_X; S) = \mathbb{E}_{\theta|b}\left[\text{AleatoricRisk}(\theta)\right].$$

For any parameters θ, the aleatoric risk is non-negative (see Lemma 1), showing that even with perfect knowledge of the parameters θ, the optimal policy $\pi^\star_\theta$ may still incur regret. The aleatoric *Bayes'* risk of S_t measures how much aleatoric risk is expected to

Fig. 3. Refer to Problem 1 and Example 1. Left: The aleatoric risk over $\theta \in [0,1]$ assuming $n = 10$ rounds, shown with three belief distribution densities. The aleatoric risk is highest for fair coins, for which each flip result is unpredictable, and lowest for a trick coin that always lands on one side. Right: The aleatoric Bayes' risk of belief distributions $b_t = \text{Beta}(\alpha, \beta)$ for various $\alpha, \beta \in \mathbb{R}_{>0}$. Observe that aleatoric Bayes' risk increases along the line $\alpha = \beta$ as $\alpha, \beta \to \infty$.

remain even after being given complete information on θ, assuming the parameters θ were sampled from b_t. Below, we provide an example of computing aleatoric risk and aleatoric Bayes' risk in a toy problem.

Problem 1. Suppose a coin flips heads with probability $\theta \in [0,1]$ and produces hidden outcomes in $\mathcal{X} = \{\text{Heads}, \text{Tails}\}$. An agent has two actions $\mathcal{A} = \{1,2\}$, and the action reward function is:

$$R(x,1) = \begin{cases} 1, & x = \text{Heads}, \\ 0, & x = \text{Tails}. \end{cases} \qquad R(x,2) = \begin{cases} 0, & x = \text{Heads}, \\ 1, & x = \text{Tails}. \end{cases}$$

There is no belief reward or discounting, so $R(b) = 0$ and $\gamma = 1$. The agent plays for n rounds, so $\Omega(\xi_t) = \mathcal{A}$ for $t < n$ and $\Omega(\xi_n) = \varnothing$. The initial belief distribution is $b_0 = \text{Beta}(1,1)$.

Example 1. In Problem 1, the unique X-optimal policy is to take action 1 when $x_t = \text{Heads}$ and action 2 when $x_t = \text{Tails}$. An θ-optimal policy is to always take action 1 if $\theta > 0.5$, and to take action 2 otherwise. It is straightforward to compute that, over n rounds, $V^{\pi^\star_X}(S_0;\theta) = n$ and $V^{\pi^\star_\theta}(S_0;\theta) = n\max\{(1-\theta),\theta\}$. Thus $\text{AleatoricRisk}(\theta) = n - \max\{n(1-\theta), n\theta\}$ and

$$\begin{aligned} \text{AleatoricBayesRisk}(b_0) &= \mathbb{E}_{\theta|b_0}\left[n - n\max\{(1-\theta),\theta\}\right], \\ &= \int_0^{0.5} (n - n(1-\theta))d\theta + \int_{0.5}^{1} (n - n\theta)d\theta = \frac{n}{4}. \end{aligned}$$

The aleatoric risk and aleatoric Bayes' risk for various θ and b, respectively, are shown in Fig. 3.

4.1.2. Epistemic risk

As a learning agent gradually learns more about the hidden parameters θ, it is typically able to improve its performance. While it can never eliminate the aleatoric risk inherent to the OLP, this learning process reduces the *epistemic* (knowledge-based) risk of the belief state, defined as follows.[4]

Definition 4. The **epistemic risk** of a belief state S_t is

$$\text{EpistemicRisk}(S_t) := \bar{\bar{r}}(\hat{\pi}^\star_{S_t} \| \pi^\star_\theta; S_t).$$

By Lemma 1, the epistemic risk is always non-negative. Importantly, however, the epistemic risk of an agent *can be reduced* through performing actions that lead to belief states with better estimates of the hidden parameters. The reduction of epistemic risk to 0 depends on the hidden parameters being *identifiable*.

Definition 5. Let $\xrightarrow[t\to\infty]{P}$ denote convergence in probability. The hidden parameters θ are **identifiable** from b_0 if and only if there exists an infinite sequence of actions $A \in \mathcal{A}^\infty$ such that $\forall \eta \in \Theta : |\eta - \theta| > 0$,

$$\mathbb{E}_{X|\theta}\left[b_t^{X,A}(\eta)\right] \xrightarrow[t\to\infty]{P} 0, \tag{4.2}$$

[4] In reinforcement learning, some works (e.g. [47,53,54]) refer to the epistemic risk of a learned *policy*, which results from a lack of training data in the vicinity of the S_t. In our notation, this risk would be denoted as $\bar{\bar{r}}(\pi \| \pi^\star_{S_t}; S_t)$.

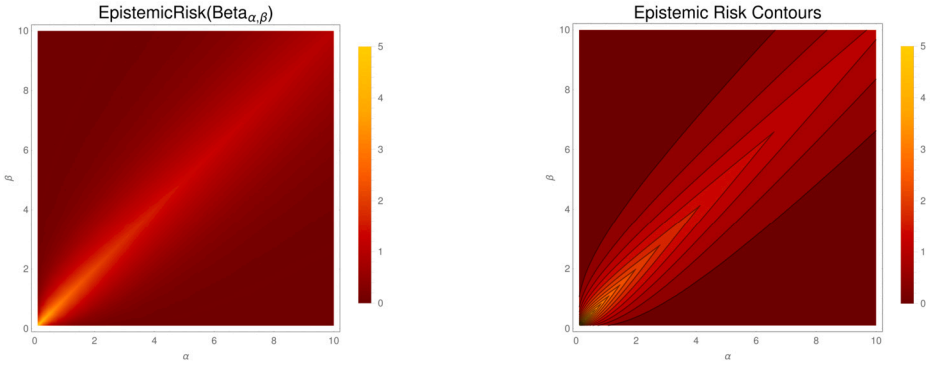


Fig. 4. Refer to Problem 1 and Example 2. Left: The epistemic risk for various initial belief distributions, assuming $n = 10$ rounds. It is largest near the origin, for which the belief distribution is highly uncertain about which action is optimal. Right: The contours highlight that identifying the best action is most difficult when $\alpha \approx \beta$; however, in contrast to the aleatoric Bayes' risk (see Fig. 3), the epistemic risk is asymptotically decreasing as $\alpha, \beta \to \infty$.

where $b_t^{X,A}(\eta)$ is defined recursively by Bayes's law as, with $y_t = \Phi(X(t), A(t))$ and $b_0^{X,A} = b_0$,

$$b_t^{X,A}(\eta) = \frac{\Pr\left(y_t \mid \theta, A(t)\right) b_{t-1}^{X,A}(\eta)}{\Pr\left(y_t \mid A(t)\right)}. \tag{4.3}$$

Theorem 1. *Suppose the hidden parameters θ are identifiable over some infinite sequence of actions. Then, over this sequence of actions,*

$$\text{EpistemicRisk}(S_t) \xrightarrow[t\to\infty]{P} 0. \tag{4.4}$$

Proof. See Appendix D. □

Theorem 1 implies that, if the belief distribution converges upon the true parameters θ, then the difference in value between an ODOL policy $\hat{\pi}^\star_{S_t}$ and any θ-optimal policy converges to 0. This reflects that $\hat{\pi}^\star_{S_t}$ generally improves over time as actions are performed and the belief state converges. Estimates of the changes in epistemic risk that would result from various actions are useful as they can guide the learning agent towards states where the ODOL policy $\hat{\pi}^\star_{S_t}$ is near-optimal; this strategy will be discussed further in Section 5. Example 2 demonstrates how to compute epistemic risk in a simple OLP, while Fig. 4 depicts how the epistemic risk in Problem 1 varies with the belief distribution.

Example 2. In Problem 1, an optimal deterministic open-loop-optimal policy $\hat{\pi}^\star_{S_0}$ is to choose action 1 if $\mathbb{E}_{\theta|b_0}[\theta] \geq 0.5$, and action 2 otherwise. For a prior belief distribution $b_0 = \text{Beta}(\alpha, \beta)$ with $\alpha, \beta \in \mathbb{R}_{>0}$, then

$$\mathbb{E}_{x|b_0}\left[R(x, \hat{\pi}^\star_{S_0}(S_t))\right] = \frac{\max\{\alpha,\beta\}}{\alpha+\beta} \implies V^{\hat{\pi}^\star_{S_0}}(S_0) = \frac{n\max\{\alpha,\beta\}}{\alpha+\beta}$$

Thus,

$$\text{EpistemicRisk}(S_0) = \mathbb{E}_{\theta\sim\text{Beta}(\alpha,\beta)}[n\max\{1-\theta,\theta\}] - \frac{n\max\{\alpha,\beta\}}{\alpha+\beta}$$

The epistemic risk for various belief distribution parameters α and β is presented in Fig. 4.

4.1.3. Process risk

Lastly, the *process* risk of any policy π is defined as the excess risk of π over an ODOL policy $\hat{\pi}^\star_{S_t}$. Accordingly, the process risk of any *open-loop* policy is non-negative.

Definition 6. The **process risk** of a policy π in some belief state S_t is

$$\text{ProcessRisk}(\pi; S_t) := \overline{\overline{r}}(\pi \| \hat{\pi}^\star_{S_t}; S_t).$$

Unlike aleatoric and epistemic risk, the process risk can be negative; specifically, the process risk is negative for any closed-loop policy that is able to, in expectation, leverage new observations to make better decisions than $\hat{\pi}^\star_{S_t}$. Suppose that the agent is following some policy which cycles through various actions, gradually identifying the hidden parameters. Proposition 1 shows that if the parameters are identifiable, the process risk of any other policy the agent might consider using becomes non-negative. This represents improvement in the ODOL policies; perhaps unsurprisingly, it indicates that once the agent has complete information

Table 3
Definitions and interpretations of various endogenous Bayesian risk functions.

Risk Function	Notation	Key Determinant	Description
Aleatoric Bayes' Risk	$\bar{\bar{r}}\left(\pi_\theta^\star \| \pi_X^\star; B_t\right)$	g_θ	Inherent OLP randomness
Epistemic Risk	$\bar{\bar{r}}\left(\hat{\pi}_{S_t}^\star \| \pi_\theta^\star; B_t\right)$	b_t	Uncertainty of b_t about θ
Process Risk	$\bar{\bar{r}}\left(\pi \| \hat{\pi}_{S_t}^\star; B_t\right)$	π	Relative risk of the policy π
Total Bayes' Risk	$\bar{\bar{r}}\left(\pi \| \pi_X^\star; B_t\right)$		Cumulative risk of the above

about the hidden parameters θ, the posterior ODOL policy would perform at least as well as *any* closed-loop policy, including the optimal closed-loop policy.

Proposition 1. *Suppose the hidden parameters θ are identifiable through some infinite sequence of actions. Over this sequence of actions, then $\forall \pi \in \Pi$, $\exists c_\pi \geq 0$ such that*

$$\text{ProcessRisk}\left(\pi; S_t\right) \xrightarrow[t\to\infty]{P} c_\pi. \tag{4.5}$$

Furthermore, the process risk of any policy is bounded below by the process risk of $\pi_b^\star$,

$$\text{ProcessRisk}(\pi; S_t) \geq \text{ProcessRisk}(\pi_b^\star; S_t)\ \forall \pi \in \Pi, \tag{4.6}$$

and so, over this same sequence of actions,

$$V^{\hat{\pi}_{S_t}^\star}(S_t) \xrightarrow[t\to\infty]{P} V^{\pi_b^\star}(S_t) \tag{4.7}$$

Proof. Eq. (4.5) follows from expanding $\bar{\bar{r}}(\pi \| \pi_\theta^\star; S) = \text{EpistemicRisk}(S) + \text{ProcessRisk}(\pi; S)$ and applying Lemma 1, followed by Theorem 1, and finally the continuous mapping theorem. Eq. (4.6) follows from expanding the definition of process risk and applying Lemma 1, and then Eq. (4.7) follows from Eq. (4.5) while noting that $\text{ProcessRisk}(\pi_b^\star; S_t) \leq \text{ProcessRisk}(\hat{\pi}_{S_t}^\star; S_t) = 0$. □

Note that in Proposition 1, the sequence of actions taken is independent of the policy π for which the risk is computed. This proposition reinforces that ODOL policies improve over time as long as the belief distribution converges on θ, and furthermore that under such conditions they asymptotically match the performance of even the best closed-loop policy $\pi_b^\star$.

Together, the aleatoric, epistemic, and process risks decompose the total risk according to

$$\text{TotalRisk}(\pi; S) \equiv \text{AleatoricBayesRisk}(S) + \text{EpistemicRisk}(S) + \text{ProcessRisk}(\pi; S). \tag{4.8}$$

A summary of the definition and interpretation of each risk is presented in Table 3. The process risk of a policy is the only component of the total risk that depends on the policy under consideration, π. Thus, a policy with less process risk in some state than another policy also has less total risk than that policy. While computing the process risk of a general policy is just as hard as computing that policy's value, computing the process risk of policies "similar" to the ODOL policy, such as lookahead policies, can be easier. This provides an efficient way to compare such policies without needing to explicitly calculate each policy's value or total risk. This observation will be leveraged in Section 5 to enable computationally efficient estimates of Bayes' risk and to efficiently identify the best policy to follow from any particular belief state.

4.2. Computing policy value

Computing the risk of one policy relative to another is equivalent to comparing their respective values, and the main obstacle to computing the value of a policy is the size of its *reachable belief set*.

Definition 7. The **reachable belief set** (RBS) from an initial state S_0 is the set of all belief states that could result from some sequence of actions $A \in \bigcup_{t=1}^{\infty} \mathcal{A}^t$ and observations $Y \in \bigcup_{t=1}^{\infty} \mathcal{Y}^t$ (where $|A| = |Y|$).

Definition 8. The **π-reachable belief set** (π-RBS) is the smallest subset of the RBS that contains all belief states that a learning agent following the policy π, initialized at S_0, could eventually transition into.

Definition 9. We call the reachable belief set **unbounded** if $\forall h \in \mathbb{N}_{>0}$, there exists a sequence of actions $a_{1:h}$ that do *not* lead to a terminal state. Otherwise, the RBS is **bounded with decision horizon** h, where h is the length of the longest sequence of actions that results in a terminal state.

Whether the RBS is bounded or unbounded depends on the specification of ξ_0, δ, and Ω. Assuming discrete action and observation sets, the size of a bounded RBS with decision horizon h is $\mathcal{O}(|\mathcal{A}\times\mathcal{Y}|^h)$.[5] This rapid growth means it is generally impractical to compute the value of an arbitrary policy (where the π-RBS may match the full RBS) using MDP algorithms like value iteration; even describing an arbitrary policy requires $\mathcal{O}(|\mathcal{A}\times\mathcal{Y}|^h)$ values to encode the action distribution of the policy at every reachable state. Accordingly, we focus our analysis on policies which have a small π-RBS, or otherwise have some structure making it tractable to compute their value; in particular, this is the case for deterministic open-loop, m-lookahead, X-optimal, and θ-optimal policies, which will be explored in the following subsections.

4.2.1. Deterministic open-loop policies

We begin by defining the set of all terminating feasible action sequences from a state $S_t = \{b_t, \xi_t\}$,

$$A^{\infty}_{\Omega}(\xi_t) = \left\{ A \in \bigcup_{n=1}^{\infty} \mathcal{A}^n \;\middle|\; \begin{array}{l} \xi_{t+\tau} = \delta\left(\xi_{t+\tau-1}, A(\tau)\right)\ \forall \tau \in \{1,\dots,|A|\}, \\ A(\tau) \in \Omega(\xi_{t+\tau-1})\ \forall \tau \in \{1,\dots,|A|\}, \\ \Omega(\xi_{t+|A|}) = \varnothing. \end{array} \right\} \tag{4.9}$$

A deterministic open-loop policy $\hat{\pi} \in \hat{\Pi}$ takes a fixed sequence of actions $A^{\hat{\pi}}$ from state S_t, so its value is

$$V^{\hat{\pi}}(S_t) := \mathbb{E}_{X|b_t}\left[\gamma^{|A^{\hat{\pi}}|} R\left(b^{X,A^{\hat{\pi}}}_{t+|A^{\hat{\pi}}|}\right) + \sum_{\tau=1}^{|A^{\hat{\pi}}|} \gamma^{\tau} R\left(X(t+\tau), A^{\hat{\pi}}(\tau)\right)\right], \tag{4.10}$$

where $b^{X,A}_{t+|A|}$ denotes the posterior belief distribution of the terminal state reached after following the action sequence A with corresponding hidden outcomes X. Now, since the set of all feasible deterministic open-loop policies from the state S_t is isomorphic with $\mathcal{A}^{\infty}_{\Omega}(\xi_t)$, the value of the ODOL policy $\hat{\pi}^{\star}_{S_t}$ is the solution to a maximization problem over $\mathcal{A}^{\infty}_{\Omega}(\xi_t)$,

$$V^{\hat{\pi}^{\star}_{S_t}}(S_t) = \max_{A \in \mathcal{A}^{\infty}_{\Omega}(\xi_t)} \mathbb{E}_{X|b_t}\left[\gamma^{|A|} R\left(b^{X,A}_{t+|A|}\right) + \sum_{\tau=1}^{|A|} \gamma^{\tau} R\left(X(t+\tau), A(\tau)\right)\right]. \tag{4.11}$$

Eq. (4.11) is a non-linear discrete optimization problem which may be, in general, very challenging to solve. However, there are many conditions under which it is more tractable. For example, most classical bandit and partial monitoring problems are formulated with some given $T \in \mathbb{N}_{>0}$ such that $\Omega(\xi_t) = \mathcal{A}$, $\forall t < T$ (and $\Omega(\xi_t) = \varnothing$, $\forall t \geq T$) and $R(b_t) = 0\ \forall b_t$. In such problems, Eq. (4.11) simplifies to

$$V^{\hat{\pi}^{\star}_{S_t}}(S_t) = \sum_{\tau=t+1}^{T} \max_{a_\tau \in \mathcal{A}} \gamma^{\tau} \mathbb{E}_{x|b_t}\left[R\left(x, a_\tau\right)\right], \tag{4.12}$$

where the $(T-t)$ unconstrained maximization problems can be solved independently. Subsection 4.3 will explore broader conditions under which Eq. (4.11) can be simplified and solved with suitable algorithms.

4.2.2. Lookahead policies

An m-step lookahead policy π is one that operates in closed-loop for $m \in \mathbb{N}_{>0}$ actions before transitioning to the ODOL policy for the posterior belief state. This results in a π-RBS of size $O(|\mathcal{A}\times\mathcal{Y}|^m + |\mathcal{Y}|^{h-m})$. The value of an m-lookahead policy is equal to the sum of the expected action rewards over the closed loop phase and the total reward expected to be collected by the posterior ODOL policy from the posterior belief state. This fact is leveraged by the proposed EBRM algorithms and will be discussed further in Section 5.

4.2.3. X-optimal policies

An X-optimal policy is deterministic and open-loop because the belief and action rewards are deterministic with respect to the sequence of actions taken by the agent. The conditional value of $\pi^{\star}_X$ in state S_t given X is thus the solution of the maximization problem, with $b^{X,A}_{t+|A|}$ defined as previously,

$$V^{\pi^{\star}_X}(S_t; X) := \max_{A \in \mathcal{A}^{\infty}_{\Omega}(\xi_t)} \left[\gamma^{|A|} R\left(b^{X,A}_{t+|A|}\right) + \sum_{\tau=1}^{|A|} \gamma^{\tau} R\left(X(t+\tau), A(\tau)\right)\right]. \tag{4.13}$$

Subsection 4.3.2 will discuss conditions under which this problem can be further simplified.

Given a means to compute the conditional value of the X-optimal policy for a particular realization of the hidden outcomes, its unconditional value is an expectation over the belief distribution b_t,

$$V^{\pi^{\star}_X}(S_t) = \mathbb{E}_{X|b_t}\left[V^{\pi^{\star}_X}(S_t; X)\right], \tag{4.14}$$

[5] For continuous action or observation sets, there is similarly exponential growth in the *dimensionality* of the RBS.

which can be computed with sampling-based methods, even if it cannot be solved analytically.

4.2.4. θ-optimal policies

A θ-optimal policy is generally closed-loop, as despite knowing the exact distribution from which the hidden outcomes are generated, maximizing the belief reward requires choosing actions based on how prior observations have shaped the belief distribution. However, under certain conditions on the belief reward function $R(b_t)$, there exists a deterministic open-loop θ-optimal policy. These conditions will be discussed in Subsection 4.3.2, and enable efficient computation of the conditional value function $V^{\pi^\star_\theta}(S_t;\theta)$. Given a means to compute this conditional value, the unconditional value of $\pi^\star_\theta$ is given by the expectation,

$$V^{\pi^\star_\theta}(S_t) = \mathbb{E}_{\theta|b_t}\left[V^{\pi^\star_\theta}(S_t;\theta)\right]. \tag{4.15}$$

4.3. Efficient construction of optimal deterministic open-loop policies

This subsection will explore approaches to efficiently solve Eq. (4.11) and construct an ODOL policy $\hat{\pi}^\star_{S_t}$ for the belief state S_t, under the following simplifying assumptions.

Assumption 1 *(Bounded RBS)*. We assume that the RBS is bounded with some decision horizon $h \in \mathbb{N}_{>0}$.

Assumption 2 *(Discrete Action Set)*. We assume that $\mathcal{A}$ is discrete and finite with cardinality $K \in \mathbb{N}_{>0}$. Without further loss of generality, we can assume that $\mathcal{A} = [K]$.

Assumption 3 *(Order-Independent Feasibility)*. We assume that, if $A \in \mathcal{A}^\infty_\Omega(\xi_t)$ and A' is a permutation of A, then $A' \in \mathcal{A}^\infty_\Omega(\xi_t)$. This is equivalent to the following conditions on Ω and δ:

1. $\Omega(\delta(\xi,a)) \subseteq \Omega(\xi) \quad \forall \xi, a : a \in \Omega(\xi)$
2. $a_2 \in \Omega\left(\delta\left(\xi,a_1\right)\right) \iff a_1 \in \Omega\left(\delta\left(\xi,a_2\right)\right) \quad \forall \xi, a_1, a_2 : \{a_1,a_2\} \subseteq \Omega(\xi)$

The key result of these assumptions is that the feasibility of a deterministic open-loop policy is determined by its *action counts* $N \in \mathbb{N}^K$, where an action sequence A is "consistent with N" if and only if

$$N(k) = \sum_{\tau=1}^{|A|} \mathbb{1}(A(\tau) = k) \quad \forall k \in [K]. \tag{4.16}$$

For any $N \in \mathbb{N}^K$, we can easily construct an action sequence A consistent with N by taking $N(1)$ copies of action 1, followed by $N(2)$ copies of action 2, and so on. If we let A' denote any permutation of A then, by Assumption 3, $A \in \mathcal{A}^\infty_\Omega(\xi_t) \iff A' \in \mathcal{A}^\infty_\Omega(\xi_t)$. We thus define $\mathbb{N}^K_\Omega(\xi_t) \subseteq \mathbb{N}^K$ such that

$$N \in \mathbb{N}^K_\Omega(\xi_t) \iff A \in \mathcal{A}^\infty_\Omega(\xi_t). \tag{4.17}$$

So, we can determine whether $N \in \mathbb{N}^K_\Omega(\xi_t)$ by constructing A and testing if $A \in \mathcal{A}^\infty_\Omega(\xi_t)$. Proposition 2 and its corollary will show that we can just as easily construct the *optimal* action sequence consistent with N. First, however, we show that the expected posterior belief reward following any action sequence N is a function of only the action counts, as seen in Lemma 2.

Lemma 2. *Let A be any action sequence consistent with $N \in \mathbb{N}^K_\Omega(\xi_t)$ from some initial state $S_t = \{b_t, \xi_t\}$, and let A' be any permutation of A. Then,*

$$\mathbb{E}_{X|b_t}\left[R\left(b^{X,A}_{t+|A|}\right)\right] = \mathbb{E}_{X|b_t}\left[R\left(b^{X,A'}_{t+|A'|}\right)\right], \tag{4.18}$$

where $b^{X,A}_{t+|A|}$ is the posterior belief distribution following action-observation pairs $\{(A(\tau), Y(\tau))\}_{\tau=1}^{|A|}$, with $Y(\tau) = \Phi(X(\tau), A(\tau))$ for each hidden outcome $X(\tau)$, and the prior b_t. For convenience, we thus define,

$$\bar{R}(N; b_t) := \mathbb{E}_{X|b_t}\left[R\left(b^{X,A}_{t+|A|}\right)\right]. \tag{4.19}$$

Proof. Refer to Appendix D. □

Proposition 2. *Let $k_1, \ldots, k_K$ be an ordering of the elements of $\mathcal{A}$ such that, given S_t,*

$$\mathbb{E}_{x|b_t}\left[R(x,k_1)\right] \geq \mathbb{E}_{x|b_t}\left[R(x,k_2)\right] \geq \cdots \geq \mathbb{E}_{x|b_t}\left[R(x,k_K)\right]. \tag{4.20}$$

For any $N \in \mathbb{N}^K_\Omega(\xi_t)$, an optimal action sequence $A^\star_N \in \mathcal{A}^\infty_\Omega(\xi_t)$ consistent with N can be constructed by taking $N(k_1)$ copies of action k_1, followed by $N(k_2)$ copies of action k_2, and so on. The policy $\hat{\pi}^\star_N$ described by $A^\star_N$ satisfies,

$$V^{\hat{\pi}^\star_N}(S_t) \geq V^{\hat{\pi}}(S_t) \quad \forall \hat{\pi} : A^{\hat{\pi}} \text{ is consistent with } N. \tag{4.21}$$

Accordingly, we define the action count optimal value function,

$$\hat{V}(N; S_t) := V^{\hat{\pi}^\star_N}(S_t) \tag{4.22}$$

$$= \gamma^{n_K} \bar{R}(N; b_t) + \sum_{i=1}^{K} \frac{\gamma^{n_{i-1}} \left(1 - \gamma^{N(k_i)}\right)}{1-\gamma} \mathbb{E}_{x|b_t}\left[R(x, k_i)\right], \tag{4.23}$$

where $n_0 := 0$ *and* $n_i := \sum_{j=1}^{i} N(k_j)$, $\forall i \in [K]$. *Note that* $\lim_{\gamma \to 1} \left[\gamma^{n_{i-1}} \left(1 - \gamma^{N(k_i)}\right)(1-\gamma)^{-1}\right] = N(k_i)$.

Proof. Follows from Lemma 2 and the monotonically non-increasing weight of later action rewards. □

Corollary 1. *The value of the ODOL policy* $\hat{\pi}^\star_{S_t}$ *in state* S_t *is*

$$V^{\hat{\pi}^\star_{S_t}}(S_t) = \max_{N \in \mathbb{N}^K_\Omega(\xi_t)} \hat{V}(N; S_t). \tag{4.24}$$

Corollary 1 means an exhaustive search for the ODOL policy can be done by evaluating the $O(h^K)$ elements of $\mathbb{N}^K_\Omega(\xi_t)$, rather than all $O(K^h)$ sequences in $\mathcal{A}^\infty_\Omega(\xi_t)$. This is often an improvement as, typically, $h \gg K$. Once $N^\star$ is found, an ODOL policy can be constructed by the approach used in Proposition 2.

4.3.1. Common conditions for integer programming solutions

The maximization problem in Eq. (4.24) is still a non-linear integer program without a general polynomial-time solution. However, for many common feasibility criteria, reward functions, and discount factors, it can be further simplified to be efficiently solved or approximated by existing algorithms.

The first element to consider is the feasibility criterion; most integer programming algorithms permit only *linear* constraints on the optimization variable. This is equivalent to requiring that there exist a weight matrix $W \in \mathbb{R}^{M \times K}$ and budget vector $c \in \mathbb{R}^M$ such that $N \in \mathbb{N}^K_\Omega(\xi_t) \iff WN \preceq c$. This requirement is satisfied for *temporal* and *knapsack* feasibility criteria, defined below.

Definition 10. A **temporal feasibility criterion** is one defined as, given some horizon $T \in \mathbb{N}_{>0}$,

$$\xi_0 = 0, \qquad \delta\left(\xi_{t-1}, a_t\right) = \xi_{t-1} + 1, \qquad \Omega(\xi_t) = \begin{cases} \mathcal{A}, & \xi_t < T \\ \varnothing, & \text{otherwise.} \end{cases}$$

Definition 11. A **knapsack feasibility criterion** is defined as, given a budget $c \in \mathbb{R}_{>0}$, a weight vector $w \in \mathbb{R}^K_{>0}$, and bounds $u_1, \ldots, u_K \in \mathbb{N} \cup \{+\infty\}$,

$$\xi_0 = \mathbf{0}^K, \qquad \delta\left(\xi_{t-1}, a\right) = \xi_{t-1} + \chi_a, \qquad k \in \Omega(\xi_t) \iff \begin{cases} \langle w, \xi_t \rangle + w(k) \leq c, \\ \quad \text{and } \xi_t(k) < u_k, \end{cases}$$

where the characteristic vector $\chi_k \in \{0,1\}^K$ satisfies $\chi_k(i) = 1 \iff i = k$.

Next, we consider how the action reward interacts with the feasibility criterion and discount factor. If action rewards are negative, the agent may be driven towards a terminal state as quickly as possible if the feasibility criterion permits it. As this may not be desirable behaviour in all cases, care must be taken when defining action rewards or shifting their values by a constant. Furthermore, negative action rewards can lead to the agent instead seeking *delays*, particularly if $\gamma < 1$; if taking an action with negative expected reward is required to satisfy the feasibility criterion, the optimal solution may precede that action with an arbitrarily long sequence of low or zero reward actions, in order to discount it.

Lastly, we consider how the belief reward interacts with the feasibility criterion and discount factor. As with action reward, if the belief reward can be negative, the agent may seek out or delay termination; as belief reward is earned only at termination, care must also be taken if actions can reduce the expected posterior belief reward. The problem is thus simpler if the belief reward function is non-negative and *adaptive monotone* [55]: $\forall b_t, \forall a$, $\mathbb{E}_{x|b_t}\left[R(b^{x,a}_{t+1})\right] \geq R(b_t) \geq 0$. Such belief reward functions are common, as belief rewards generally measure how much "information" has been learned about one or more of the hidden parameters; as the hidden parameters θ are fixed, taking an action never causes a learning agent to lose information. By Jensen's inequality, adaptive monotonicity is satisfied if $R(b)$ is convex in b; such belief rewards include the negative entropy of b, and the log generalized-precision $\log\det\Sigma^{-1}$, where Σ is the covariance matrix of b. Even with a non-negative, adaptive monotone belief reward function, termination-seeking behaviour may arise for a discount factor $\gamma < 1$; if the agent cannot increase its terminal belief reward by a factor of at least γ^{-1} with each action, it will seek out a terminal state as quickly as possible.

Table 4
Solvers for Eq. (4.24) under common feasibility criteria and expected belief reward properties.

Feasibility Criterion	γ	$\bar{R}(N;b)$	Problem Class	Solver	Approx. Factor	Runtime Complexity
Temporal	$(0,1]$	$0, \forall N$	Concave Maximization	Greedy	1	$O(K)$
Temporal	1	Concave & Modular	Separable Concave Maximization	Iterative Greedy	1	$O(Kh)$
Temporal	$(0,1]$	M-Concave [57]	Integer M-Concave Maximization [57]	Steepest Ascent [58]	1	$O(K^2h)$
Temporal	$(0,1]$	Concave & Submodular	Monotone Submodular Maximization [59]	Iterative Greedy	$1-e^{-1}$	$O(Kh)$
Knapsack	1	Linear & Modular	(Un-)Bounded Knapsack [60]	Dynamic Program	1	$O(Kh)$
Knapsack	$(0,1]$	$0, \forall N$	Submodular Cost, Submodular Knapsack [61]	Iterative Greedy	$1-e^{-1}$	$O(Kh)$
Knapsack	1	Concave & Submodular				

Based on the issues discussed above, we assume a non-negative adaptive monotone belief reward and either non-negative action rewards or a temporal feasibility criterion.[6] Given these assumptions, the choice of integer programming algorithm depends primarily on the discount factor and form of the expected posterior belief reward as a function of N. For example, efficient algorithms are known for $\bar{R}(N;b_t)$ that are concave and modular or submodular in N.[7] The contribution of action reward as a function of N is linear and modular when $\gamma = 1$, and generally $M^\natural$-*concave* [56] and submodular otherwise. A non-exhaustive list of algorithms for efficiently solving Eq. (4.24) given various combinations of feasibility criterion, belief reward, and discount factor is presented in Table 4. The approximation factor measures the ratio between the value of the policy found and the value of the true ODOL policy.

4.3.2. Applications to solving θ- and X-optimal policies

The condition which most simplifies finding and solving the value of θ-optimal policies is an *observation-independent* belief reward, such that the reward assigned to a posterior belief distribution is a function of only the initial belief distribution and the actions taken. As a common example, if the posterior belief distribution b is normal with covariance matrix Σ, then a belief reward function satisfying $\exists f : R(b) = f(\Sigma)$ is observation-independent. Under this condition, finding the θ-optimal policy reduces to solving for action counts, so one of the solvers from Table 4 may apply.

While there is always a deterministic open-loop X-optimal policy for a given X, the main source of complexity in finding it and solving its value is, likewise, the belief reward function. A strong condition which simplifies finding and solving the value of X-optimal policies is a *separable* belief reward and a temporal feasibility criterion, such that $\exists f : R(b_{t+1}^{x,a}) = R(b_t) + f(x,a)$. The problem still cannot be reduced to solving for action counts, however it can be simplified into $O(h)$ univariate optimization problems.

5. The endogenous Bayesian risk minimization approach

We propose Endogenous Bayesian Risk Minimization (EBRM) as a general-purpose approach to optimal online learning. The goal of an EBRM policy is to, at each belief state S_t, find and imitate a policy that is optimal from that state. To make this approach computationally tractable, we constrain our search to a small set of simple candidate policies $\Pi_t \subset \Pi$. In each state, the EBRM policy imitates the policy in Π_t with the maximum value,

$$\pi_{\text{EBRM}}(S_t) := \arg\max_{\pi \in \Pi_t} V^\pi(S_t) \tag{5.1}$$

The set of candidate policies is the only "hyperparameter" of an EBRM algorithm.

By imitating the candidate policy with the highest value in each decision step, the value of the best candidate policy is a lower bound on the value of the EBRM policy, and the Bayes' risk of the best candidate policy similarly upper bounds the Bayes' risk of EBRM. Thus an EBRM policy accumulates, in expectation, more reward than any individual candidate policy would.

5.1. Greedy-EBRM: EBRM using one-step lookahead candidate policies

Greedy-EBRM is the simple yet highly effective online learning algorithm which results from choosing the candidate policy set composed of all deterministic 1-step lookahead policies,

[6] These assumptions are also driven by practical considerations as solvers for many classes of integer programs, such as knapsack problems, typically expect non-negative objectives.

[7] Submodularity is equivalent to only non-positive off-diagonal elements in the Hessian of $\bar{R}(N;b_t)$, while supermodularity is equivalent to only non-negative off-diagonal elements. A modular function is both submodular and supermodular.

$$\Pi_t = \{\pi_{t,a}\}_{a\in\mathcal{A}}, \tag{5.2}$$

$$\pi_{t,a}(S_t) = \text{Take action } a\text{, observe } y\text{, then proceed according to the posterior ODOL policy.} \tag{5.3}$$

The value of a deterministic 1-step lookahead policy is therefore

$$V^{\pi_{t,a}}(S_t) = \mathbb{E}_{x|b_t}\left[R(x,a) + \max_{N\in\mathbb{N}^K_\Omega\left(\xi^a_{t+1}\right)} \gamma\hat{V}\left(N; S^{x,a}_{t+1}\right)\right], \tag{5.4}$$

where $\hat{V}(N; S^{x,a}_{t+1})$, defined in Eq. (4.24), is the value of the best open-loop policy for $S^{x,a}_{t+1} = \{b^{x,a}_{t+1}, \xi^a_{t+1}\}$ consistent with action counts N. The maximization problem is equivalent to the one in Eq. (4.24), but with a shorter decision horizon, so it can be calculated using the same algorithm that found the ODOL policy. However, there is generally no closed-form solution to the *expectation* of the maximum, and so the expectation often requires sampling-based methods to compute.

Fortunately, we can use the epistemic and process risk functions to help eliminate sub-optimal policies and obtain probabilistic bounds on the expectation in Eq. (5.4). First, we observe that

$$\pi^\star_t \in \arg\max_{\pi\in\Pi_t} V^\pi(S_t) \iff \pi^\star_t \in \arg\min_{\pi\in\Pi_t} \text{ProcessRisk}(\pi; S_t). \tag{5.5}$$

We then decompose the process risk of each $\pi_{t,a}$ into two terms: the *immediate risk* of a, which represents the "opportunity cost" of choosing action a instead of following the ODOL policy, and the *expected value of information* gained from the observation y which would be generated by the action a.

Algorithm 1: Greedy-EBRM.

Given: OLP Specification
Input: Belief State S_t
`InitialRisk`$(S_t) \leftarrow$ `AleatoricBayesRisk`(S_t) + `EpistemicRisk`(S_t)
foreach $a \in \mathcal{A}$:
  `ProcessRisk`$(\pi_{t,a}; S_t) \leftarrow$ `ImmediateRisk`$(S_t, a) - \gamma \cdot$ `ExpectedVoI`(S_t, a)
end foreach
$a^\star_{t+1} \leftarrow \arg\min_{a\in\mathcal{A}}$ [`ProcessRisk`$(\pi_{t,a}; S_t)$] // Optimal action
$U_t \leftarrow$ `InitialRisk`(S_t) + `ProcessRisk`$(\pi_{t,a^\star_{t+1}}; S_t)$ // Upper bound on Greedy-EBRM Bayes' risk
return $a^\star_t, U_t$

We begin by defining the immediate risk of an action a, which quantifies how much better the current ODOL policy is expected to perform than *the best deterministic open-loop policy that begins by picking* a. Equivalently, this measures how much additional regret is expected to be incurred over the ODOL policy by being forced to take action a as the next action. Importantly, both policies are constructed according only to the current belief state, without considering possible future observations.

Definition 12. The **immediate risk** of performing an action $a \in \mathcal{A}$ in a belief state $S_t \in \mathcal{S}$ is defined as

$$\text{ImmediateRisk}(S_t, a) := V^{\hat{\pi}^\star_{S_t}}(S_t) - \left(\mathbb{E}_{x|b_t}\left[R(x,a)\right] + \max_{N\in\mathbb{N}^K_\Omega\left(\xi^a_{t+1}\right)} \mathbb{E}_{x|b_t}\left[\gamma\hat{V}\left(N; S^{x,a}_{t+1}\right)\right]\right). \tag{5.6}$$

The term on the right side of Eq. (5.6) represents the value of the best deterministic open-loop policy which begins by taking action a.[8] At least one of these policies is an ODOL policy for the state S_t; therefore, computing all K immediate risks requires computing the value of exactly K deterministic open-loop policies, and it holds $\forall S_t, \forall a$ that $\text{ImmediateRisk}(S_t, a) \geq 0$.

Next, the expected value of information of an action a, defined below, measures how much the value of an ODOL policy for the posterior state $S^{x,a}_{t+1}$ constructed with the additional observation $\Phi(x,a)$ surpasses, in expectation, one that is constructed without it.

Definition 13. The **expected value of information** gained by performing action a in belief state S_t is

$$\text{ExpectedVoI}(S_t, a) := \mathbb{E}_{x|b_t}\left[\max_{N\in\mathbb{N}^K_\Omega\left(\xi^a_{t+1}\right)} \hat{V}\left(N; S^{x,a}_{t+1}\right)\right] - \max_{N\in\mathbb{N}^K_\Omega\left(\xi^a_{t+1}\right)} \mathbb{E}_{x|b_t}\left[\hat{V}\left(N; S^{x,a}_{t+1}\right)\right]. \tag{5.7}$$

While the expected value of information is driven by *uncertainty* in the belief distribution and the amount of *information* gained from an observation, it is measured in units of real reward (value). It is also closely related to epistemic risk, as shown in Lemma 3.

[8] Recall that, even if the ODOL policy $\hat{\pi}^\star_{S_t}$ would take action a at least once, it only begins with a if $a \in \arg\max_a \mathbb{E}_{x|b}[R(x,a)]$.

Lemma 3. *The expected value of information is bounded according to,*

$$0 \leq \text{ExpectedVoI}(S_t, a) \leq \text{EpistemicRisk}\left(\mathbb{E}_{x|b_t}\left[S_{t+1}^{x,a}\right]\right) - \mathbb{E}_{x|b_t}\left[\text{EpistemicRisk}\left(S_{t+1}^{x,a}\right)\right]. \tag{5.8}$$

Proof. Refer to Appendix E. □

The expected value of information represents the rate of convergence of the value of the ODOL policy to the value of the θ-optimal policy, as described in Theorem 1. Together, the expected value of information and immediate risk measure the process risk of any 1-step lookahead policy:

$$\text{ProcessRisk}(\pi_{t,a}; S_t) = \text{ImmediateRisk}(S_t, a) - \gamma \cdot \text{ExpectedVoI}(S_t, a). \tag{5.9}$$

The first practical result of this analysis, presented in Proposition 3, enables Greedy-EBRM to eliminate from consideration policies in the candidate policy set that are known to be sub-optimal.

Proposition 3. *If the immediate risk of an action $a \in \mathcal{A}$ exceeds the epistemic risk of the state $\mathbb{E}_{x|b_t}\left[S_{t+1}^{x,a}\right] = \{b_t, \xi_{t+1}^a\}$, then the deterministic 1-step lookahead policy $\pi_{t,a}$ is not a minimizer of the process risk. That is,*

$$\text{ImmediateRisk}(S_t, a) > \gamma \cdot \text{EpistemicRisk}\left(\mathbb{E}_{x|b_t}\left[S_{t+1}^{x,a}\right]\right) \implies \pi_{t,a} \notin \arg\min_{\pi \in \Pi_t} \text{ProcessRisk}(\pi; S_t). \tag{5.10}$$

Proof. Refer to Appendix E. □

Applying the bound in Proposition 3 requires, for each action $a \in \mathcal{A}$, both the immediate risk of the action and the values of the ODOL and θ-optimal policies for the state $\mathbb{E}_{x|b_t}\left[S_{t+1}^{x,a}\right] = \{b_t, \xi_{t+1}^a\}$, in order to compute the epistemic risk bound. Then, estimating the process risk of each 1-step lookahead policy $\pi_{t,a}$ (excluding those eliminated from consideration under Proposition 3) can be done using sampling-based methods to estimate the expected value of information gained by action a. Using each sample to estimate the value of information requires finding and computing the value of an ODOL policy, as shown in Lemma 4.

Lemma 4. *Let $x_i \overset{\text{iid}}{\sim} \Pr\left(x \mid b_t\right)$ be independently and identically distributed hidden outcomes conditioned on the belief distribution b_t of state $S_t = \{b_t, \xi_t\}$, and let*

$$N_{t+1}^a \in \underset{N \in \mathbb{N}_\Omega^K\left(\xi_{t+1}^a\right)}{\arg\max} \mathbb{E}_{x|b_t}\left[\hat{V}\left(N; S_{t+1}^{x,a}\right)\right]. \tag{5.11}$$

The estimated value of information gained by action $a \in \mathcal{A}$ in this state, based on $n \in \mathbb{N}_{>0}$ samples, is

$$\nu_a(n) := \frac{1}{n}\sum_{i=1}^{n}\left[\max_{N \in \mathbb{N}_\Omega^K\left(\xi_{t+1}^a\right)} \hat{V}\left(N; S_{t+1}^{x_i,a}\right) - \hat{V}\left(N_{t+1}^a; S_{t+1}^{x_i,a}\right)\right]. \tag{5.12}$$

By construction it holds that $\nu_a(n) \geq 0$, and by the linearity of expectation $\mathbb{E}\left[\nu_a(n)\right] = \text{ExpectedVoI}(S_t, a)$.

Following from Lemma 4, the value of information has finite variance if and only if $\exists \sigma_a \in \mathbb{R}$ such that

$$\mathbb{E}_{x|b_t}\left[\left(\max_{N \in \mathbb{N}_\Omega^K\left(\xi_{t+1}^a\right)} \hat{V}\left(N; S_{t+1}^{x,a}\right) - \hat{V}\left(N_{t+1}^a; S_{t+1}^{x,a}\right) - \text{ExpectedVoI}(S_t, a)\right)^2\right] \leq \sigma_a^2. \tag{5.13}$$

We further call the value of information bounded by $M_a \in \mathbb{R}$ if and only if

$$\Pr\left(\max_{N \in \mathbb{N}_\Omega^K\left(\xi_{t+1}^a\right)} \hat{V}\left(N; S_{t+1}^{x,a}\right) - \hat{V}\left(N_{t+1}^a; S_{t+1}^{x,a}\right) \leq M_a\right) = 1. \tag{5.14}$$

Assuming the value of information has finite variance, Theorem 2 provides probabilistic bounds on whether $\pi_{t,a} \in \arg\min_\pi \text{ProcessRisk}(\pi; S_t)$. These bounds depend on the number of samples used to estimate the value of information of each action. Corollary 2 shows how the number of samples can be picked to reach any desired level of confidence that a policy $\pi_{t,a}$ is sub-optimal, before rejecting it. Under the stronger condition that the value of information is bounded, Theorem 3 provides a lower bound on the number of samples required to bound the expected *risk* of rejecting a policy $\pi_{t,a}$ to some value $\epsilon > 0$. Proofs for these results are provided in Appendix E.

Theorem 2. *Suppose that for each $a \in \mathcal{A}$, $n_a > 0$ samples have been used to estimate the expected value of information gained by action a in the belief state S_t. Then, define*

$$k \in \arg\min_{a'} \left[\text{ImmediateRisk}(S_t, a') - \gamma \cdot \nu_{a'}(n_{a'})\right], \tag{5.15}$$

$$\mu_a := \left(\text{ImmediateRisk}(S_t, a) - \text{ImmediateRisk}(S_t, k)\right) - \gamma\left(\nu_a(n_a) - \nu_k(n_k)\right). \tag{5.16}$$

If the values of information gained by actions a and k have finite variances σ_a^2 and σ_k^2, respectively, then

$$\mu_a > 0 \implies \Pr\left(\pi_{t,a} \notin \arg\min_{\pi \in \Pi_t} \text{ProcessRisk}(\pi; S_t)\right) \geq \frac{\mu_a^2 n_a n_k}{\gamma^2\left(\sigma_a^2 n_k + \sigma_k^2 n_a\right) + \mu_a^2 n_a n_k}. \tag{5.17}$$

Corollary 2. *It follows algebraically from Theorem 2 that, $\forall \epsilon > 0$ and $\forall a : \mu_a > 0$,*

$$n_k \geq \left\lceil \frac{2\sigma_k^2 \gamma^2 (1-\epsilon)}{\mu_a^2 \epsilon} \right\rceil, \; n_a \geq \left\lceil \frac{2\sigma_a^2 \gamma^2 (1-\epsilon)}{\mu_a^2 \epsilon} \right\rceil \implies \Pr\left(\pi_{t,a} \in \arg\min_{\pi \in \Pi_t} \text{ProcessRisk}(\pi; S_t)\right) \leq \epsilon. \tag{5.18}$$

Theorem 3. *Suppose that the value of information gained by each action $a \in \mathcal{A}$ is bounded by M_a, and $n_a = n_k = n$. Then, following the notation of Theorem 2 and Corollary 2, by Hoeffding's inequality [62],*

$$\Pr\left(\pi_{t,a} \notin \arg\min_{\pi \in \Pi_t} \text{ProcessRisk}(\pi; S_t)\right) \geq 1 - \exp\left(\frac{-2n\mu_a^2}{\gamma^2\left(M_a + M_k\right)^2}\right). \tag{5.19}$$

Under these assumptions, $\forall \epsilon > 0$ and $\forall a : \mu_a > 0$, if

$$n \geq \left\lceil \frac{\gamma^2\left(M_a + M_k\right)^2}{2\mu_a^2} \ln\left(\frac{\gamma\left(M_a + M_k\right)}{\epsilon}\right) \right\rceil \tag{5.20}$$

then the expected risk of eliminating policy $\pi_{t,a}$ is bounded above by ϵ.

In practice, tighter bounds may be possible by taking into account the problem-specific structure of the action count value function $\hat{V}(N; S_t)$. It is worth noting that while the number of samples required to confidently reject a policy $\pi_{t,a}$ grows as $\mu_a \to 0$, small values of μ_a indicate that the expected difference in process risk (value) compared to the best reference policy $\pi_{t,k}$ is small; thus, the expected risk of rejecting $\pi_{t,a}$ may be small even if there is a non-trivial probability that it is optimal. Furthermore, even a single sample of the expected value of information for each action is sufficient to provide an unbiased estimate of the best policy in the policy set. Thus, these results are best applied in settings where risk quantification is critical; in most cases, the number of samples is likely chosen based on the computational resources available.

5.2. Overcoming the myopia of one-step lookahead policy sets

The expected value of information for 1-step lookahead policies cannot account for the value of information from *multiple* actions. Consequentially, these policies tend to have poor performance when multiple actions must be taken in order to produce any change in the epistemic risk; an excellent discussion of this issue with respect the Knowledge Gradient algorithm [21] for MAB problems is presented in [23]. We present a simple instance of the problem here.

Problem 2 *(Apple Tasting).* The apple tasting problem is characterized by $\mathcal{X} = \{1, 2\}$, $\mathcal{A} = \{1, 2\}$ and

$$W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad O = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

such that $R(x, a) = W(x, a)$ and $\Phi(x, a) = O(x, a)$. Suppose $x_t = 1$ with probability $\theta \in [0, 1]$, and $b_t(\theta) = \text{Beta}(\theta; \alpha_t, \beta_t)$. If the agent takes action 1, it receives a binary observation that it can use to update its belief distribution according to

$$\alpha_{t+1} = \alpha_t + y_{t+1}, \quad \beta_{t+1} = \beta_t + (1 - y_{t+1}).$$

However, if the agent takes action 2 it receives the same observation regardless of x_t, and so gains no information. There is a temporal feasibility criterion and no belief reward or discounting.

Example 3 *(Failure of Greedy-EBRM).* The ODOL policy for Problem 2 is to always take action 1 if $\mathbb{E}[\theta] \geq 0.5$, and to otherwise take action 2. Suppose that at time t, the belief state is characterized by $(\alpha_t, \beta_t) = (1, 3)$, so $\mathbb{E}[\theta] = 0.25$. If the agent takes action 1, the possible posterior belief distribution parameters are $(2, 3)$ or $(1, 4)$; in either case, the posterior ODOL policy will be unchanged as it will still hold that $\mathbb{E}[\theta] < 0.5$. Similarly, if the agent takes action 2, its belief state is not updated and so the ODOL policy

is unchanged. Therefore, as the ODOL policy in every possible posterior state is unchanged, the expected value of information from either action is zero; the agent will thus continue to select action 2 forever, gaining no new information and incurring linear regret.[9]

Accordingly, we present an asymptotic value of information approximation as a simple alteration to Greedy-EBRM that guarantees, over enough decisions, convergence on the optimal policy if the parameters θ are identifiable, as discussed in Theorem 1. This technique is discussed in the following subsection.

5.2.1. AsympGreedy-EBRM: leveraging the asymptotic value of information

The main limitation of Greedy-EBRM is that the expected value of information does not capture the value of information gained only from multiple actions. AsympGreedy-EBRM overcomes this issue by relying on a secondary, *asymptotic* measure of the value of information provided by each action.

First we introduce the Fisher information matrix corresponding to action a, denoted $\mathcal{I}_a(\theta)$, which is defined with respect to the log-likelihood function $\ell(\theta \mid a, y)$ as

$$\mathcal{I}_a(\theta_0) := \begin{bmatrix} \mathbb{E}_y\left[\left(\frac{\partial}{\partial\theta_1}\ell(\theta \mid a,y)\right)^2 \middle| \theta=\theta_0\right] & \cdots & \mathbb{E}_y\left[\left(\frac{\partial}{\partial\theta_1}\ell(\theta \mid a,y)\right)\left(\frac{\partial}{\partial\theta_D}\ell(\theta \mid a,y)\right) \middle| \theta=\theta_0\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}_y\left[\left(\frac{\partial}{\partial\theta_D}\ell(\theta \mid a,y)\right)\left(\frac{\partial}{\partial\theta_1}\ell(\theta \mid a,y)\right) \middle| \theta=\theta_0\right] & \cdots & \mathbb{E}_y\left[\left(\frac{\partial}{\partial\theta_D}\ell(\theta \mid a,y)\right)^2 \middle| \theta=\theta_0\right] \end{bmatrix}, \tag{5.21}$$

where $D := \dim\Theta$. Informally, the matrix $\mathcal{I}_a$ encodes how changes in the hidden parameters change the likelihood of the random variable y conditioned on an action a. There are cases where one or more hidden parameters do not play a role in the likelihood of the observation produced by a particular action; for example, in Problem 2, the hidden parameter does not affect the likelihood of the observation y for $a=2$. In such cases, $\mathcal{I}_a$ is singular. However, the *combined* Fisher information matrix,

$$\mathcal{I}_{\mathcal{A}}(\theta) := \sum_{a\in\mathcal{A}} \mathcal{I}_a(\theta), \tag{5.22}$$

is non-singular if and only if the parameters θ are *identifiable*, as defined in Subsection 4.1.2. In fact, the Bernstein-von Mises theorem [63] implies that, if the parameters are identifiable, the action set is finite with $K=|\mathcal{A}|$, and every action is taken at least $n\in\mathbb{N}_{>0}$ times, then the belief distribution converges to a multivariate Gaussian such that

$$\lim_{n\to\infty} b_{K\cdot n}(\theta) \to \mathcal{N}\left(\theta; \theta^\star, n^{-1}\mathcal{I}_{\mathcal{A}}(\theta^\star)^{-1}\right), \tag{5.23}$$

where $\theta^\star$ is the true value of the hidden parameters.

Importantly, the Fisher information matrices $\{\mathcal{I}_a\}_{a\in\mathcal{A}}$ provide an indication of which actions provide information, and which parameters they provide information about. Thus, even if the 1-step expected value of information for an action a is zero, a non-zero matrix $\mathcal{I}_a$ (even if singular) indicates that the action still provides information about one or more parameters. In fact, the Bernstein-von Mises theorem implies the *asymptotic* value of this information. This concept is formalized in Theorem 4.

Theorem 4. *Let Σ_t denote the covariance of the belief distribution b_t and let $\hat{\theta}_t := \arg\max_\theta b_t(\theta)$. Suppose $\mathcal{I}_{\mathcal{A}}(\hat{\theta})$ is non-singular, $\mathcal{A}=[K]$, $\Theta=\mathbb{R}^D$, and the regularity conditions of the Bernstein-von Mises theorem are satisfied [63].*[10] *As $t\to\infty$, if every action is taken at least $\lfloor tK^{-1}\rfloor$ times, then the epistemic risk of state $S_t=\{b_t,\xi_t\}$ decreases in expectation according to*

$$\text{EpistemicRisk}(S_t) - \mathbb{E}_{x|a}\left[\text{EpistemicRisk}(S_{t+1}^{x,a})\right] \xrightarrow[t\to\infty]{P} \text{EpistemicRisk}(\tilde{S}_t) - \text{EpistemicRisk}(\tilde{S}_{t+1}^a), \tag{5.24}$$

$$\tilde{b}_t = \mathcal{N}\left(\hat{\theta}_t, \Sigma_t\right), \tag{5.25}$$

$$\tilde{b}_{t+1}^a = \mathcal{N}\left(\hat{\theta}_t, \left(\Sigma_t^{-1} + \mathcal{I}_a(\hat{\theta}_t)\right)^{-1}\right), \tag{5.26}$$

where $\tilde{S}_t=\{\tilde{b}_t,\xi_t\}$ and $\tilde{S}_{t+1}^a=\{\tilde{S}_{t+1}^a, \delta(\xi_t,a)\}$.

Proof. As the number of actions tends to infinity, the relative weight of the prior goes to zero, and so $\hat{\theta}\to\theta^\star$ and $\Sigma_t\to n^{-1}\mathcal{I}_{\mathcal{A}}(\theta^\star)^{-1}$; then, the theorem holds by application of Bernstein-von Mises theorem. □

We define the quantity in the right side of Eq. (5.24) to be the **asymptotic value of information**,

$$\text{AsymptoticVoI}(S_t,a) := \text{EpistemicRisk}(\tilde{S}_t) - \text{EpistemicRisk}(\tilde{S}_{t+1}^a). \tag{5.27}$$

[9] As applied to Problem 2, Greedy-EBRM reduces to the Knowledge Gradient algorithm [21].

[10] Among these, the one most of note is that $\forall\theta\in\Theta$, we require the prior to satisfy $b_0(\theta)>0$.

The asymptotic value of information calculation assumes that the parameters θ have approximately converged on their true value, and thus measures how the epistemic risk will change as the *uncertainty* of the belief distribution decreases while the expectation of the parameters stays constant. As long as the parameters are identifiable and the epistemic risk is non-zero, the asymptotic value of information will be strictly positive, as seen in Corollary 3.

Corollary 3. *If the parameters θ are identifiable, then*

$$\text{EpistemicRisk}(S_t) > 0 \implies \exists a \in \mathcal{A} : \text{AsymptoticVoI}(S_t, a) > 0. \tag{5.28}$$

Proof. By contradiction; if $\forall a$, $\text{AsymptoticVoI}(S_t, a) = 0$, then epistemic risk has converged. By Theorem 1, the epistemic risk must then be zero, since the parameters are identifiable. □

Corollary 3 implies that, for any OLP in which the parameters are identifiable, the asymptotic value of information will guide the agent towards complete knowledge of the parameters even if no single action is sufficient to gain useful information. Furthermore, the asymptotic value of information is measured in units of reward (value), and parallels the expected value of information in modelling how the epistemic risk changes as actions are taken. We therefore introduce, in Algorithm 2, AsympGreedy-EBRM, a modification of Greedy-EBRM for problems where the expected value information from any single action may be zero, even if significantly more information could be gained from a bit more exploration.

Algorithm 2: AsympGreedy-EBRM.

Given: OLP Specification
Input: Belief State S_t
$\texttt{InitialRisk}(S_t) \leftarrow \texttt{AleatoricBayesRisk}(S_t) + \texttt{EpistemicRisk}(S_t)$
foreach $a \in \mathcal{A}$:
  $\texttt{ProcessRisk}(\pi_{t,a}; S_t) \leftarrow \texttt{ImmediateRisk}(S_t, a) - \gamma \cdot \max\{\texttt{ExpectedVoI}(S_t, a), \texttt{AsymptoticVoI}(S_t, a)\}$
end foreach
$a^\star_{t+1} \leftarrow \arg\min_{a \in \mathcal{A}} \left[\texttt{ProcessRisk}(\pi_{t,a}; S_t)\right]$ // Optimal action
$U_{t+1} \leftarrow \texttt{InitialRisk}(S_t) + \texttt{ProcessRisk}(\pi_{t,a^\star_{t+1}}; S_t)$ // Upper bound on AsympGreedy-EBRM Bayes' risk
return $a^\star_{t+1}, U_{t+1}$

Remark 1. The assumption $\Theta = \mathbb{R}^D$ in Theorem 4 does not always hold; in this case, while the belief distribution still converges locally to a multivariate Gaussian, it may not be possible to compute the epistemic risk for a Gaussian belief. In such cases, a heuristic solution is for the agent to use the original belief distribution family for the asymptotic belief distributions, with $\tilde{b}_t = b_t$ and the parameters of $\tilde{b}^a_{t+1}$ chosen to maximize its similarity to $\mathcal{N}(\hat{\theta}_t, (\Sigma_t^{-1} + \mathcal{I}_a(\hat{\theta}_t))^{-1})$, such as by moment matching.

5.2.2. Other approaches

An alternative approach to overcome the myopia of 1-step lookahead policies is to consider the larger set of m-step lookahead policies, as for a sufficiently large m the best m-step lookahead policy is equivalent to the optimal closed-loop policy $\pi^\star_b$.[11] However the computational complexity of finding the best m-step lookahead policy is exponential in m. This large set can be reduced to only linear lookahead policies, which consider taking the same action multiple times but not combinations of actions; this strategy is used by the KG* algorithm [22], however it has been shown to remain overly myopic [23].

Alternatively, one may include other non-myopic policies in the candidate policy set. The EBRM solution outperforms any of the individual policies in the candidate policy set, so adding more policies generally improves the EBRM solution. The challenge of this approach is finding policies for which the value can be computed efficiently. In any case, the epistemic and immediate risk functions can be used to lower bound the process risk of any policy, even a stochastic one, as seen in Proposition 4.

Proposition 4. *The process risk, in the belief state $S_t \in \mathcal{S}$, of any policy π is bounded below as*

$$\text{ProcessRisk}(\pi; S_t) \geq \mathbb{E}_{a \sim \pi(S_t)} \left[\text{ImmediateRisk}(S_t, a) - \mathbb{E}_{x|b_t}\left[\text{EpistemicRisk}(S^{x,a}_{t+1})\right]\right]. \tag{5.29}$$

Proof. The process risk of an action is lower bounded by the regret incurred immediately by action a and the minimum possible process risk of any policy in the posterior belief state. The process risk of the posterior belief state $S^{x,a}_{t+1}$ is bounded below by $-\text{EpistemicRisk}(S^{x,a}_{t+1})$ (refer to Lemma 5 in Appendix E). □

[11] This occurs by $m = h - 1$, where h is, as previously, the decision horizon.

Table 5
Multi-armed bandit problems (MAB) characteristics.

OLP Component	MAB Definition
Hidden Outcomes	$x_t \in \mathcal{X} \subseteq \mathbb{R}^K$
Actions	$a_t \in \mathcal{A} = \{1, \dots, K\}$
Observations	$y_t \in \mathcal{Y} \subseteq \mathbb{R}$
Observation Fn.	$\Phi(x_t, a_t) = x_t(a_t)$

Table 6
OLP component specifications for various stochastic bandits. p.s.d.: Positive semi-definite.

OLP Component		Gaussian Bandit	Bernoulli Bandit	Beta Bandit
θ	Hidden Parameters	$\mu \in \mathbb{R}^K$	$\mu \in [0,1]^K$	
	- Known Parameters	$\Sigma \in \mathbb{R}^{K\times K}$ (p.s.d.)	None	$\nu \in \mathbb{R}_{>0}$
f_θ	Process Model	$\mathcal{N}(\mu, \Sigma)$	Bernoulli(μ)	Beta$(\mu\nu, (1-\mu)\nu)$
b_t	Belief Distribution	$\mathcal{N}\left(\hat{\mu}_t, \hat{\Sigma}_t\right)$	Beta$\left(\alpha_t, \beta_t\right)$	
	- Belief Parameters	$\hat{\mu}_t \in \mathbb{R}^K$, $\hat{\Sigma}_t \in \mathbb{R}^{K\times K}$	$\alpha_t \in \mathbb{R}^K_{>0}$, $\beta_t \in \mathbb{R}^K_{>0}$	

Table 7
OLP specification for bandit optimization and best-arm identification.

OLP Component		Bandit Optimization	Best-Arm Identification
$R\left(x_t, a_t\right)$	Action Reward	$x_t(a_t)$	0
$R\left(b_t\right)$	Belief Reward	0	$\max_k \mathbb{E}_{\mu\|b_t}[\mu(k)] - \max_k \hat{\mu}_t(k)$ $\max_k P\left(\mu(k) > \mu(j)\ \forall j \neq k \mid b_t\right)$
γ	Discount Factor	$(0, 1]$	N/A
$\Omega(\xi)$	Feasibility Criterion	Temporal with horizon $T \in \mathbb{N}_{>0}$	

5.3. Anytime-EBRM: AsympGreedy-EBRM for unknown time horizons

Many classical OLPs are studied in the context of an *infinite* time horizon with no feasibility constraints (i.e. $\forall \xi, \Omega(\xi) = \mathcal{A}$), to make it easier to study the asymptotic behaviour of online learning algorithms. This is closely related to the case of an OLP that stops randomly, where algorithms are expected to minimize accumulated regret over *any* time horizon.

The AsympGreedy-EBRM algorithm cannot generally be implemented for an infinite time horizon without discounting, since it leads to infinite risk values. So, we introduce a modified *Anytime*-EBRM algorithm, which assumes that the remaining time is always equal to the current time (i.e., that it is always "halfway done"). Thus, at time t, it considers an artificial temporal feasibility criterion Ω_t such that,

$$\Omega_t(\xi) = \begin{cases} \mathcal{A}, & \xi < 2t + 1, \\ \varnothing, & \text{otherwise.} \end{cases} \tag{5.30}$$

This provides a means to study the asymptotic behaviour of the AsympGreedy-EBRM approach and how it performs without knowledge of the time horizon. We compare Anytime-EBRM to the other EBRM algorithms and baselines in Section 6.

6. Experimental results

In this section we explore the performance of EBRM algorithms across a range of common benchmark OLPs as well as some novel experiments that reflect real-world applications. Different OLPs are characterized by different specifications of the components listed in Table 1.

6.1. Bandit optimization and best-arm identification

Stochastic multi-armed bandits are a longstanding benchmark with which to evaluate online learning algorithms. The distinguishing features of stochastic MABs are presented in Table 5. We will begin by evaluating AsympGreedy-EBRM algorithms alongside popular online learning heuristics across a variety of stochastic bandits, described in Table 6, as well as both archetypal OLP goals of *bandit optimization* and *best-arm identification* as described in Table 7.

The defining characteristic of a bandit problem is "bandit feedback"; this refers to the observation function $\Phi\left(x_t, a_t\right) = x_t(a_t)$, which indicates that after each action the agent is provided with only the hidden outcome corresponding to that action. The bandit optimization (BDO) goal is to choose actions that maximize the sum of these observations; this is a classic exploration-exploitation problem, as the agent must explore different actions before it can begin to identify and exploit the action with the largest observation mean. The metric we use for performance in bandit optimization is (exogenous) average expected regret,

Table 8
Bayes-EBRM Algorithm Components.

EBRM Function	Bayes-EBRM
ImmediateRisk	$\max_k \hat{\mu}_t(k) - \hat{\mu}_t(a)$
EpistemicRisk	$(T-t)\,\mathbb{E}_{\mu\mid b_t}\left[\max_k \mu(k) - \max_k \hat{\mu}_t(k)\right]$
ExpectedVoI	$(T-t-1)\,\mathbb{E}_{x\mid b_t}\left[\max_k \hat{\mu}^{x,a}_{t+1}(k) - \max_k \hat{\mu}_t(k)\right]$
AsymptoticVoI	$(T-t-1)\left(\mathbb{E}_{\mu\mid \bar{b}_t}\left[\max_k \mu(k)\right] - \mathbb{E}_{\mu\mid \bar{b}^a_{t+1}}\left[\max_k \mu(k)\right]\right)$

Table 9
Epi-EBRM Algorithm Components.

EBRM Function	Epi-EBRM
ImmediateRisk	0
EpistemicRisk	$\mathbb{E}_{\mu\mid b_t}\left[\max_k \mu(k) - \max_k \hat{\mu}_t(k)\right]$
ExpectedVoI	$\mathbb{E}_{x\mid b_t}\left[\max_k \hat{\mu}^{x,a}_{t+1}(k) - \max_k \hat{\mu}_t(k)\right]$
AsymptoticVoI	$\mathbb{E}_{\mu\mid \bar{b}_t}\left[\max_k \mu(k)\right] - \mathbb{E}_{\mu\mid \bar{b}^a_{t+1}}\left[\max_k \mu(k)\right]$

Table 10
Anytime-EBRM Components for Multi-Armed Bandits (MABs).

EBRM Function	Anytime-EBRM
ImmediateRisk	$\max_k \hat{\mu}_t(k) - \hat{\mu}_t(a)$
EpistemicRisk	$(t+1)\,\mathbb{E}_{\mu\mid b_t}\left[\max_k \mu(k) - \max_k \hat{\mu}_t(k)\right]$
ExpectedVoI	$t \cdot \mathbb{E}_{x\mid b_t}\left[\max_k \hat{\mu}^{x,a}_{t+1}(k) - \max_k \hat{\mu}_t(k)\right]$
AsymptoticVoI	$t \cdot \left(\mathbb{E}_{\mu\mid \bar{b}_t}\left[\max_k \mu(k)\right] - \mathbb{E}_{\mu\mid \bar{b}^a_{t+1}}\left[\max_k \mu(k)\right]\right)$

$$t^{-1} \cdot \overline{\mathfrak{R}}(A_t; S_0, \theta) = \frac{1}{t}\sum_{\tau=1}^{t}\left(\max_{k\in[K]} \mathbb{E}_{x\mid\theta}\left[R(x,k) - R(x,a_t)\right]\right).$$

A lower average expected regret value is better, as it represents the difference between the average reward of the θ-optimal policy and the average reward of the online learning algorithm. For algorithms which are asymptotically optimal at BDO, the average expected regret should tend towards 0. In figures, we show the average *Bayes'* regret relative to the θ-optimal policy as a function of t; the average Bayes regret is the mean average expected regret across many independent trials, wherein each trial we sample $\theta \sim b_0$. In tables we show the mean *cumulative* expected regret of each algorithm at the end of the trial, which is, in expectation, the Bayes regret of that algorithm relative to the θ-optimal policy.

The goal of best-arm identification (BAI) is to identify the action with the largest observation mean, without regard to which actions are taken to do so. As seen in Table 7, success in this objective can be measured in two ways. The first is by $\max_k \hat{\mu}_t(k)$, which represents the mean action reward of the best action identified by time t [64]. This encourages identifying the action with the largest mean, and the penalty of identifying a different action is proportional to the difference in their respective action reward means. We will focus on the "Epistemic Uncertainty",

$$\mathbb{E}_{\mu\mid b_t}\left[\max_{k\in[K]} \mu(k)\right] - \max_{k\in[K]} \hat{\mu}_t(k),$$

which represents how much the observation mean of the true best action is expected to exceed the largest observed observation mean. As such, this metric converges to 0 as the agent becomes increasingly certain that no action has a larger observation mean than the best one it has identified.

Another popular metric of success in best-arm identification is the maximum probability for which any particular action is best [20]; this metric is intuitive and representative of many real-world problems, however it is difficult to optimize when the best $n > 1$ actions perform very similarly. In fact, this metric fluctuates asymptotically if the top n actions have equal observation means. This makes it poorly suited for use as a belief reward; we will, however, use it as an evaluation metric in some experiments. Since confidence values are often very close to 1, we display this metric in figures as "Log Uncertainty", computed as,

$$\ln\left(1 - \max_{k\in[K]} P(\mu(k) > \mu(j)\ \forall j \neq k \mid b_t)\right).$$

6.1.1. AsympGreedy-EBRM algorithms

All AsympGreedy-EBRM algorithms operate according to Algorithm 2; however, the immediate risk and value of information of each action depend on the OLP specification. The definitions of these functions for bandit optimization and best-arm identification objectives are presented in Tables 8 and 9. We show the corresponding functions for bandit optimization with Anytime-EBRM in Table 10. As a result of the choice of metric used for best-arm identification, epistemic risk for both tasks differs only by a scaling

that depends on the time horizon; as such, the expected value of information and asymptotic value of information are also similarly defined for the two problems. The key difference is that in bandit optimization problems there is an immediate risk to taking an action with a lower observation mean, which does not apply to best-arm identification problems.

To distinguish between AsympGreedy-EBRM algorithms for bandit optimization and best-arm identification, we label the former as *Bayes-EBRM* and the latter as *Epi-EBRM*. This naming convention is based on how AsympGreedy-EBRM for best-arm identification is equivalent to a bandit optimization algorithm that chooses actions to minimize *only* the posterior epistemic risk (i.e., maximize value of information). Conversely, Bayes-EBRM chooses actions to minimize the Bayes' risk by identifying the 1-step lookahead candidate policy with minimal process risk.

6.1.2. Baseline algorithms

In order to provide a baseline against which to compare the EBRM algorithms, we evaluated several reference online learning algorithms. These algorithms are presented below, in groups distinguished by their design goals and whether they take the time horizon (for temporal feasibility criteria) into account.

General online learning algorithms These algorithms, like the EBRM algorithms, can be applied to any OLPs which can be described by the BMDP formulation presented in Section 3.

- *Thompson Sampling* (TS; also known as Posterior Sampling) [3]: Samples a set of hidden parameters θ from the belief distribution b_t, and then chooses a_{t+1} to be whichever action is optimal (for bandit optimization) according to the sampled parameters. It can be easily extended to a wide range of OLPs by instead choosing the first action of the ODOL policy for the sampled parameters; we call this approach *Generalized TS*. Thompson sampling, first proposed in 1933, is remarkably effective given its simplicity [65,66]. While classical Thompson sampling requires no tuning, recent works have added hyperparameters that can be tuned for better performance on specific bandit problems [67].

Best-arm identification Best-arm identification is one of the earliest fields of online learning, and is often called "pure exploration". Oftentimes, bandit optimization algorithms are used with their hyperparameters tuned to encourage exploration. We consider three dedicated BAI algorithms.

- *KG Explore* [29,68]: Chooses the arm with the highest *knowledge gradient*. This "gradient" represents how much the largest posterior arm mean is expected to exceed the largest prior (current) arm mean, multiplied by the remaining time $(T - t)$. KG Explore is the Greedy-EBRM algorithm for BAI, as this gradient is exactly equal the expected value of information in Eq. (5.7); unlike Epi-EBRM, KG Explore does not consider the asymptotic value of information.
- *KG* Explore* [22]: Improves upon KG Explore by computing the knowledge gradient based on repeating each action $m = \{1, \dots, T - t\}$ times, and chooses the m for which the knowledge gradient weighted by m^{-1} is largest. This is a more effective heuristic, but expensive to compute for long time horizons [23].
- *Top-Two Expected Improvement* (TTEI) [20]: Operates by biasing sampling towards the two actions with the highest mean observations; the probability that the heuristic chooses the second-best action is controlled by a hyperparameter, but TTEI generally works well even when this hyperparameter is set to its default value. It has demonstrated superior performance when compared to various alternatives, including KG Explore [20].

Finite-time bandit optimization These algorithms are designed to minimize regret in stochastic bandit optimization problems, and take into account the finite time horizon T.

- *UCB-MOSS* [69]: Chooses the action (arm) with the highest observation mean, with an added bias proportional to the remaining time horizon $(T - t)$ and inversely proportional to the number of times that action has been used so far. This bias is designed to be minimax optimal in stochastic bandit optimization problems with binary rewards.
- *Knowledge Gradient* (KG) [21]: The same as KG Explore, but adds the observation mean to the gradient. KG is the Greedy-EBRM algorithm for BDO and, unlike Bayes-EBRM, does not consider the asymptotic value of information.
- *KG** [22]: The same as KG* Explore, but adds the observation mean to the gradient. It still requires $O(Kh)$ computations for each decision, making it expensive to compute for long time horizons [23].

Asymptotic bandit optimization These algorithms are designed to incur the minimum worst-case regret in stochastic bandit optimization problems with unknown time horizons (including infinite horizons). Thus they are generally expected to underperform finite-time bandit optimization algorithms when T is given.

- *Double Sampling* (DS) [66]: A recent improvement upon Thompson sampling that makes modifications for a more efficient exploration-exploitation trade-off. It performs similarly to TS when the probability of any particular action being optimal is low, but chooses the action with the highest observation mean when the probability of that action being optimal is high. DS uses a variable number of samples to make each decision, inversely proportional to its confidence in the best action.
- *UCB1* [13]: Perhaps the most popular heuristic for bandit optimization problems; as an upper confidence bound algorithm, it chooses the action with the highest observation mean, biased by an amount that is inversely proportional to the number of times

Table 11
Bayes Regret. 10-arm Bernoulli Bandit, 1000 Turns. Average over 2000 trials. [(1)]Results from [23].

Algorithm	Mean ± Std. Err.	10%	25%	50%	75%	90%	95%
Bayes-EBRM	**17.2** ± 0.3	6.9	9.0	13.2	20.1	30.7	40.6
UCB-MOSS	51.2 ± 0.2	41.3	44.8	49.8	55.3	63.0	68.4
KG[(1)]	51.0 ± 1.5	0.7	2.9	11.9	82.3	159.0	204.2
KG*[(1)]	18.4 ± 0.6	2.9	5.4	8.7	16.3	46.9	76.6
Anytime-EBRM	19.8 ± 0.4	3.6	8.6	15.9	24.7	37.9	51.4
IDS[(1)]	18.0 ± 0.4	3.6	7.4	13.3	22.5	35.6	51.9
V-IDS[(1)]	18.1 ± 0.4	5.2	8.1	13.5	22.3	36.5	48.8
DS	23.5 ± 0.4	9.4	12.7	18.7	29.2	43.7	52.0
TS	28.4 ± 0.3	13.1	17.5	25.1	35.3	47.9	56.1
Any-MOSS	53.9 ± 0.2	43.7	47.7	52.1	57.9	64.5	71.6
UCB1	133.6 ± 0.4	106.7	119.9	135.0	148.0	158.1	164.1

that action has been chosen in the past. This bias is often multiplied by a scalar hyperparameter, which can be tuned for better performance on specific OLPs; a larger scalar encourages exploration over exploitation.

- *Information Directed Sampling* (IDS/V-IDS) [23]: Chooses actions by minimizing the squared regret that the agent incurs per bit of information gained by the agent about the optimal action. This works well across a variety of bandit optimization problems, although it can be computationally expensive to compute the number of bits of information expected to gained by some action; variance-based IDS (V-IDS) instead uses a lower bound estimate of this quantity that is more efficient to compute.
- Any-MOSS (MOSS-anytime) [70]: A variant of UCB-MOSS, this heuristic is designed to be minimax optimal for asymptotic bandit optimization problems.

6.1.3. Experimental methodology

All trials used a temporal feasibility criterion with finite horizon $T \in \mathbb{N}_{>0}$. Each trial is characterized by a specific set of hidden parameters θ and hidden outcomes X. In some experiments, the hidden parameters are fixed, while in others they were randomly generated from the prior belief distribution b_0. Regardless, T hidden outcomes $x_1, \ldots, x_T$ were randomly sampled from the stochastic process f_θ. In each round $t = 1, \ldots, T$ of each trial, each algorithm chose one action a_t, and received the observation $\Phi(x_t, a_t)$. The same hidden outcome x_t was used to generate the observations and action rewards for all algorithms. The number of trials varied between experiments, as some experiments required more in order to achieve sufficiently low standard errors. All figures show 95% confidence intervals of the mean.

6.1.4. Results and discussion

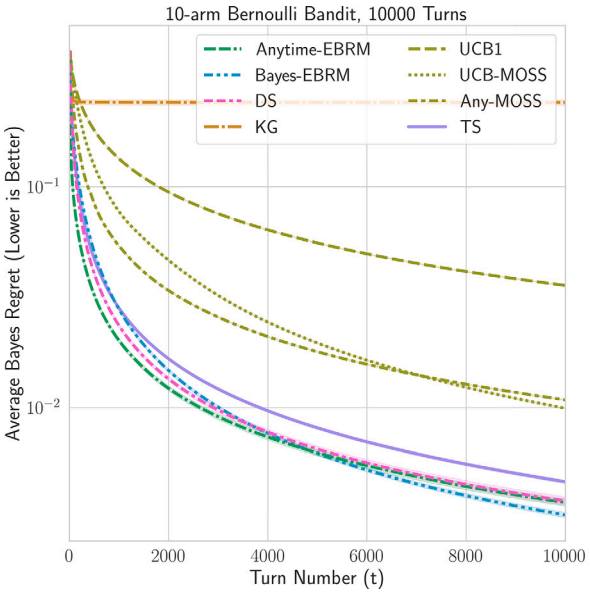
We evaluated the performance of the AsympGreedy-EBRM algorithms in a variety of numerical experiments, many of which are baselines established by or based on prior works.

One of the most comprehensive comtemporary comparisons of online learning algorithms for bandit optimization was made by Russo et al. [23]. Table 11 replicates their experiment consisting of a 10-arm Bernoulli bandit with hidden parameters sampled according to $\mu \sim b_0$, with $b_0 = \text{Beta}(1, 1)$, and a time horizon of $T = 1000$. The columns show the Bayes regret computed from 2000 trials, followed by various percentiles of Bayes regret. The algorithms which make use of the time horizon T are listed first, with a double line separating them from the algorithms designed for asymptotic performance.

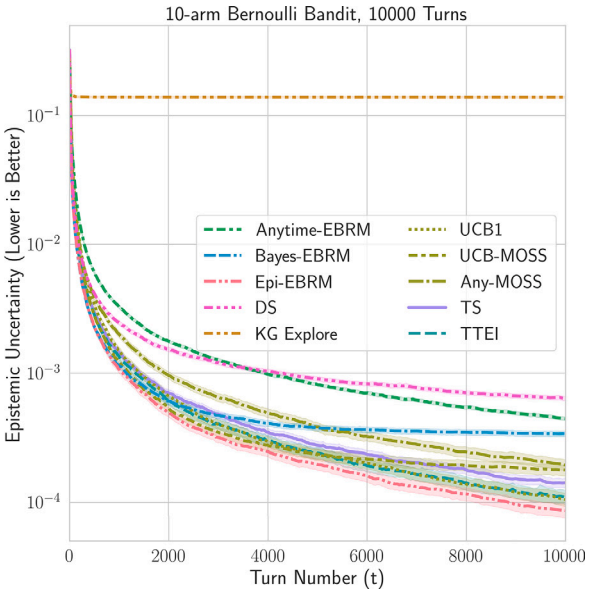
As Bernoulli bandits with beta priors are among the simplest and most well studied OLPs, it is unsurprising that Bayes-EBRM makes only a slight improvement upon the previous state-of-the-art for bandit optimization in this problem. More interestingly, it has more consistent performance than the other leading heuristics, including KG*, IDS, and V-IDS; this is particularly noticeable in the 90th and 95th percentile results. Anytime-EBRM performs comparably to IDS/V-IDS, demonstrating that the value of information computed by the AsympGreedy-EBRM approach is similarly useful to the IDS value of information function, and the performance improvement of Bayes-EBRM is mostly driven by taking the time horizon into account. The KG heuristic has been noted to perform particularly poorly for bandit problems with binary rewards [23].

In a similar Bernoulli bandit experiment with a time horizon of $T = 10^4$, shown in Fig. 5, we find that Bayes-EBRM continues to outperform the other algorithms in bandit optimization. Similarly, for best-arm identification, Epi-EBRM achieves lower epistemic uncertainty than any baseline algorithm. This figure demonstrates aspects of the exploration-exploitation trade-off; in general, algorithms which focus on minimizing Bayes regret will explore less and have higher epistemic uncertainty [64]. However, for a finite-time horizon there is an optimal amount of exploration required to do well at bandit optimization. As seen in Fig. 5, Bayes-EBRM spends approximately the first 4000 turns reducing its epistemic uncertainty, and performs worse on the bandit optimization objective than Anytime-EBRM or DS until this point; then, it surpasses them both by exploiting its more complete knowledge of the best action. Most of the other algorithms over-explore, with the UCB-based algorithms having orders of magnitude larger Bayes regret.

We next consider a 10-arm Gaussian bandit problem with time horizon $T = 1000$, and b_0 a zero-mean uncorrelated multivariate Gaussian distribution with unit variances. The results are presented in Table 12 and in Fig. 6. This is another archetypal OLP, and



(a) Anytime-EBRM and DS over-exploit early in each trial, while Bayes-EBRM outperforms them at BDO by reaching the optimal level of epistemic uncertainty before exploiting the best arm.



(b) The stark contrast in BAI performance between Epi-EBRM and KG Explore highlights the importance of using the asymptotic value of information to overcome the myopia of one-step lookahead strategies.

Fig. 5. Performance of algorithms in BDO and BAI objectives while making decisions in a 10-arm Bernoulli bandit problem over 10^4 turns; lower values indicate better performance. Shaded regions indicate 95% confidence intervals over 5000 trials.

Table 12
Bayes Regret. 10-arm Gaussian Bandit, 1000 Turns. Average over 5000 trials. [(1)]Results from [23].

Algorithm	Mean ± Std. Err.	10%	25%	50%	75%	90%	95%
Bayes-EBRM	**49.6** ± 0.6	28.4	34.0	42.1	53.8	69.1	85.5
UCB-MOSS	51.2 ± 0.6	29.6	34.0	40.2	50.2	71.0	107.3
KG	63.0 ± 1.7	16.3	20.3	25.8	35.7	141.9	303.5
KG*	52.4 ± 1.3	18.6	23.3	29.4	39.4	79.0	196.3
Anytime-EBRM	57.2 ± 0.8	24.7	30.9	41.6	60.5	98.8	145.6
V-IDS[(1)]	58.4 ± 1.7	24.0	30.3	39.2	56.3	104.6	158.1
DS	66.5 ± 0.8	31.2	40.6	54.8	75.5	105.9	136.2
TS	69.5 ± 0.5	39.4	48.9	61.7	81.4	106.2	125.5
Any-MOSS	58.0 ± 0.8	30.8	35.1	41.3	53.8	96.4	152.2
UCB1	94.4 ± 0.4	64.3	74.5	90.3	109.5	129.7	143.7

Bayes-EBRM again demonstrates leading performance by a small margin.[12] We also experiment with varying the time horizon as in [23]. Each of the values in Table 13 represents the Bayes regret estimated over 2000 trials, so the table represents 20000 trials for each of 9 algorithms. We observe that Bayes-EBRM is the only algorithm which performs consistently well across all time horizons; KG and V-IDS each perform similarly well to Bayes-EBRM for short time horizons, but begin to perform much worse from $T = 500$. Eventually, UCB-MOSS manages to outperform Bayes-EBRM, despite poor performance over short time horizons.

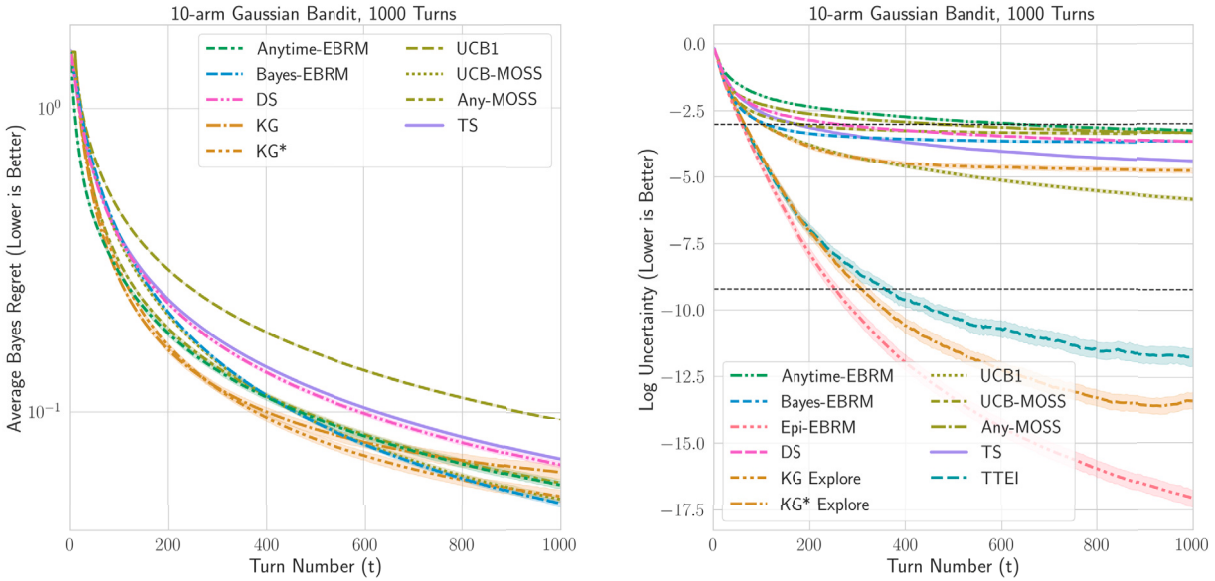
In best-arm identification on the 10-arm Gaussian bandit problem, we find that Epi-EBRM significantly outperforms the previous state-of-the-art beyond the first 100 turns. TTEI performs well up to this point, but its focus on the top-two arm candidates causes it to neglect the other 8 arms which may each still have a non-negligible chance of being the largest.

In Appendix B, we present the results of additional experiments involving a 2-arm Bernoulli bandit, a 5-arm Gaussian Bandit, and a 10-arm Beta bandit with time horizons of $T = 200$, $T = 100$, and $T = 10^4$, respectively (see Figs. B.12–B.14). These results are qualitatively similar to the previous results, demonstrating that the AsympGreedy-EBRM algorithms match or surpass state-of-the-art performance in bandit optimization and best-arm identification regardless of the type of bandit, the number of arms, or the time horizon.

[12] IDS is missing from this comparison as the authors note that it is too computationally expensive, and V-IDS achieves comparable performance [23].

Table 13
Bayes Regret. 10-arm Gaussian Bandit, 10 time horizons. Average over 2000 trials. [(1)]Results from [23].

Algorithm	Time Horizon T									
	10	25	50	75	100	250	500	750	1000	2000
Bayes-EBRM	**9.1**	**14.8**	**19.7**	**23.1**	**24.9**	**33.5**	**40.8**	**45.4**	**47.6**	58.1
UCB-MOSS	15.3	21.2	26.2	29.2	31.1	38.0	43.2	46.0	49.6	**57.2**
KG	9.1	15.0	19.8	23.6	25.7	36.0	42.1	53.6	61.5	83.7
KG*	9.1	14.9	19.8	23.6	25.7	34.9	40.3	47.8	51.6	61.3
Anytime-EBRM	9.3	15.3	21.4	25.9	29.3	39.9	45.9	51.7	54.6	67.8
V-IDS[(1)]	9.8	16.1	21.1	24.5	27.3	36.7	48.2	52.8	58.3	68.4
DS	12.0	21.6	29.2	34.1	37.0	47.9	54.6	61.5	65.0	76.6
TS	12.1	21.8	29.8	34.7	37.9	49.4	58.2	64.3	67.8	80.6
Any-MOSS	15.3	20.9	25.5	28.9	30.9	40.1	47.0	51.7	56.3	67.3
UCB1	15.3	24.1	34.3	40.9	45.9	63.0	77.5	86.5	93.4	112.5



(a) KG and KG* initially over-exploit the BDO objective, resulting in lower initial Bayes regret but ultimately performing worse than Bayes-EBRM and UCB-MOSS, which wait to be more confident in the best arm before exploiting it.

(b) The upper dashed line represents 95% confidence in having found the best arm, while the lower dashed line represents 99.99% confidence. Epi-EBRM achieves vastly higher confidence in the best arm than TTEI or KG* Explore over long horizons.

Fig. 6. Performance of algorithms in BDO and BAI objectives while making decisions in a 10-arm Gaussian bandit problem over 1000 turns. Shaded regions indicate 95% confidence intervals over 2000 trials.

6.1.5. Hyperparameter tuning

As noted in previous sections, some online learning heuristics have hyperparameters that can be tuned in order to achieve better performance on specific OLPs. NoTeS is a tuning algorithm designed specifically to optimize the hyperparameters of online learning algorithms in order to minimize Bayes regret (risk) [67]. It is an iterative algorithm, which reports the best hyperparameters found by the time it reaches a user-defined tuning budget (number of iterations). It outperformed various baseline algorithms in being able to find the lowest Bayes regret tuning in the smallest tuning budget [67].

Table 14 presents the average Bayes regret of UCB1 and TS for 2- and 10-arm Bernoulli bandit problems with $T = 200$ and $T = 10^4$ respectively, alongside the results for Bayes-EBRM, Anytime-EBRM, DS, and UCB-MOSS. Bayes-EBRM and Anytime-EBRM outperform UCB1 and TS in both experiments, even when the latter are tuned over 1000 iterations. As usual, Bayes-EBRM demonstrates better performance than Anytime-EBRM by taking the time horizon into account.

6.1.6. Computational decision complexity comparison

Most of the baseline algorithms are designed to be fast, and use $O(K)$ computations to compute some simple heuristic; often, these require taking a sample from b_t, or computing functions of its mean or covariance. The exceptions are KG*, which use $O(KT)$ computations, and DS, which uses a tunable number of samples from b_t for each decision. The EBRM algorithms require sampling to compute the expected value of information function for each action; in bandit problems, each action only provides information about

Table 14
Bayes Regret, Bernoulli Bandits. Values averaged over 10^4 trials. *Results from [67].

Algorithm	2 Arms, 200 Turns, $\mu = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}$					10 Arms, 10^4 Turns, $\mu(k) \sim \text{Beta}(1,1)$				
	Tuning Budget					Tuning Budget				
	Initial	50*	200*	1000*	Max*	Initial	50*	200*	1000*	Max*
UCB1	10.2	5.3	4.7	4.4	4.2	357.7	63.6	52.5	49.2	47.2
TS	5.5	5.1	4.8	4.6	4.3	46.3	74.9	42.9	36.4	33.8
Bayes-EBRM	**3.8**	—	—	—	—	**32.8**	—	—	—	—
Anytime-EBRM	4.2	—	—	—	—	37.2	—	—	—	—
DS	5.2	—	—	—	—	38.4	—	—	—	—
UCB-MOSS	7.4	—	—	—	—	99.4	—	—	—	—

Table 15
Task-EBRM Algorithm Components.

EBRM Component	Task-EBRM
$\text{ImmediateRisk}(B_t, a)$	$\max_k \hat{\mu}_t(k) - \hat{\mu}_t(a)$
$\text{EpistemicRisk}(B_t)$	$(N+T-t)\,\mathbb{E}_{\mu \mid b_t}\left[\max_k \mu(k) - \max_k \hat{\mu}_t(k)\right]$
$\text{ExpectedVoI}(B_t, a)$	$(N+T-t-1)\,\mathbb{E}_{x \mid b_t}\left[\max_k \hat{\mu}_{t+1}^{x,a}(k) - \max_k \hat{\mu}_t(k)\right]$
$\text{AsymptoticVoI}(B_t, a)$	$(N+T-t-1)\,\mathbb{E}_{\mu \mid \tilde{b}_{t+1}^a}\left[\max_k \mu(k)\right] - \mathbb{E}_{\mu \mid \tilde{b}_t}\left[\max_k \mu(k)\right]$

one hidden parameter and constructing an ODOL policy from this information has cost $O(1)$, so the cost of generating a sample is $O(1)$. Thus, assuming a fixed number of samples, EBRM bandit decisions have a time complexity of $O(K)$. The actual decision time, however, depends heavily on the cost of each sample. In practice, for bandit problems we can accurately compute the expected value of information by using integration by quadrature at a relatively small, fixed number of sample locations in Θ.

In Fig. 7, we explore how the choice of the fixed number of samples used to estimate the expected value of information for each action affects the performance and running time of Bayes-EBRM. In general, we find that BDO performance is largely insensitive to the number of samples used. Furthermore, unlike KG*, the time for an EBRM algorithm to make a bandit decision is independent of the time horizon T.

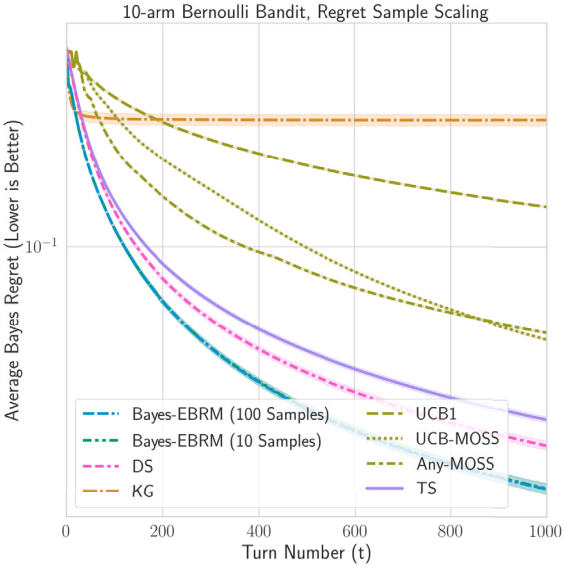
6.2. Combined belief- and action- rewards

As noted in Section 1, and highlighted in Fig. 1, online learning algorithms must balance between multiple competing OLP objectives. For all of the baseline algorithms discussed thus far, this balance is an implicit, fixed parameter of the algorithm. However, EBRM-based approaches are uniquely capable of achieving OLP objectives that can be expressed as the sum of action-based and belief-based rewards which meet the conditions described in Section 3. We demonstrate this by revisiting the example in Fig. 1.

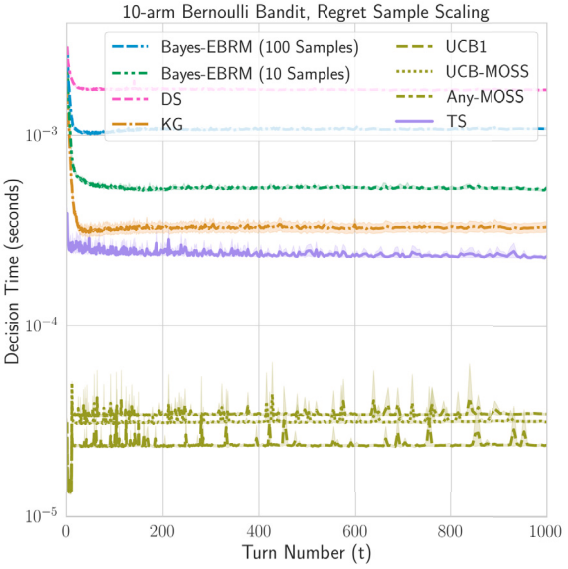
Suppose there is a clinical trial to be conducted with a study group of 423 patients. The four treatments to be tested have *a priori* unknown success rates, $\mu_1 = \mu_2 = \mu_3 = 0.3$ and $\mu_4 = 0.5$. After the trial is completed, the best performing treatment will be administered to N patients *outside of the trial* who are likewise awaiting treatment. The goal is to have as many successful outcomes as possible across all $423 + N$ patients; the *Task Regret* represents the number of unsuccessful treatment outcomes. This example is adapted from the multi-arm trial setting considered in [2].

The AsympGreedy-EBRM algorithm for this problem is trivial to design, and presented in Table 15 as *Task-EBRM*. Note that in this example, epistemic uncertainty and Bayes regret are equivalent to the x- and y-axis labels in Fig. 1, respectively. This algorithm is parameterized by N, and directly attempts to maximize the total number of successful patient outcomes. For $N = 0$ Task-EBRM is equivalent to Bayes-EBRM, while as $N \to \infty$ its behaviour approaches that of Epi-EBRM. We evaluate this algorithm for various N alongside the baseline algorithms, and present the results in Fig. 8.

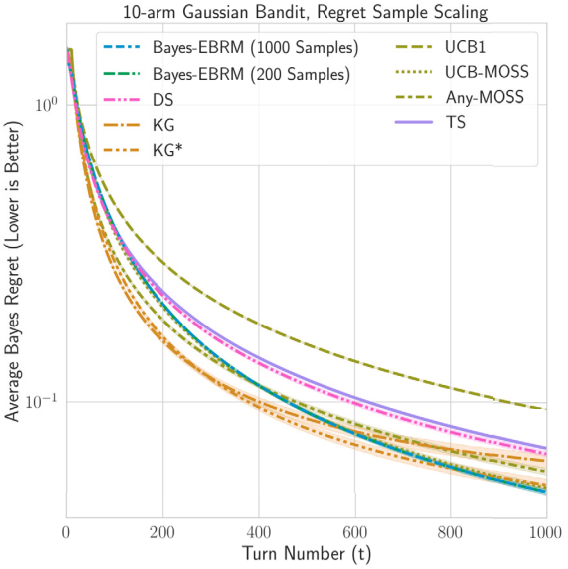
As expected, the relative performance of most algorithms varies greatly for different values of N, while Task-EBRM consistently achieves the least task regret. For example, TTEI performs relatively poorly even when $N = 2500$, but is the best of the baseline algorithms as $N \to \infty$. Similarly, DS and Anytime-EBRM perform very well for $N \leq 500$, but perform poorly for $N \geq 2500$. UCB-MOSS and Any-MOSS are the most versatile of the baseline algorithms, but choosing either of these still results in up to three times as many unsuccessful patient outcomes as $N \to \infty$. In fact, by evaluating Task-EBRM at additional values of N we find that AsympGreedy-EBRM *dominates* the baseline algorithms, in that for any of the baseline algorithms, there is a value of $N \geq 0$ such that Task-EBRM simultaneously achieves *both* lower epistemic uncertainty *and* less Bayes regret. This is presented in Fig. 1, where the x-axis shows epistemic uncertainty scaled by 10^3. These results support that AsympGreedy-EBRM algorithms are the superior solutions for real-world OLPs.



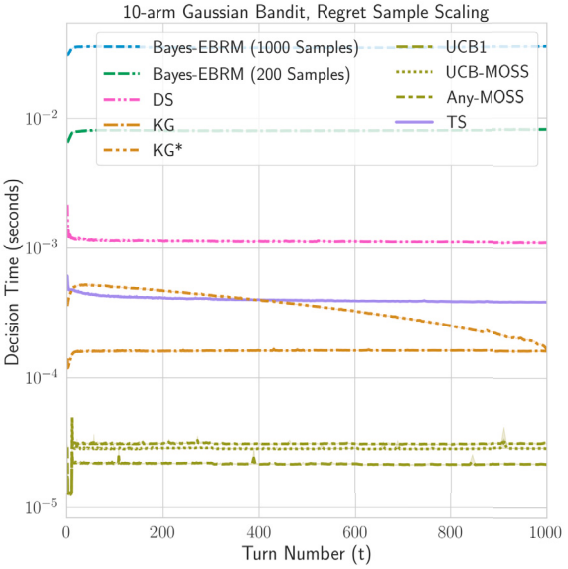
(a) The quality of Bayes-EBRM decisions with 100 samples is indistinguishable from its performance with only 10 in this Bernoulli bandit problem, resulting in a single blue-green curve.



(b) The Bayes-EBRM methods for this 10-arm Bernoulli bandit problem use comparable amounts of computation to the non-UCB baselines.



(c) The quality of Bayes-EBRM decisions with 1000 samples is indistinguishable from its performance with 200 in this Gaussian bandit problem, producing a single solid blue-green curve.



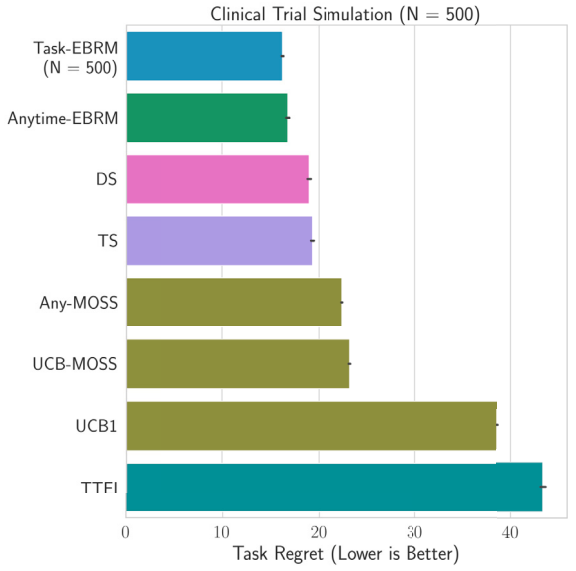
(d) Bayes-EBRM decisions with 200 samples require a fifth as much computation time as those with 1000 in this problem. We also see the proportionality of the KG* decision time to $T - t$.

Fig. 7. We re-evaluate the performance of Bayes-EBRM for the Bernoulli and Gaussian bandit experiments, from Figs. 5 and 6 respectively, each with 10-arms and $T = 10^3$, with different numbers of samples used to estimate the expected value of information. We further report the average time for each algorithm to make a decision as a function of the turn number.

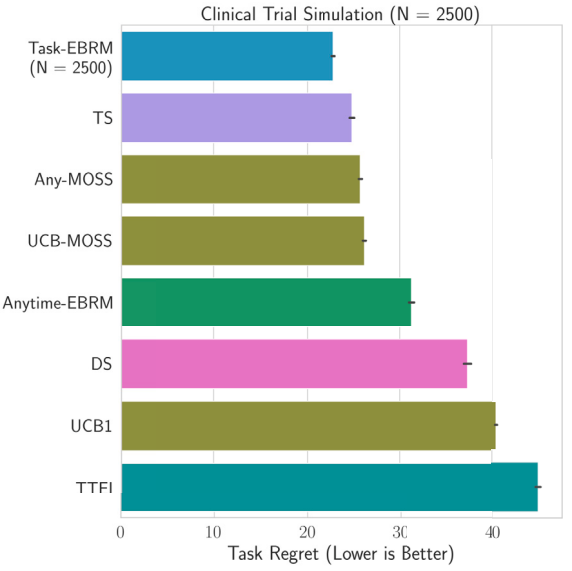
6.3. Partial monitoring and dynamic pricing

Partial monitoring is a more general class of online learning than bandit optimization, and enables the study of more interesting reward and observation models. Stochastic partial monitoring is a specific case of the OLP structure presented in Section 3, characterized by finite action and outcome sets with an "observation matrix" H and a "loss matrix" L, which define the observation model and action rewards, respectively. The full specification, less problem-specific components, is given in Table 16.

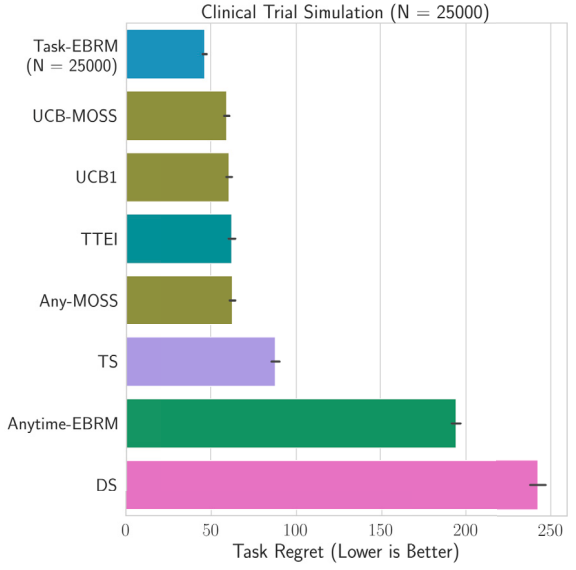
Dynamic Pricing is a prototypical partial monitoring problem which has complexities beyond that of bandit optimization [7]. In it, each of the K actions represent a "sales price", and each of the $M = K$ hidden outcomes represents a customer's "willingness to



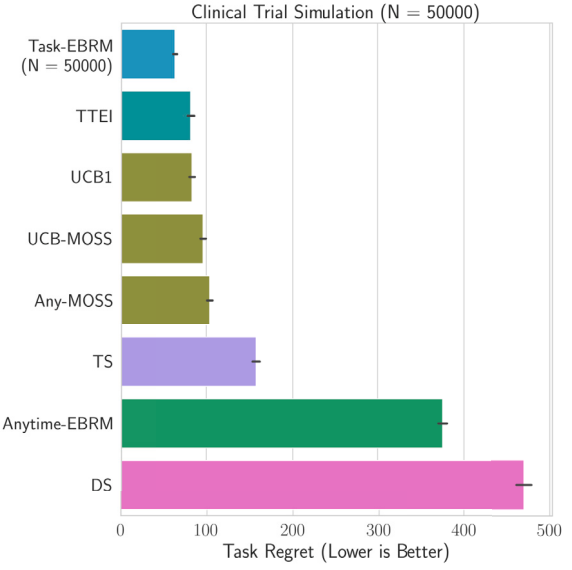
(a) At $N = 500$, we see similar relative performance as for the BDO objective in the Bernoulli bandit experiment from Figure 5, as Task-EBRM approximates Bayes-EBRM.



(b) At $N = 2500$, the lack of exploration performed by DS and Anytime-EBRM relative to Task-EBRM results in more unsuccessful outcomes among post-trial patients.



(c) At $N = 25000$, the large amount of exploration preferred by the UCB-based algorithms and TTEI begins to result in relatively strong performance.



(d) By $N = 50000$, we see similar relative performance as for the BAI objective in the Bernoulli bandit experiment from Figure 5, as Task-EBRM approximates Epi-EBRM.

Fig. 8. Performance of the AsympGreedy-EBRM and baseline algorithms on the task described in 6.2, for various sizes of global patient population N. Task regret indicates the number of failed treatments among $423 + N$ patients.

pay". If the sales price is higher than a customer's willingness to pay, there is no sale and the agent incurs some fixed loss $c \in \mathbb{R}_{>0}$. Otherwise, a sale is made, and the agent incurs a loss based on how much lower the price was than the customer's willingness to pay. Importantly, the customer's willingness to pay is never directly revealed. The dynamic pricing specification of H and L is thus, as given in [71] and where "y" denotes a sale and "n" denotes no sale,

$$H = \begin{bmatrix} \mathrm{y} & \cdots & \cdots & \mathrm{y} \\ \mathrm{n} & \mathrm{y} & \cdots & \mathrm{y} \\ \vdots & \ddots & \ddots & \vdots \\ \mathrm{n} & \cdots & \mathrm{n} & \mathrm{y} \end{bmatrix}, \qquad L = \begin{bmatrix} 0 & 1 & \cdots & K-1 \\ c & 0 & \cdots & K-2 \\ \vdots & \ddots & \ddots & \vdots \\ c & \cdots & c & 0 \end{bmatrix}. \tag{6.1}$$

Table 16
Partial Monitoring Problem Characteristics.

	OLP Component	Partial Monitoring Definition
$\mathcal{X}$	Hidden Outcomes	$\mathcal{X} = \{1, \dots, M\}$
$\mathcal{A}$	Actions	$\mathcal{A} = \{1, \dots, K\}$
Θ	Hidden Parameters	$\Theta = \Delta^M := \{\theta \in \mathbb{R}^M_{\geq 0} : \|\theta\|_1 = 1\}$
f_θ	Process Model	$\Pr(x \mid \theta) = \theta(x)$
Φ	Observation Model	$\Phi(x, a) = H(x, a)$
	- Known Parameters	$H \in \mathcal{Y}^{M \times K}$
$R(x, a)$	Action Reward	$R(x, a) = -L(x, a)$
	- Known Parameters	$L \in \mathbb{R}^{M \times K}$
$R(b)$	Belief Reward	$R(b) = 0 \ \forall b$
$\Omega(\xi)$	Feasibility Criterion	Temporal with horizon $T \in \mathbb{N}_{>0}$

Dynamic pricing is more difficult than bandit optimization because the problem is not *locally observable* [6]. In short, this means that the relative expected reward (equivalently, loss) of some pairs of actions a_1, a_2 cannot be determined without taking a third action a_3. This presents an issue if a_1, a_2 are candidates for the best action, while a_3 has much lower expected reward. In our formulation, this is a case where the (expected) value of information of various actions is highly correlated with their immediate risk, as the actions with low risk provide little or no information. A more thorough analysis of the issue, and its related implications to the difficulty of partial monitoring problems, is given by [6].

Letting $\mu = \langle -L, \theta \rangle$ and $\hat{\mu}_t = \mathbb{E}_{\theta|b_t}[\mu]$, the immediate risk, epistemic risk, EVoI and AVoI of the AsympGreedy algorithm for dynamic pricing match those of the Bayes-EBRM algorithm, as given in Table 8. The ODOL policy in any belief state is to choose the action with the highest expected reward, $\arg\max_{k \in [K]} \hat{\mu}_t(k)$.

6.3.1. Experimental methodology

To generate each dynamic pricing problem instance, we generated a hidden parameter vector from a uniform distribution over the K-dimensional probability simplex Θ. The initial belief distribution, however, was taken to be a normal distribution with mean $\mathbb{E}[\theta]$ and an identity covariance matrix I_K. While this prior is unbiased, it is not the actual distribution from which hidden parameters are drawn. We generated 5000 problem instances for each experiment, and set the fixed cost for no sale to $c = 2$.

We use the BPM (Bayes-update Partial Monitoring) approach presented in [71] to generate a Gaussian posterior belief distribution following each new action-observation pair. Samples taken from these belief distributions may lie outside of the probability simplex Θ; in such cases, as in [71], we project these samples to the nearest point in Θ. We compare against the BPM-TS algorithm presented in [71], which is the generalized Thompson sampling strategy (discussed in Subsection 6.1.2) using the same BPM update rule.

6.3.2. Results and discussion

Figs. 9 and 10 show average Bayes regret and epistemic uncertainty for each of the two experiments, with time horizons $T = 10^3$ and $T = 10^4$, respectively. The results parallel those of the bandit experiments; the Bayes- and Anytime-EBRM algorithms outperform the baseline by effectively managing the trade-off between exploration and exploitation, balancing immediate risk with the expected and asymptotic values of information. Knowledge of the time horizon enables Bayes-EBRM to outperform Anytime-EBRM, but the gap shrinks over longer horizons. The Greedy-EBRM algorithm, which ignores the asymptotic value of information, fails to sufficiently explore and incurs linear regret.

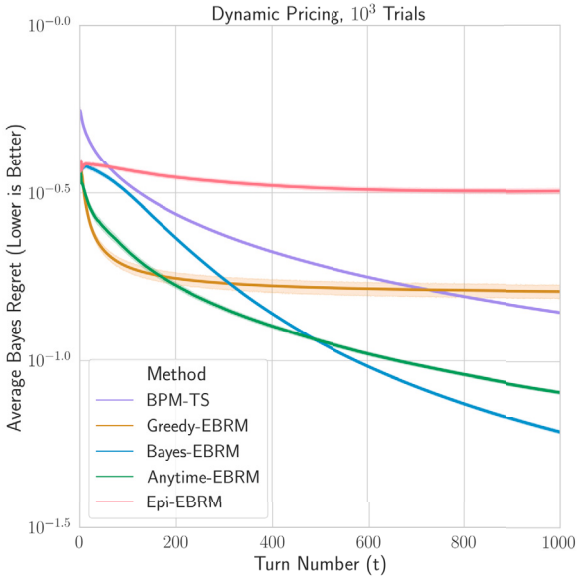
6.3.3. Impact of prior misspecification

As noted previously, the normal prior used in the dynamic pricing experiments does not match the true uniform distribution from which the hidden parameters are drawn. To determine the sensitivity of the algorithms to the choice of prior, we explored scaling the prior covariance matrix to sI_K, for some $s > 0$.

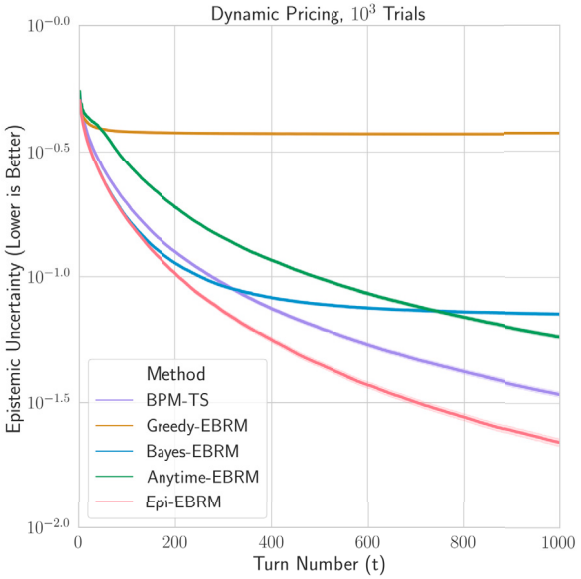
The results for $s \in \{2, 10, 0.5, 0.1\}$ are presented in Fig. 11. In general, the most algorithms show little performance change over the wider priors $s \in \{2, 10, 0.5\}$, but performance degrades when $s = 0.1$. This is to be expected, as a wide prior can be compensated for by taking a few actions with high value of information, and any corresponding immediate risk incurred has little impact over longer time horizons. Conversely, a narrow prior can cause algorithms to underestimate the value of exploration and, in the case of EBRM, encourages following a sub-optimal ODOL policy.

7. Conclusions & future work

The belief-space Markov Decision Process model in Section 3 provides a standard way to model online learning problems with combined action- and belief-based rewards, action-based costs and various feasibility criterion. The notion of measuring risk with respect to X- and θ-optimal policies as well as ODOL policies presents new ways to understand online learning problems and analyze policies through aleatoric, epistemic, and process risks. The EBRM-approach of searching for policies with minimal process risk has been shown to be feasible and highly effective at solving bandit problems, with AsympGreedy-EBRM algorithms matching or

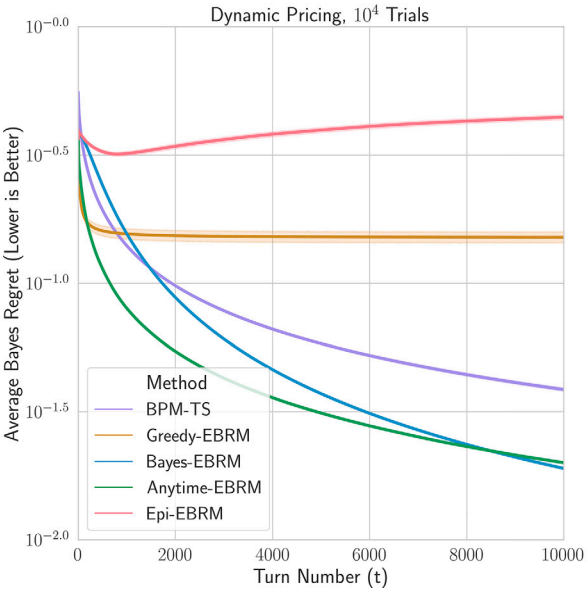


(a) The Anytime-EBRM algorithm excels at minimizing Bayes regret at the beginning of each experiment, but insufficient exploration causes it to fall behind Bayes-EBRM for $t > 500$.

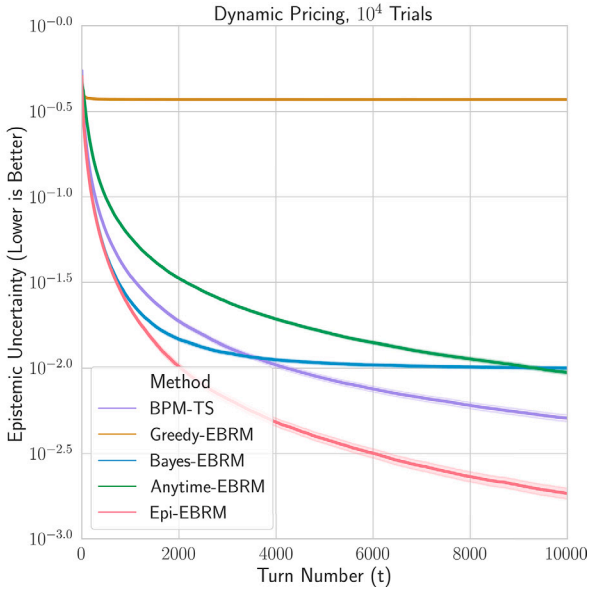


(b) By reasoning explicitly about the value of information, Epi-EBRM reduces the agent's epistemic uncertainty significantly faster than BPM-TS.

Fig. 9. Results of the dynamic pricing experiment with $T = 10^3$. The Greedy-EBRM algorithm fails to sufficiently explore, suffering linear regret. The Anytime-EBRM algorithm, like BPM-TS, lacks knowledge of the time horizon but achieves superior performance by explicitly reasoning about immediate risks and the expected and asymptotic values of information. By further taking the time horizon into account, Bayes-EBRM significantly outperforms the other methods.

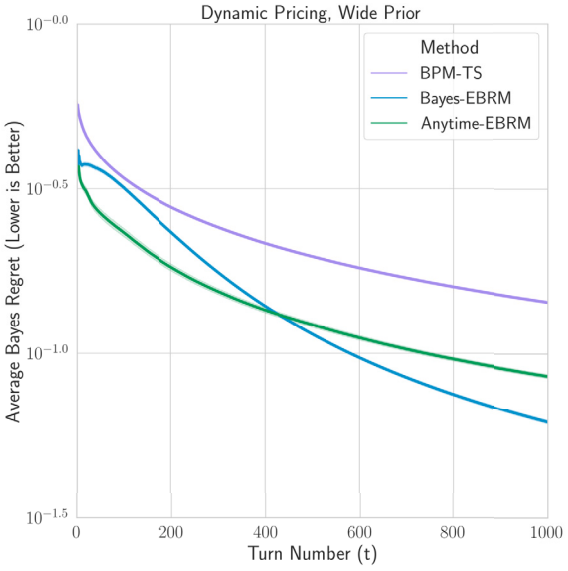


(a) While broadly similar to the results in Figure 9(a), the difference in average Bayes regret between the Anytime-EBRM and Bayes-EBRM algorithms is reduced over the extended time horizon.

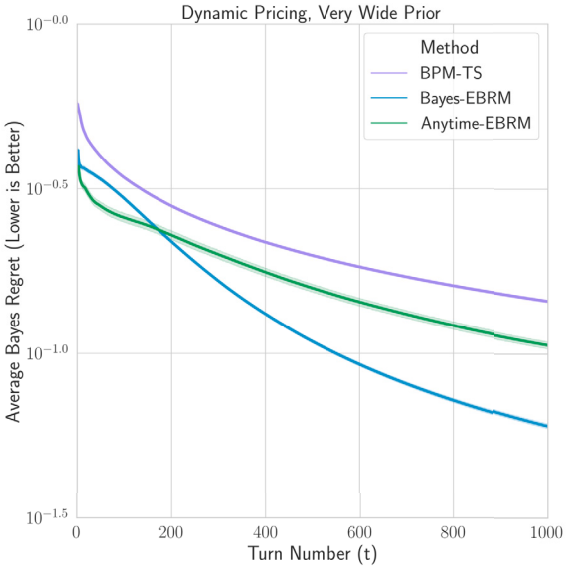


(b) Epi-EBRM continues to most effectively reduce epistemic uncertainty, while the behaviour of BPM-TS gradually shifts away from exploration over time.

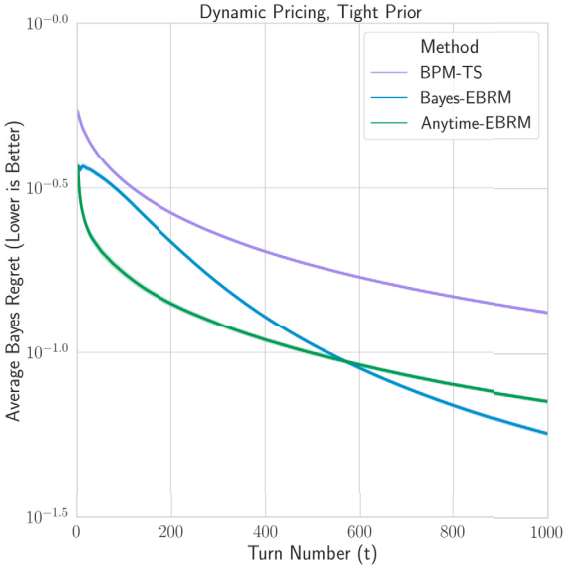
Fig. 10. Results of the dynamic pricing experiment with $T = 10^4$.



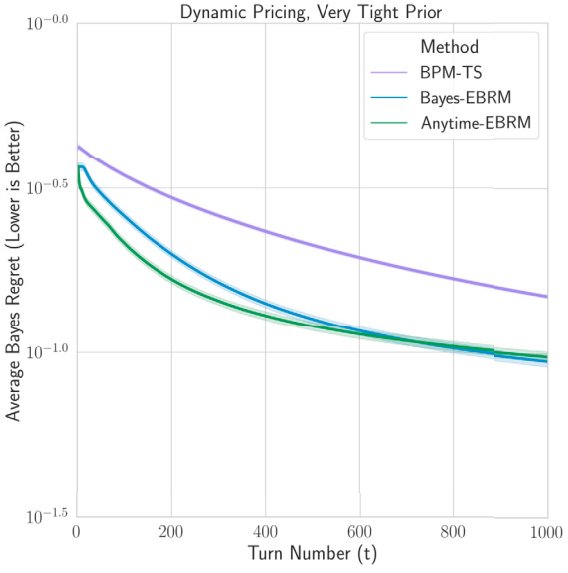
(a) At $s = 2$, we see very similar performance as for the baseline case $s = 1$, shown in Figure 9, across all algorithms.



(b) At $s = 10$, Anytime-EBRM alone shows a noticeable increase in average regret.



(c) At $s = 0.5$, Anytime-EBRM and Bayes-EBRM each demonstrate a slight increase in performance.



(d) At $s = 0.1$, Anytime-EBRM and Bayes-EBRM are both negatively impacted.

Fig. 11. Performance of the EBRM algorithms and BPM-TS on the dynamic pricing task for various priors, differing only in the scale of their respective covariance matrices, s. In general, "wider" priors result in better performance than overly "tight" or "narrow" priors, which can cause the algorithms to under-explore.

exceeding the state of the art in every experiment. The proposed approach is unique in that deriving the immediate risk and value-of-information functions corresponding to a particular online learning problem is enough to characterize the AsympGreedy-EBRM algorithm for that problem.

The EBRM approach represents a change in direction from previous online learning algorithms, which were each designed for optimal performance over some archetypal class of online learning problem and objective, to the design of an algorithm that is *parameterized by* the online learning problem specification itself. While more complicated, this ensures that the behaviour of an EBRM algorithm is always aligned with the task goals, and eliminates the need for hyperparameters. As such, EBRM approaches are a highly effective compromise between POMDP solutions like reinforcement learning, which can optimally solve complex OLPs but often have significant setup and computational costs, and online learning heuristics, which are easy to implement and compute but cannot capture the complexities of specific OLPs.

7.1. Future work

The success of AsympGreedy-EBRM based algorithms in the experiments in Section 6, and specifically the clinical trial experiment in Subsection 6.2, motivates developing EBRM algorithms for other impactful real-world applications. Furthermore, designing new EBRM candidate policy sets, and techniques to compute the process risk of policies beyond 1-step lookahead policies, will enable the development of new EBRM-based algorithms that may provide further performance improvements upon Greedy-EBRM, and represents a new direction of research in online learning theory. A promising initial step in this direction is the incorporation of information-directed sampling techniques [23] and the related theoretical analyses into new techniques to estimate changes in epistemic risk.

There are a wide variety of real-world online learning problems with complexities that could benefit from EBRM approaches. For example, problems with continuous action spaces, typically modelled as linear bandits, should see similar improvements through EBRM-based algorithms. However, they introduce new computational challenges; in particular, the one-step lookahead candidate policy set is infinite given a continuous action space, and so the agent cannot check every policy. Developing techniques to identify the one-step lookahead policy with minimal process risk in a continuous policy set is thus one research priority.

Action costs and cost budgets are additional problem complexities that have been the subject of limited study but have valuable real-world applications. For example, autonomous vehicles and robots may use online learning to learn hidden parameters in their environment; actions could represent using a sensor to take a measurement, or asking a question of a human in order to learn their preferences or objectives. Sensing or communicating such information generally has costs, which may be based on energy used or time spent communicating, and real-world systems have limits on these costs. EBRM-based approaches are well suited to capture these costs and budgets in the action-based reward and feasibility criterion, respectively.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

We acknowledge support for this project from the National Science Foundation (NSF-NRI grants 1734400 and 2133029), and the Woods Hole Oceanographic Institution (WHOI) Investment in Science Fund.

Appendix A. Contrasting risk and regret

A **regret function** measures how much less reward *was* attained, in a posterior sense, by some policy than would have been by the X-optimal policy. That is, if $A_t^\pi = \{a_\tau^\pi\}_{\tau=1}^t$ is the sequence of actions generated by a policy π from an initial belief state S_0, and $S_t^\pi = \{b_t^\pi, \xi_t^\pi\}$ is the corresponding terminal belief state, then for some hidden outcomes X the instance regret is

$$\mathfrak{R}(A_t^\pi; S_0, X) := V^{\pi^\star_X}(S_0; X) - \left(\gamma^t R(b_t^\pi) + \sum_{\tau=1}^{t} \gamma^\tau R(x_\tau, a_\tau^\pi)\right). \tag{A.1}$$

The expected and Bayesian regrets are similarly defined,

$$\overline{\mathfrak{R}}(A_t^\pi; S_0, \theta) := \mathbb{E}_{X|\theta}\left[\mathfrak{R}(A_t^\pi; S_0, X)\right], \tag{A.2}$$

$$\overline{\overline{\mathfrak{R}}}(A_t^\pi; S_0, b_t) := \mathbb{E}_{\theta|b_t}\left[\overline{\mathfrak{R}}(A_t^\pi; S_0, \theta)\right]. \tag{A.3}$$

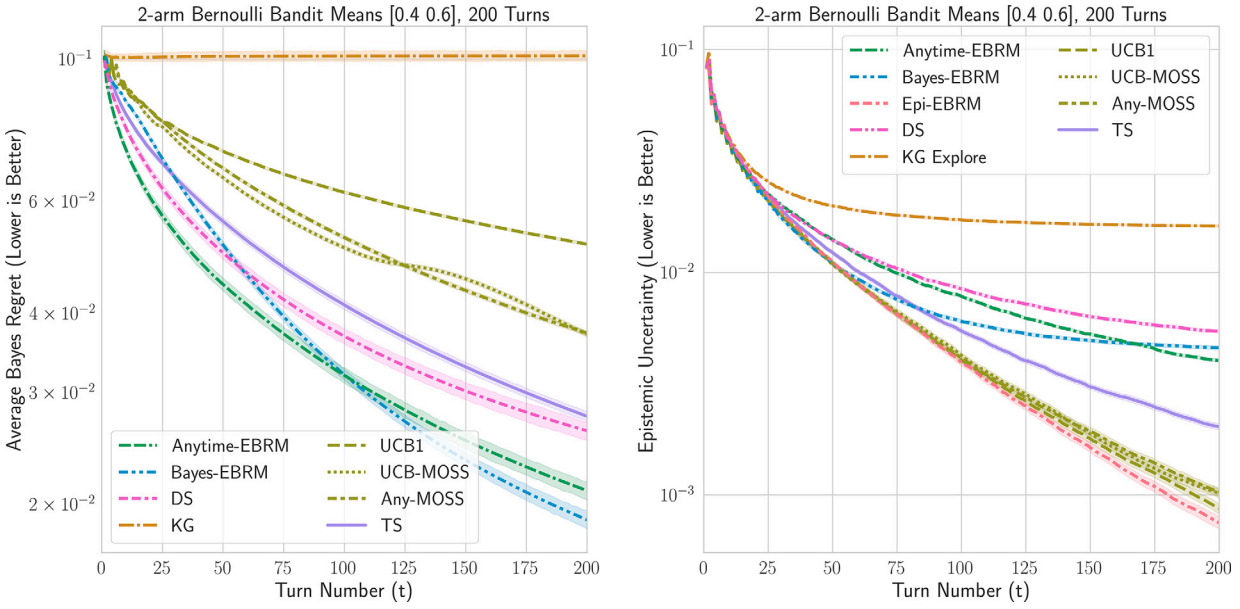
These quantities are defined with respect to a particular sequence of actions and observations made by the agent, unlike risk which is computed based on all possible sequences of actions the agent *might* perform from some initial state. The risk of a policy relative to the X-optimal policy is therefore equal to the *expected* amount of regret that will be incurred by that policy.

Once incurred, regret is "permanent"; by the Tower rule, the Bayesian regret of a sequence of actions does not change, in expectation, due to future actions and observations:

$$\mathbb{E}_{x|b_t}\left[\overline{\overline{\mathfrak{R}}}(A_t^\pi; S_0, b_{t+1}^{x,a})\right] = \overline{\overline{\mathfrak{R}}}(A_t^\pi; S_0, b_t),\ \forall a \in \mathcal{A}. \tag{A.4}$$

Conversely, the agent can reduce its risk (i.e., its future regret) by taking an action that provides information enabling it to make better decisions in the future.

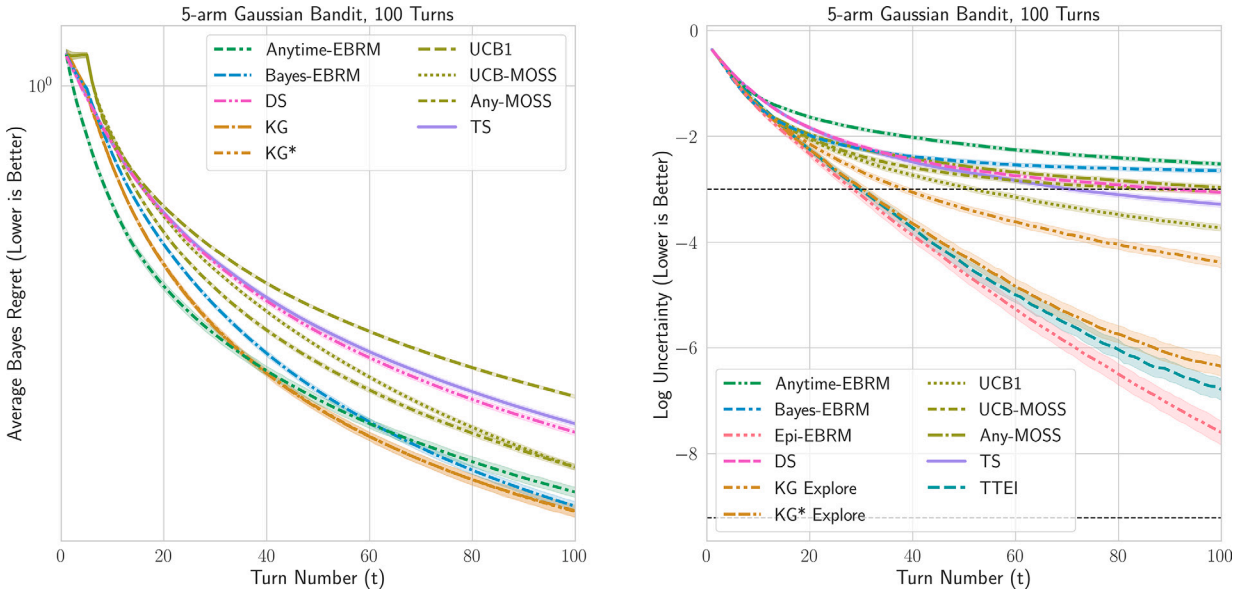
Appendix B. Additional experiments



(a) The AsympGreedy-EBRM approaches are highly successful at reducing the average Bayes regret, even over relatively few turns.

(b) In this simple setting, the UCB algorithms are nearly as effective at reducing epistemic uncertainty as Epi-EBRM.

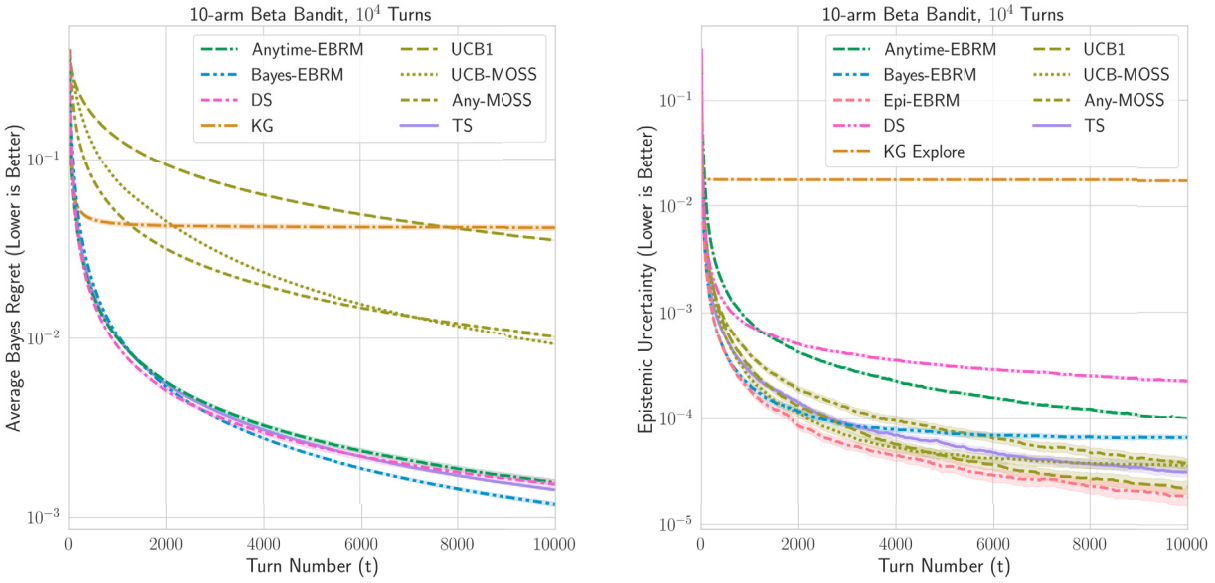
Fig. B.12. The results of this 2-arm Bernoulli bandit experiment from [67] clearly demonstrate the exploration-exploitation trade-off, where the algorithms with the lowest average Bayes regret tend to have the higher epistemic uncertainty, and vice-versa. In particular, observe how Bayes-EBRM rapidly switches from exploration to exploitation around $t = 75$ in order to maximally leverage its accumulated information over the remaining time.



(a) The y-axis shows average sample regret, the bandit optimization metric for which lower is better. Asymp-EBRM has the least average regret initially, before being overtaken by KG and Bayes-EBRM.

(b) The y-axis shows log uncertainty; lower values indicate higher confidence in the best-arm identification. The upper dashed line represents 95% confidence in having found the best arm, while the lower dashed line represents 99.99% confidence.

Fig. B.13. Performance of algorithms in bandit optimization and best-arm identification while making decisions in 5-arm Gaussian Bandit problem over 100 turns. Shaded regions indicate 95% confidence intervals over 2000 trials.



(a) TS, DS, Asymp-EBRM and Bayes-EBRM perform nearly identically over the first 4000 turns, after which Bayes-EBRM begins to significantly outperform the rest.

(b) The y-axis shows epistemic uncertainty; lower values indicate better BAI performance. Epi-EBRM achieves the lowest epistemic uncertainty across trials.

Fig. B.14. Performance of algorithms in BDO and BAI objectives while making decisions in a 10-arm Beta bandit problem over 10^4 turns. Shaded regions indicate 95% confidence intervals over 2000 trials.

Appendix C. Example clinical research trial scenarios

In this section we describe some theoretical clinical research scenarios to highlight the limitations of existing online learning heuristics.

Problem 3 *(Villar et al., [2]).* Consider 423 patients sequentially assigned to various treatments within a randomized controlled trial. Which assignments best address the competing objectives of (1) maximizing the number of successful patient outcomes by assigning as many as possible to the best treatments, and (2) achieving sufficient statistical power to detect a significant difference between treatments?

The study which presented Problem 3 found that several leading online learning algorithms each failed to balance between these two objectives [2]. This is due to their inability to account for a second variable, like statistical power in a weighted objective (e.g. maximize a weighted sum of statistical power and successful patient outcomes). The EBRM approach presented in this work can directly model this type of objectives.

Problem 4. The goal of a similar trial to that in Problem 3 is to determine which treatment to distribute nationally/globally to a much larger population of N patients. Which assignments result in the highest number of successful patient outcomes across all $(N + 423)$ patients who will receive treatments?

Problem 4 is concerned with ensuring that the treatment recommended for broad distribution is the most effective one, or nearly so. This is a different metric than statistical power or confidence, and one which is not directly targeted by any existing algorithm. Furthermore, this problem introduces an explicit weighting; for $N = 10^4$, identifying a treatment that has a 1% higher success rate will result in 100 more successful outcomes across the patient population. In Section 6, we see that the EBRM approach results in the most successful patient outcomes in this type of scenario, for various N. Importantly, the EBRM algorithm does not require any "tuning" to achieve these results; it simply takes N as a parameter.

Problem 5. As an extension to Problem 4, suppose that there is are costs to enrolling patients in the trial, and a unique cost for administering each type of treatment. What are the optimal trial size and treatment assignments in order to maximize successful patient outcomes given a budget constraint?

Most existing online learning heuristics lack the ability to model heterogeneous costs across different actions (treatments). Recent work has begun to model such complexities (e.g., [72]), however this approach required developing another heuristic which focuses

on maximizing reward per unit cost. The EBRM approach automatically takes into account various feasibility criterion, including those driven by action costs.

Appendix D. Proof of Section 4 results

Proof of Theorem 1

First observe that the identifiability of the hidden parameters and the continuous mapping theorem [73] together imply that

$$\mathbb{E}_{x|b_t}[f(\cdot)] \xrightarrow[t\to\infty]{P} \mathbb{E}_{x|\theta}[f(\cdot)].$$

As the belief reward is bounded and continuous, then, also by the continuous mapping theorem, there exists some $c \in \mathbb{R}$ such that

$$R(b_t) \xrightarrow[t\to\infty]{P} c.$$

As the belief reward converges in probability to a constant as $t \to \infty$, we have that

$$\begin{aligned}\text{EpistemicRisk}(S_t) &= \mathbb{E}_{\theta|b_t}\left[V^{\pi^\star_\theta}(S_t;\theta)\right] - V^{\hat{\pi}^\star_t}(S_t) \\ &\xrightarrow[t\to\infty]{P} \sum_{n=1}^{\infty} \mathbb{E}_{x|\theta}\left[\gamma^{n-1} R(x, \pi^\star_\theta(S_{t+n-1}))\right] - \sum_{n=1}^{\infty} \mathbb{E}_{x|\theta}\left[\gamma^{n-1} R(x, \hat{\pi}^\star_t(S_{t+n-1}))\right]\end{aligned}$$

As each sum is independent of any future observations or hidden variables, they can each be maximized by some deterministic open-loop policy $\hat{\pi} \in \hat{\Pi}$ and thus

$$\mathbb{E}_{x|\theta}\left[R(x, \hat{\pi}^\star_t(S_t))\right] \xrightarrow[t\to\infty]{P} \mathbb{E}_{x|\theta}\left[R(x, \pi^\star_\theta(S_t))\right].$$

Proof of Lemma 2

Let σ be the permutation function that satisfies $A' = \sigma(A)$, and let $X' = \sigma(X)$ be the corresponding permutation of X. First, we observe,

$$x_\tau \overset{\text{iid}}{\sim} P(x \mid b_t) \implies \Pr\left(X' \mid b_t\right) = \Pr\left(X \mid b_t\right).$$

Furthermore, denoting $y'_\tau = \Phi\left(x'_\tau, a'_\tau\right)$ and $y_\tau = \Phi\left(x_\tau, a_\tau\right)$ for $\tau = 1, \ldots, |A|$,

$$\begin{aligned}x_\tau \overset{\text{iid}}{\sim} P(x \mid b_t) &\implies \Pr(y'_{1:|A'|} \mid \theta, a'_{1:|A'|}) = \Pr\left(y_{1:|A|} \mid \theta, a_{1:|A|}\right), \\ &\implies b^{X',A'}_{t+|A'|}(\theta) = b^{X,A}_{t+|A|}(\theta), \quad \forall \theta \in \Theta.\end{aligned}$$

Appendix E. Proofs of Section 5 results

Proof of Lemma 3

The adaptive monotonicity of $R(b)$ and the convexity of the expectation operator imply

$$\mathbb{E}_{x|b_t}\left[\max_{N \in \mathbb{N}^K_\Omega\left(\xi^a_{t+1}\right)} \hat{V}\left(N; S^{x,a}_{t+1}\right)\right] \geq \max_{N \in \mathbb{N}^K_\Omega\left(\xi^a_{t+1}\right)} \mathbb{E}_{x|b_t}\left[\hat{V}\left(N; S^{x,a}_{t+1}\right)\right] \geq \max_{N \in \mathbb{N}^K_\Omega\left(\xi^a_{t+1}\right)} \hat{V}\left(N; \mathbb{E}_{x|b_t}\left[S^{x,a}_{t+1}\right]\right).$$

The lemma follows from expanding Eq. (5.8) with the definition of each term and applying this inequality.

Proof of Proposition 3

We introduce Lemma 5 to simplify the proof.

Lemma 5. *The process risk of the best deterministic* 1*-step lookahead policy is bounded,* $\forall S_t$*,*

$$-\text{EpistemicRisk}(S_t) \leq \min_{a \in \mathcal{A}} \text{ProcessRisk}(\pi_{t,a}; S_t) \leq 0. \tag{E.1}$$

The lower bound of Lemma 5 follows from, $\forall \pi$ and $\forall S_t$, (see Lemma 1)

$$0 \leq \bar{\bar{r}}(\pi \| \pi^\star_\theta; S_t) = \text{EpistemicRisk}(S_t) + \text{ProcessRisk}(\pi; S_t).$$

The upper bound follows from Lemma 3, using the fact that $\text{ImmediateRisk}(S_t, \pi_{t,a}) = 0$ if $a = \hat{\pi}^\star_{S_t}(S_t)$. According to Lemma 5, the best 1-step lookahead policy has non-positive process risk, but by Lemma 3,

$$\begin{aligned}\text{ProcessRisk}(S_t, a) &= \text{ImmediateRisk}(S_t, a) - \gamma \cdot \text{ExpectedVoI}(S_t, a) \\ &\geq \text{ImmediateRisk}(S_t, a) - \gamma \cdot \text{EpistemicRisk}\left(\mathbb{E}_{x|b_t}\left[S^{x,a}_{t+1}\right]\right).\end{aligned}$$

The proposition follows.

Proof of Theorem 2

Consider $\mathbb{E}\left[\mu_a\right] = \text{ProcessRisk}\left(\pi_{t,a}; S_t\right) - \text{ProcessRisk}\left(\pi_{t,k}; S_t\right)$, and by construction $\mu_a \geq 0$. Therefore,

$$\mathbb{E}\left[\mu_a\right] > 0 \implies \text{ProcessRisk}\left(\pi_{t,k}; S_t\right) < \text{ProcessRisk}\left(\pi_{t,a}; S_t\right). \tag{E.2}$$

Next, $\text{Var}\left[\mu_a\right] = \gamma^2(\sigma_a^2 n_a^{-1} + \sigma_k^2 n_k^{-1})$ follows from the independence of $\nu_a(n_a)$ and $\nu_k(n_k)$, which are generated from IID samples of the hidden outcomes. The theorem follows from Cantelli's inequality [74] applied to $\Pr\left(\mu_a - \mathbb{E}\left[\mu_a\right] < \mu_a\right)$.

Proof of Theorem 3

Following from the proof of Theorem 2, observe that $\forall n > 0$,

$$\max_{N \in \mathbb{N}^K_\Omega\left(\xi^a_{t+1}\right)} \hat{V}\left(N; S^{x,a}_{t+1}\right) - \hat{V}\left(N^a_{t+1}; S^{x,a}_{t+1}\right) \leq M_a \implies \nu_a(n) \leq M_a, \tag{E.3}$$

$$\implies \mu_a \leq \left(\text{ImmediateRisk}(S_t, a) - \text{ImmediateRisk}(S_t, k)\right) + \gamma M_k, \tag{E.4}$$

$$\implies \mu_a \geq \left(\text{ImmediateRisk}(S_t, a) - \text{ImmediateRisk}(S_t, k)\right) - \gamma M_a. \tag{E.5}$$

Next, we consider that for $n_a = n_k = n$, then $n\mu_a$ is equal to the sum of n IID samples of the difference in process risk between policies $\pi_{t,a}$ and $\pi_{t,k}$,

$$\begin{aligned}n\mu_a = \sum_{i=1}^{n} \Bigg[&\text{ImmediateRisk}(S_t, a) - \gamma \left(\max_{N \in \mathbb{N}^K_\Omega\left(\xi^a_{t+1}\right)} \hat{V}\left(N; S^{x_i,a}_{t+1}\right) - \hat{V}\left(N^a_{t+1}; S^{x_i,a}_{t+1}\right) \right) \\ &- \left(\text{ImmediateRisk}(S_t, k) - \gamma \left(\max_{N \in \mathbb{N}^K_\Omega\left(\xi^k_{t+1}\right)} \hat{V}\left(N; S^{x_i,k}_{t+1}\right) - \hat{V}\left(N^k_{t+1}; S^{x_i,k}_{t+1}\right) \right) \right) \Bigg],\end{aligned} \tag{E.6}$$

with $\mathbb{E}\left[n\mu_a\right] = n(\text{ProcessRisk}(\pi_{t,a}; S_t) - \text{ProcessRisk}(\pi_{t,k}; S_t))$. Then, applying Hoeffding's inequality,

$$\Pr\left(n\mu_a - \mathbb{E}\left[n\mu_a\right] \geq c\right) \leq \exp\left(\frac{-2c^2}{n(\gamma M_k + \gamma M_a)^2}\right). \tag{E.7}$$

The theorem follows from taking $c = n\mu_a$ and recognizing that the largest amount of regret which could be incurred by rejecting policy $\pi_{t,a}$ in favour of $\pi_{t,k}$ is $\gamma\left(M_a + M_k\right)$, and the probability of incurring this regret is $\Pr\left(\pi_{t,a} \in \arg\min_{\pi \in \Pi_t} \text{ProcessRisk}(\pi; S_t)\right) \leq \Pr\left(\mathbb{E}\left[n\mu_a\right] < 0\right)$.

References

[1] T. Lattimore, C. Szepesvári, Bandit Algorithms, 1 ed., Cambridge University Press, 2020, https://doi.org/10.1017/9781108571401, https://www.cambridge.org/core/product/identifier/9781108571401/type/book.

[2] S.S. Villar, J. Bowden, J. Wason, Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges, Stat. Sci. 30 (2015) 199–215, https://doi.org/10.1214/14-STS504, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4856206/.

[3] W.R. Thompson, On the likelihood that one unknown probability exceeds another in view of the evidence of two samples, Biometrika 25 (1933) 285–294, https://doi.org/10.2307/2332286, http://www.jstor.org/stable/2332286.

[4] T. Lattimore, C. Szepesvári, An information-theoretic approach to minimax regret in partial monitoring, in: A. Beygelzimer, D. Hsu (Eds.), Proceedings of Machine Learning Research, vol. 99, 2019, pp. 2111–2139, https://proceedings.mlr.press/v99/lattimore19a.html, arXiv:1902.00470.

[5] G. Bartók, D. Pál, C. Szepesvári, Minimax regret of finite partial-monitoring games in stochastic environments, in: Proceedings of the 24th Annual Conference on Learning Theory, JMLR Workshop and Conference Proceedings, 2011, pp. 133–154, https://proceedings.mlr.press/v19/bartok11a.html, ISSN: 1938-7228.

[6] G. Bartók, D.P. Foster, D. Pál, A. Rakhlin, C. Szepesvári, Partial monitoring—classification, regret bounds, and algorithms, Math. Oper. Res. 39 (2014) 967–997, https://doi.org/10.1287/moor.2014.0663, http://pubsonline.informs.org/doi/10.1287/moor.2014.0663.

[7] R. Kleinberg, T. Leighton, The value of knowing a demand curve: bounds on regret for online posted-price auctions, in: 44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings, ISSN 0272-5428, IEEE Computer Society, Cambridge, MA, USA, 2003, pp. 594–605, https://doi.org/10.1109/SFCS.2003.1238232, https://www.computer.org/csdl/proceedings-article/focs/2003/20400594/12OmNvkpl3b.

[8] G. Bartok, N. Zolghadr, C. Szepesvári, An adaptive algorithm for finite stochastic partial monitoring, in: Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012, pp. 1–20.

[9] M. Aziz, E. Kaufmann, M.-K. Riviere, On multi-armed bandit designs for dose-finding clinical trials, J. Mach. Learn. Res. 22 (2021) 38, https://www.jmlr.org/papers/volume22/19-228/19-228.pdf, arXiv:1903.07082.
[10] S.L. Scott, A modern Bayesian look at the multi-armed bandit, Appl. Stoch. Models Bus. Ind. 26 (2010) 639–658, https://onlinelibrary.wiley.com/doi/10.1002/asmb.874.
[11] R.I. Brafman, M. Tennenholtz, R-max – a general polynomial time algorithm for near-optimal reinforcement learning, J. Mach. Learn. Res. 3 (2002) 213–231.
[12] T. Jaksch, R. Ortner, P. Auer, Near-optimal regret bounds for reinforcement learning, J. Mach. Learn. Res. 11 (2010) 38.
[13] P. Auer, N. Cesa-Bianchi, P. Fisher, Finite-time analysis of the multiarmed bandit problem, in: Machine Learning, vol. 47, Kluwer Academic Publishers, 2002, pp. 235–256.
[14] E. Kaufmann, O. Cappe, A. Garivier, On Bayesian upper confidence bounds for bandit problems, in: N.D. Lawrence, M. Girolami (Eds.), Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 22, PMLR, La Palma, Canary Islands, 2012, pp. 592–600.
[15] K. Jamieson, M. Malloy, R. Nowak, S. Bubeck, Lil' UCB: an optimal exploration algorithm for multi-armed bandits, J. Mach. Learn. Res. 35 (2014) 423–439, https://proceedings.mlr.press/v35/jamieson14.html.
[16] K. Jamieson, R. Nowak, Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting, in: 2014 48th Annual Conference on Information Sciences and Systems (CISS), 2014, https://doi.org/10.1109/CISS.2014.6814096, http://ieeexplore.ieee.org/document/6814096/.
[17] K. Misra, E.M. Schwartz, J. Abernethy, Dynamic Online Pricing with Incomplete Information Using Multiarmed Bandit Experiments, Mark. Sci. 38 (2019) 226–252, http://pubsonline.informs.org/doi/10.1287/mksc.2018.1129.
[18] D. Russo, Technical Note—A Note on the Equivalence of Upper Confidence Bounds and Gittins Indices for Patient Agents, Oper. Res. 69 (2020) 273–278, https://doi.org/10.1287/opre.2020.1987, https://pubsonline.informs.org/doi/abs/10.1287/opre.2020.1987.
[19] D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions, J. Glob. Optim. 13 (1998) 455–492, https://doi.org/10.1023/A:1008306431147.
[20] C. Qin, D. Klabjan, D. Russo, Improving the expected improvement algorithm, in: 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017, p. 11.
[21] I.O. Ryzhov, W.B. Powell, P.I. Frazier, The knowledge gradient algorithm for a general class of online learning problems, Oper. Res. 60 (2012) 180–195, https://doi.org/10.1287/opre.1110.0999, http://pubsonline.informs.org/doi/abs/10.1287/opre.1110.0999.
[22] I.O. Ryzhov, P.I. Frazier, W.B. Powell, On the robustness of a one-period look-ahead policy in multi-armed bandit problems, Proc. Comput. Sci. 1 (2010) 1635–1644, https://doi.org/10.1016/j.procs.2010.04.183, https://www.sciencedirect.com/science/article/pii/S1877050910001845.
[23] D. Russo, B. Van Roy, Learning to optimize via information-directed sampling, Oper. Res. 66 (2018) 230–252, https://doi.org/10.1287/opre.2017.1663, http://pubsonline.informs.org/doi/10.1287/opre.2017.1663.
[24] J. Kirschner, T. Lattimore, A. Krause, Information Directed Sampling for Linear Partial Monitoring, in: J. Abernethy, S. Agarwal (Eds.), Proceedings of Thirty Third Conference on Learning Theory, in: Proceedings of Machine Learning Research, vol. 125, PMLR, 2020, pp. 2328–2369, https://proceedings.mlr.press/v125/kirschner20a.html.
[25] J. Kirschner, T. Lattimore, C. Vernade, C. Szepesvári, Asymptotically Optimal Information-Directed Sampling, in: M. Belkin, S. Kpotufe (Eds.), Proceedings of Thirty Fourth Conference on Learning Theory, in: Proceedings of Machine Learning Research, vol. 134, PMLR, 2021, pp. 2777–2821, https://proceedings.mlr.press/v134/kirschner21a.html.
[26] Y. Wang, W.B. Powell, Finite-time analysis for the knowledge-gradient policy, SIAM J. Control Optim. 56 (2018) 1105–1129, https://doi.org/10.1137/16M1073388, https://epubs.siam.org/doi/10.1137/16M1073388, arXiv:1606.04624.
[27] D. Russo, Simple Bayesian algorithms for best-arm identification, Oper. Res. 68 (2020) 1625–1647, https://doi.org/10.1287/opre.2019.1911, https://pubsonline.informs.org/doi/abs/10.1287/opre.2019.1911, arXiv:1602.08448.
[28] S.S. Gupta, K.J. Miescke, Bayesian look ahead one-stage sampling allocations for selection of the best population, J. Stat. Plan. Inference 54 (1996) 229–244, https://doi.org/10.1016/0378-3758(95)00169-7, https://www.sciencedirect.com/science/article/pii/0378375895001697.
[29] P.I. Frazier, W.B. Powell, S. Dayanik, A knowledge-gradient policy for sequential information collection, SIAM J. Control Optim. 47 (2008) 2410–2439, https://doi.org/10.1137/070693424, http://epubs.siam.org/doi/10.1137/070693424.
[30] R. Bellman, A Markovian decision process, J. Math. Mech. 6 (1957) 679–684, https://www.jstor.org/stable/24900506.
[31] L.P. Kaelbling, M.L. Littman, A.R. Cassandra, Planning and acting in partially observable stochastic domains, Artif. Intell. 101 (1998) 99–134, https://doi.org/10.1016/S0004-3702(98)00023-X, https://www.sciencedirect.com/science/article/pii/S000437029800023X.
[32] C. Cai, X. Liao, L. Carin, Learning to explore and exploit in POMDPs, in: Advances in Neural Information Processing Systems, vol. 22, Curran Associates, Inc., 2009, https://proceedings.neurips.cc/paper_files/paper/2009/hash/e58cc5ca94270acaceed13bc82dfedf7-Abstract.html.
[33] A. Sharma, J. Harrison, M. Tsao, M. Pavone, Robust and adaptive planning under model uncertainty, in: Proceedings of the Twenty-Ninth International Conference on Automated Planning and Scheduling, Association for the Advancement of Artificial Intelligence, 2019, https://doi.org/10.1609/icaps.v29i1.3505, arXiv:1901.02577 [cs].
[34] M. Ghavamzadeh, S. Mannor, J. Pineau, A. Tamar, Bayesian reinforcement learning: a survey, Found. Trends Mach. Learn. 8 (2015) 359–483, https://doi.org/10.1561/2200000049, arXiv:1609.04436.
[35] Q. Liu, A. Chung, C. Szepesvári, C. Jin, When is partially observable reinforcement learning not scary?, in: Proceedings of Thirty Fifth Conference on Learning Theory, in: PMLR, ISSN 2640-3498, 2022, pp. 5175–5220, https://proceedings.mlr.press/v178/liu22f.html.
[36] G. Arcieri, C. Hoelzl, O. Schwery, D. Straub, K.G. Papakonstantinou, E. Chatzi, Bridging POMDPs and Bayesian decision making for robust maintenance planning under model uncertainty: an application to railway systems, Reliab. Eng. Syst. Saf. 239 (2023) 109496, https://doi.org/10.1016/j.ress.2023.109496, https://www.sciencedirect.com/science/article/pii/S0951832023004106.
[37] P. Sharma, B. Kraske, J. Kim, Z. Laouar, Z. Sunberg, E. Atkins, Risk-aware Markov decision process contingency management autonomy for uncrewed aircraft systems, J. Aerosp. Inform. Syst. (2024) 1–15, https://doi.org/10.2514/1.I011235.
[38] R. Meshram, A. Gopalan, D. Manjunath, Optimal recommendation to users that react: online learning for a class of POMDPs, in: 2016 IEEE 55th Conference on Decision and Control (CDC), 2016, pp. 7210–7215, https://doi.org/10.1109/CDC.2016.7799381, https://ieeexplore.ieee.org/abstract/document/7799381.
[39] M. Komorowski, L.A. Celi, O. Badawi, A.C. Gordon, A.A. Faisal, The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care, Nat. Med. 24 (2018) 1716–1720, https://doi.org/10.1038/s41591-018-0213-5, https://www.nature.com/articles/s41591-018-0213-5.
[40] M.O. Duff, A. Barto, Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes, PhD Thesis, University of Massachusetts Amherst, 2002, https://www.gatsby.ucl.ac.uk/~yael/Okinawa/DuffThesis.pdf.
[41] J. Garcıa, F. Fernández, A comprehensive survey on safe reinforcement learning, J. Mach. Learn. Res. 16 (2015) 1437–1480.
[42] X. Huo, F. Fu, Risk-aware multi-armed bandit problem with application to portfolio selection, R. Soc. Open Sci. 4 (2017) 171377, https://doi.org/10.1098/rsos.171377, https://royalsocietypublishing.org/doi/full/10.1098/rsos.171377.
[43] Y. Shen, M.J. Tobia, T. Sommer, K. Obermayer, Risk-sensitive reinforcement learning, Neural Comput. 26 (2014) 1298–1328, https://doi.org/10.1162/NECO_a_00600.
[44] M. Rigter, B. Lacerda, N. Hawes, Risk-averse Bayes-adaptive reinforcement learning, in: Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., 2021, pp. 1142–1154, https://proceedings.neurips.cc/paper_files/paper/2021/hash/08f90c1a417155361a5c4b8d297e0d78-Abstract.html.
[45] S. Al-Hussaini, N. Dhanaraj, J.M. Gregory, R. Jomy Joseph, S. Thakar, B.C. Shah, J.A. Marvel, S.K. Gupta, Seeking human help to manage plan failure risks in semi-autonomous mobile manipulation, J. Comput. Inf. Sci. Eng. 22 (2022), https://doi.org/10.1115/1.4054088.

[46] B. Charpentier, R. Senanayake, M. Kochenderfer, S. Günnemann, Disentangling epistemic and aleatoric uncertainty in reinforcement learning, http://arxiv.org/abs/2206.01558, https://doi.org/10.48550/arXiv.2206.01558, arXiv:2206.01558 [cs], 2022.
[47] P. Festor, G. Luise, M. Komorowski, A.A. Faisal, Enabling risk-aware reinforcement learning for medical interventions through uncertainty decomposition, http://arxiv.org/abs/2109.07827, https://doi.org/10.48550/arXiv.2109.07827, arXiv:2109.07827 [cs], 2022.
[48] X. Lu, B. Van Roy, V. Dwaracherla, M. Ibrahimi, I. Osband, Z. Wen, Reinforcement learning, bit by bit, in: Now Foundations and Trends, 2023.
[49] Y. Lin, Y. Ren, E. Zhou, Bayesian risk Markov decision processes, Adv. Neural Inf. Process. Syst. 35 (2022) 17430–17442, https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f7d90b1198fec96defd80b5ebd5bc81-Abstract-Conference.html.
[50] A. Guez, D. Silver, P. Dayan, Scalable and efficient Bayes-adaptive reinforcement learning based on Monte-Carlo tree search, J. Artif. Intell. Res. 48 (2013) 841–883, https://doi.org/10.1613/jair.4117, https://www.jair.org/index.php/jair/article/view/10853.
[51] G. Lee, B. Hou, A. Mandalika, J. Lee, S. Choudhury, S.S. Srinivasa, Bayesian policy optimization for model uncertainty, http://arxiv.org/abs/1810.01014, https://doi.org/10.48550/arXiv.1810.01014, arXiv:1810.01014 [cs], 2019.
[52] H. Eriksson, C. Dimitrakakis, Epistemic risk-sensitive reinforcement learning, in: Proceedings of the 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2019, pp. 339–344, http://arxiv.org/abs/1906.06273, arXiv:1906.06273 [cs, stat].
[53] M. Rigter, B. Lacerda, N. Hawes, One risk to rule them all: addressing distributional shift in offline reinforcement learning via risk-aversion, http://arxiv.org/abs/2212.00124, https://doi.org/10.48550/arXiv.2212.00124, arXiv:2212.00124 [cs], 2023.
[54] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, S. Udluft, Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning, in: International Conference on Machine Learning, in: PMLR, 2018, pp. 1184–1193.
[55] D. Golovin, A. Krause, Adaptive submodularity: theory and applications in active learning and stochastic optimization, J. Artif. Intell. Res. 42 (2011) 60.
[56] K. Murota, Discrete Convex Analysis, Discrete Mathematics and Applications, Society for Industrial and Applied Mathematics, 2003, https://doi.org/10.1137/1.9780898718508, https://epubs.siam.org/doi/book/10.1137/1.9780898718508.
[57] K. Murota, 6. M-convex functions, in: Discrete Convex Analysis, Discrete Mathematics and Applications, Society for Industrial and Applied Mathematics, 2003, pp. 133–176, https://doi.org/10.1137/1.9780898718508.ch6, https://epubs.siam.org/doi/10.1137/1.9780898718508.ch6.
[58] K. Murota, 10. Algorithms, in: Discrete Convex Analysis, Discrete Mathematics and Applications, Society for Industrial and Applied Mathematics, 2003, pp. 281–322, https://doi.org/10.1137/1.9780898718508.ch10, https://epubs.siam.org/doi/10.1137/1.9780898718508.ch10.
[59] G.L. Nemhauser, L.A. Wolsey, M.L. Fisher, An analysis of approximations for maximizing submodular set functions—I, Math. Program. 14 (1978) 265–294, https://doi.org/10.1007/BF01588971.
[60] H. Kellerer, U. Pferschy, D. Pisinger, The bounded knapsack problem, in: H. Kellerer, U. Pferschy, D. Pisinger (Eds.), Knapsack Problems, Springer, Berlin, Heidelberg, 2004, pp. 185–209, https://doi.org/10.1007/978-3-540-24777-7_7.
[61] R. Iyer, J. Bilmes, Submodular optimization with submodular cover and submodular knapsack constraints, in: Proceedings of the 26th International Conference on Neural Information Processing Systems, vol. 2, NIPS'13, Lake Tahoe, Nevada, Curran Associates Inc., Red Hook, NY, USA, 2013, pp. 2436–2444.
[62] W. Hoeffding, On sequences of sums of independent random vectors, in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 2, in: Contributions to Probability Theory, vol. 2, University of California Press, Berkeley, CA, 1961, pp. 213–227, https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fourth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/On-Sequences-of-Sums-of-Independent-Random-Vectors/bsmsp/1200512603.
[63] L.L. Cam, Asymptotic Methods in Statistical Decision Theory, Springer Series in Statistics, Springer, New York, NY, 1986, https://doi.org/10.1007/978-1-4612-4946-7, http://link.springer.com/10.1007/978-1-4612-4946-7.
[64] S. Bubeck, R. Munos, G. Stoltz, Pure exploration in multi-armed bandits problems, in: R. Gavaldà, G. Lugosi, T. Zeugmann, S. Zilles (Eds.), Algorithmic Learning Theory, in: Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2009, pp. 23–37.
[65] D. Russo, B. Van Roy, Learning to optimize via posterior sampling, Math. Oper. Res. 39 (2014) 1221–1243, https://doi.org/10.1287/moor.2014.0650, http://www.jstor.org/stable/24541007.
[66] I. Urteaga, C.H. Wiggins, Bayesian bandits: balancing the exploration-exploitation tradeoff via double sampling, arXiv:1709.03162 [cs, stat], 2018, http://arxiv.org/abs/1709.03162, arXiv:1709.03162.
[67] C.-W. Hsu, B. Kveton, O. Meshi, M. Mladenov, C. Szepesvári, Empirical Bayes regret minimization, arXiv:1904.02664 [cs, stat], 2020, http://arxiv.org/abs/1904.02664, arXiv:1904.02664.
[68] P. Frazier, W. Powell, S. Dayanik, The knowledge-gradient policy for correlated normal beliefs, INFORMS J. Comput. 21 (2009) 599–613, https://doi.org/10.1287/ijoc.1080.0314, http://pubsonline.informs.org/doi/abs/10.1287/ijoc.1080.0314.
[69] J.-Y. Audibert, S. Bubeck, Regret bounds and minimax policies under partial monitoring, J. Mach. Learn. Res. 11 (2010) 2785–2836, http://jmlr.org/papers/v11/audibert10a.html.
[70] R. Degenne, V. Perchet, Anytime optimal algorithms in stochastic multi-armed bandits, in: Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 1587–1595, https://proceedings.mlr.press/v48/degenne16.html, ISSN: 1938-7228.
[71] H.P. Vanchinathan, G. Bartók, A. Krause, Efficient partial monitoring with prior information, in: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 27, Curran Associates, Inc., 2014, https://proceedings.neurips.cc/paper_files/paper/2014/file/0a113ef6b61820daa5611c870ed8d5ee-Paper.pdf.
[72] S. Cayci, A. Eryilmaz, R. Srikant, Budget-constrained bandits over general cost and reward distributions, in: S. Chiappa, R. Calandra (Eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, PMLR, vol. 108, 2020, pp. 4388–4398, https://proceedings.mlr.press/v108/cayci20a.html, arXiv:2003.00365.
[73] H.B. Mann, A. Wald, On stochastic limit and order relationships, Ann. Math. Stat. 14 (1943) 217–226, https://doi.org/10.1214/aoms/1177731415, https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-14/issue-3/On-Stochastic-Limit-and-Order-Relationships/10.1214/aoms/1177731415.full.
[74] F.P. Cantelli, Sui confini della probabilità, in: Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de Settembre di 1928, vol. 6, 1929 (Comunicazioni, Sezione IV (A)-V-VII), 1928, pp. 47–60, https://dialnet.unirioja.es/servlet/articulo?codigo=3183299.