Bayesian Decision Making observing an Expert

Anonymous Author(s)
Affiliation
Address
email

Abstract

Learning agents are increasingly deployed alongside existing experts, such as human operators or previously trained agents. While Bayesian methods like Thompson Sampling offer principled approaches to trade-offs between reward learning and information gain, it is unclear how a learner should optimally incorporate expert information, which differs in kind from its own action-outcome experiences. We study this problem of online Bayesian learning next to an expert in multi-armed bandits. We consider: (i) *an offline setting*, where the learner receives a dataset of outcomes from the expert's optimal arm before interaction, and (ii) *a simultaneous setting*, where the learner must choose at each step whether to update its beliefs using its own experience or the expert's concurrent outcome. We formalize how expert data influences the learner's posterior, and prove that pretraining on expert outcomes tightens information-theoretic regret bounds by the mutual information between the expert data and the optimal arm. For the simultaneous setting, we propose an information-directed rule where the learner processes the data source that maximizes the one-step information gain about the optimal arm. We empirically validate our findings, showing that the value of expert information is highest in asymmetric environments where it can significantly prune the parameter space, and we demonstrate that our information-directed agent successfully leverages this to accelerate learning.

1 Introduction

Many learning systems are deployed *next to other learners*: an agent learning online may co-exist with a party that already knows how to act well in the same environment (a human operator, a calibrated controller, or a previously trained policy). Examples of this are clinical decision support (learning beside clinicians), robotics (learning beside a safe supervisor), and online platforms (learning beside a well-tuned baseline). While Bayesian bandit algorithms and Thompson Sampling in particular offer efficient exploration strategies with information–theoretic regret guarantees [Thompson, 1933, Russo and Van Roy, 2014, 2016], it remains unclear how a Bayesian learner should optimally use expert information that differs in kind from its own action–outcome experience. In particular, we study *online Bayesian learning next to an expert* in multi-armed bandits. The learner interacts with a bandit with unknown characteristics, while an expert (who knows the optimal action $A^*(\theta^*)$) reveals observable outcomes. We consider two settings of access to expert information: (i) an *offline* dataset of outcomes from the optimal arm collected before interaction, and (ii) *simultaneous* learning where, at each round, the learner may process either its own action–outcome pair or an expert outcome. These raise two basic questions. **How should expert data be incorporated in a Bayesian bandit?** Intuitively, knowledge of the optimal action distribution should prune parameter values that cannot induce that optimal arm distribution. We formalize this intuition and show that using an expert dataset to update the prior via the likelihood of the optimal arm yields a posterior that converges to the ideal update that conditions directly on a known optimal distribution (Proposition 1). Moreover, probabilities of arm optimality computed under this posterior coincide with the usual Bayesian posterior over the optimal

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

arm (Proposition 2). Plugging these identities into the information–ratio framework of Russo and Van Roy [2016], we obtain that the expected regret of Thompson Sampling with the expert-updated prior is governed by the *reduction in entropy* of the optimal-arm random variable (Theorem 1). In expectation over expert data, pretraining on expert outcomes strictly decreases the entropy of A^*, and thus tightens the regret bound. **When learning online, which source should the learner pay attention to?** In the simultaneous mode, the learner can observe expert outcomes but not actions. Alternatively, the learner's experiences contain both actions and outcomes. Because rewards do not depend on which source is processed, an information–theoretic decision rule emerges: *at each round, process the source that maximizes the mutual information (MI) with the optimal arm*. We propose a simple particle-based estimator that compares one-step posterior entropies after hypothetically conditioning on either source and chooses the larger information gain (Algorithm 1). The rule hinges on Bayesian experimental design [Lindley, 1956], with the *target* being the optimal action rather than the parameter.

Contributions We study an online Bayesian learning next to an expert. (i) We show that the proposed Bayesian inference on finite expert datasets converge to the ideal infinite-information update (Proposition 1) and that arm-optimality probabilities computed under this posterior equal the standard posterior over A^* (Proposition 2). (ii) Leveraging Russo and Van Roy [2016], we show tighter Bayesian regret bounds and a clean measure of the value of expert data (Theorem 1). (iii) We propose a particle-based algorithm to choose between expert and self information by maximizing one-step mutual information about A^* (Algorithm 1), and we discuss bias/variance trade-offs of one-sample estimates. (iv) Empirically we show expert information provides no gains in a symmetric bandit case (as expert outcomes add no discriminative information about θ), substantial regret reductions from offline expert data in asymmetric worlds, and dramatic improvements in strongly asymmetric worlds where expert outcomes nearly identify θ^*.

Main Insight Expert information is most valuable when it moves probability mass between optimal arms; its value is exactly the reduction in uncertainty about the optimal action. Framing *who to learn from* as information acquisition about the optimal arm yields both interpretable theory and practical algorithms that result in agents knowing when to listen to the expert.

1.1 Related Work

Bandits and Beliefs Thompson Sampling [Thompson, 1933] has been a prevalent Bayesian algorithm for online learning for decades [Agrawal and Goyal, 2012, Chapelle and Li, 2011, Russo et al., 2018]. Russo and Van Roy [2016] made the explicit connection between the regret bounds and efficiency of Thompson Sampling and information theoretic quantities on the agent decision rules. There are also many examples of multi-agent bandit problems [Brânzei and Peres, 2021, Chang and Lu, 2025] where the question of agent information is introduced. To the best of our knowledge, these works do not consider how to incorporate expert samples in a Bayesian update and how this affects Thompson Sampling regrets. Additionally, our work traces back to early game-theoretic and theory-of-mind ideas. Works as Geanakoplos and Polemarchakis [1982], Moses and Nachum [1990] discussed the implications of agents with different belief structures sharing information to learn. Our work considers how do these ideas apply to a reward maximisation (online learning) problem.

Learning from Experts and Demonstrations Our work is also connected to the broad literature on learning from expert feedback. Particularly, imitation learning and inverse reinforcement learning focus on inferring a policy or reward function from an expert's actions [Abbeel and Ng, 2004, Ross et al., 2011]. Our approach differs fundamentally: we do not observe the expert's actions, but rather the outcomes generated by their known-optimal policy. This shifts the inference problem from "what did the expert do?" to "what must the world be like for the expert's policy to be optimal?". Furthermore, our setting diverges from the (frequentist) *bandits with expert advice* framework [Cesa-Bianchi et al., 1997, Auer et al., 2002], RL with expert information [Gimelfarb et al., 2018] or best arm selection problems with offline data [Agrawal et al., 2023, Yang et al., 2025, Cheung and Lyu, 2024]. Here, we assume a single, observable expert, and the central point is the optimal integration of their information with the learner's Bayesian framework.

Active Learning and Information Sources Our results on deciding to learn from an expert echo a form of Bayesian experimental design [Lindley, 1956] and are closely related to Information-Directed

Sampling (IDS), which selects actions to optimize the trade-off between immediate reward and information gain about the optimal action [Russo and Van Roy, 2014]. However, where standard IDS chooses an arm to pull, our agent makes a meta-decision about which data stream to process. This connects to Arumugam and Van Roy [2021] where the authors propose rate distortion to allow online learners to choose samples to learn from. Additionally, there are connections to recent work on regret bounds for online learning from expert feedback [Plaut et al., 2025a,b], where authors study the setting where agents can ask experts for which action is best.

2 Single-Agent Bandit Problem

To first formally define the sequential decision-making problem considered, we introduce some notions and necessary concepts from information theory.

Preliminaries We define our problem on a probability space $(\Omega, \mathcal{F}, P)$ with all quantities including the true parameter of the bandit, and the agent's sampled parameters, actions, and outcomes, being considered random variables on this space. For a discrete random variable X, $\mathbb{E}[X]$ is the expected value of X, and the entropy of a (discrete) random variable X with probability mass function $p(x)$ is $\mathbb{H}(X) := -\sum_{x\in\mathcal{X}} p(x)\log p(x) = \mathbb{E}[-\log p(X)]$. The conditional entropy of X given another random variable Y is $\mathbb{H}(X|Y) := \mathbb{E}[-\log p(X|Y)]$, representing the remaining uncertainty in X once Y is known. The mutual information between X and Y is defined as $\mathbb{I}(X;Y) := \mathbb{H}(X) - \mathbb{H}(X|Y)$. It quantifies the reduction in uncertainty about X resulting from observing Y. Throughout the paper, we use a subscript t to denote conditioning on the history of variables up to time t, $\mathcal{H}_t = \{A_s, Y_s\}_{s<t}$. For instance, the posterior probability of X is denoted $P_t(X) := P(X \mid \mathcal{H}_t)$. Similarly, the conditional entropy of a random variable X given the history is $\mathbb{H}_t(X) := \mathbb{H}(X \mid \mathcal{H}_t)$, and the conditional mutual information is $\mathbb{I}_t(X;Y) := \mathbb{I}(X;Y \mid \mathcal{H}_t)$.

Single Agent Bandit An agent chooses actions $a \in \mathcal{A}$ at every time-step $t \in \mathbb{N}$, with $\mathcal{A}$ being a finite set of actions. Each action produces a (possibly random) outcome $Y_{t,a} \in \mathcal{Y}$, and the agent obtains a reward $R(Y_{t,a})$, with $R : \mathcal{Y} \to \mathbb{R}$. The outcomes are drawn from distributions p_a, of which the agents do not have knowledge of. We assume the outcome distribution $p_\theta := (p_{\theta,a})_{a\in\mathcal{A}}$ to be parameterised by some $\theta \in \Theta$ such that for any action, the (mean) reward is a function of θ, $\mu(a,\theta) := \mathbb{E}_{y\sim p_{\theta,a}}[R(y)]$. For some parameter θ, the *optimal action* $A^* \in \mathcal{A}$ is then the action that satisfies $A^*(\theta) = \arg\max_{a\in\mathcal{A}} \mu(a,\theta)$. The objective of such agent is to maximize the expected cumulative reward (or equivalently, minimize the expected regret relative to the best action). The regret is defined as

$$Reg(T) = \sum_{t=1}^{T} R(Y_t^*) - R(Y_t),$$

where $Y_t^* \sim p_{\theta,A^*}$, and we use $Y_t \equiv Y_{t,A_t}$.

Thompson Sampling Thompson sampling is a Bayesian algorithm for bandit problems that works by sampling actions according to the (posterior) probability that they are the optimal action. Let $\mathcal{H}_t := \{A_t, Y_t\}_{1\le t\le T-1}$ be the history of the actions taken and outcomes observed up to (not including) time T. Thompson Sampling works by assuming the agent samples actions from a posterior distribution (or prior before any new observations) $P(\theta \mid \mathcal{H}_t)$ (abbreviated as $P_t(\theta)$) conditioned on $\mathcal{H}_t$ such that $P_t(A = A^*) = P_t(A = A_t)$. Then, the agent samples a parameter $\hat{\theta}_t \sim P_t(\theta)$, and selects the action that maximises expected rewards *under the model* $\hat{\theta}_t$:

$$A_t \in \arg\max_{a\in\mathcal{A}} \mu(a, \hat{\theta}_t).$$

Then, a new outcome Y_t is observed (when choosing A_t), and the belief $P_t(\theta)$ is updated according to the history $\mathcal{H}_{t+1} = \{\mathcal{H}_t, \{A_t, Y_{A_t}\}\}$ via Bayes' rule:

$$P_{t+1}(\hat{\theta}_t) = P(\hat{\theta}_t \mid \mathcal{H}_{t+1}) \propto p_{\hat{\theta}_t,A_t}(Y_t) P(\hat{\theta}_t \mid \mathcal{H}_t).$$

For Bayesian decision makers, one usually considers the expected regret

$$\mathbb{E}\left[Reg(T)\right] = \mathbb{E}\left[\sum_{t=1}^{T} R(Y_t^*) - R(Y_t)\right],$$

where the expectation is taken with respect to the outcomes and the parameters sampled $\hat{\theta}_t$. We define finally a quantity that will be of use for some of the results in the paper. From Russo and Van Roy [2016], we define the *information ratio* in a Bandit as $\Gamma_t := \mathbb{E}_t \left[Reg(T)\right]^2 / \mathbb{I}_t(A^*, (A_t, Y_t))$. In other words, it is the ratio of the squared expected regret at time t given the past history against the mutual information between the optimal action distribution and the current observation.

3 Learning from Expert Data

Consider the case where one player i has no prior knowledge of the environment, and the second player j is an *expert* (i.e. knows θ^*). Assume player j can share some information with player i. This can manifest via (i) Player i gets an initial dataset $D_N^* = \{Y_n^*\}_{1 \leq n \leq N}$ and (ii) Player i gets to observe new samples Y_t^* as they start learning.

3.1 With Expert Prior Data

Infinite Information To start our analysis, assume first that $N \to \infty$ and we can construct an unbiased density estimator with no errors, or, in other words, player i has access to the likelihood $p^*_{A^*}(Y)$. Treating this as an offline data scenario, we can interpret the knowledge of $p_{A^*}(Y)$ as an observation to be incorporated into the player's knowledge via posterior inference. Intuitively, knowing $p_{A^*}(Y)$ should restrict the set of non-zero likelihood parameters in our posterior to those which satisfy $\tilde{\Theta} := \{\theta \in \Theta : \max_{a \in \mathcal{A}} p_{\theta,a} = p^*_{A^*}\}$. Let $\mathbb{I}[p^*_{A^*} \mid \theta] = 1$ if $\max_{a \in \mathcal{A}} p_{\theta,a} = p^*_{A^*}$. Then, for the posterior to be consistent with the observed data, we want it to satisfy

$$P_1(\theta \mid p^*_{A^*}) \propto P_0(\theta) \mathbb{I}[p^*_{A^*} \mid \theta]. \tag{1}$$

We use P_0 to refer to the initial prior the player has over the parameters Θ, and P_1 as the (offline) posterior resulting from incorporating the expert data. This posterior in (1) will assign zero mass to any parameter θ which induces an optimal action distribution that does not match $p^*_{A^*}$. From the set of parameters that induce such a distribution, we cannot distinguish (have equal likelihood), so the prior will dominate the posterior mass. We show that this posterior update is consistent in the upcoming section, by showing it can be derived as a result of an infinite data limit.

Finite Information Next, consider the case where $N < \infty$, and therefore player i starts with a finite dataset $D_N^* = \{Y_n^*\}_{1 \leq n \leq N}$ of samples from the optimal arm, but cannot identify (yet) what arm these correspond to. Following the intuition in the case of infinite information, one would want to incorporate this off-line information into the prior, to afterwards proceed normally with TS, hopefully with a prior that is better informed.

Recall that, under parameter $\theta \in \Theta$, the likelihood of a given sample Y^* being sampled from the bandit θ is $p_{A^*,\theta}(Y^*)$. Then, given a set of N samples $D_N^* = \{Y_n^*\}_{1 \leq n \leq N}$, we can infer a posterior under the likelihood that the data comes from the current model as

$$P_1(\theta \mid D_N^*) \propto P_0(\theta) p_{A^*,\theta}(D_N^*). \tag{2}$$

Since the expert samples are i.i.d., we can write the right hand side as

$$P_0(\theta) p_{A^*,\theta}(D_N^*) = P_0(\theta) \prod_{i=1}^{N} p_{A^*,\theta}(Y_i^*) = P_0(\theta) \exp\big(\sum_{i=1}^{N} \log p_{A^*,\theta}(Y_i^*)\big). \tag{3}$$

Proposition 1. *Assume a countable set Θ. As the number of samples increases $N \to \infty$, the posterior update in* (3) *converges to the infinite data update in* (1)*. In other words,*

$$\lim_{N \to \infty} P_1(\theta \mid D_N^*) = P_1(\theta \mid p_A^*) \quad \textit{a.s.} \tag{4}$$

Regret Bounds with Offline Expert Data To estimate the Bayesian regret improvement of the agents when having access to offline expert data, let us first define the following concepts. The probability $\mathbb{P}_0(A = A^*)$ under measure P_0 is the probability of A being optimal under the prior distribution $P_0(\theta)$. Let $\Theta_A^* := \{\theta \in \Theta : a = \arg\max_{a' \in \mathcal{A}} \mu(a', \theta)\}$; in other words, Θ_A^* is the subset of parameters that yields a to be the optimal action. Observe we can then write $\mathbb{P}_0(A =$

$A^*) = \int_{\Theta^*_A} P_0(\theta) d\theta$. Then, define $\mathbb{H}_0(A^*)$ to be the entropy of the optimal action distribution under measure P_0, or equivalently:

$$\mathbb{H}_0(A^*) = \sum_{a \in \mathcal{A}} P_0(A = A^*) \log P_0(A = A^*).$$

To prove posterior consistency, we need to show that the probability density of $A = A^*$ under posterior $P_0(\theta \mid D^*_N)$ (denoted as $P_1(A = A^*)$ is equal to the posterior probability $P_0(A = A^* \mid D^*_N)$ under prior $P_0(\theta)$.

Proposition 2. *The probability of an action A being optimal under measure $P(\theta \mid D^*_N)$ is equal to the posterior probability:*

$$P_1(A = A^*) = P_0(A = A^* \mid D^*_N).$$

Russo and Van Roy [2016] established that the Bayesian regret of a Thompson Sampling algorithm is upper bounded by $\sqrt{\mathbb{H}_t(A^*)}$. We can now show that under expert data, the entropy of the (offline) posterior $P_0(\theta \mid D^*_N)$ is guaranteed to decrease in expectation over the observed data.

Theorem 1 (Regret Reduction from Offline Expert Data)**.** *Let TS_0 be a Thompson Sampling agent with prior $P_0(\theta)$, and TS_1 be a Thompson Sampling agent whose prior is the expert-updated posterior $P_1(\theta) = P(\theta \mid D^*_N)$. The expected Bayesian regret of TS_1, taken over all sources of randomness including the expert data D^*_N, is bounded by the regret of TS_0:*

$$\mathbb{E}[Reg_{TS_1}(T)] \leq C\sqrt{T\left(\mathbb{H}_0(A^*) - \mathbb{I}_0(A^*; D^*_N)\right)} \leq \mathbb{E}[Reg_{TS_0}(T)],$$

where C is a problem-dependent constant, $\mathbb{H}_0(A^*)$ is the prior entropy of the optimal arm and $\mathbb{I}_0(A^*; D^*_N)$ is the mutual information between the optimal arm and the expert data under P_0.

Intuitively, this means that if the mutual information between the expert data and the optimal action distribution is high (*i.e.* the expert samples allow the agent to reduce the set of possible parameters to a much smaller subset), then the resulting regret will be significantly lower.

4 Simultaneous Learning: Learning Next to an Expert

Consider now the problem where the player has no prior information on what the optimal arm distribution looks like, but as it learns, it will observe both the (action, outcome) pair (A_t, Y_t) it generated itself and the (optimal) outcome Y^*_t the expert player generated (and thus also knows $R(Y^*_t)$.

In this case, we assume that the observer player can only learn from one sample at the time. Therefore, the player needs to choose at every step t whether they learn from the expert outcome Y^*_t (which does not include action index), or their own sampled pair (A_t, Y_t). We assume the player will still receive it's own reward $R(Y_t)$, and thus the expected instantaneous regret $\mathbb{E}[R(Y^*_t) - R(Y_t)]$ does not depend on the expert sample, or on the agent's choice on which information source to incorporate. This simplifies the analysis of the decision the agent needs to make. From Russo and Van Roy [2016] and Russo and Van Roy [2014], the expected regret of (general) Bayesian online learners is bounded by $\sqrt{\overline{\Gamma}\mathbb{H}_t(A^*)T}$, where $\overline{\Gamma}$ is an upper bound for the information ratio. Given that the agent's choice over what information to incorporate does not change the immediate rewards, this choice needs to be driven by the information gain from each source. Let $D_t \in \{Y^*_t, (Y_t, A_t)\}$ be the random variable representing the data processed at time t, which can be either the expert outcome or the pair (outcome, action) from the player themselves. Then, the choice of data to learn from can be expressed through the choice:

$$\arg\min_{D_t} \ \mathbb{E}[\mathbb{H}_t(A^* \mid D_t)] = \arg\min_{D_t} \ \mathbb{H}_t(A^*) - \mathbb{I}_t(A^*, D_t) = \arg\max_{D_t} \ \mathbb{I}_t(A^*, D_t). \tag{5}$$

In other words, the agent should choose to learn from the sample that maximises the mutual information with the optimal action distribution. This is effectively a Bayesian experimental design framework [Lindley, 1956], where the *experiments* (self-generated data vs. expert data) need to be selected to maximise information gain[1].

[1] Bayesian experimental design is usually framed in terms of the information gain of model parameters θ. In our case, we care about the mutual information between $(A = A^*, D)$.

Conditional Entropy Estimations From (5), the agent can estimate their optimal information source at every time-step based on the estimated conditional entropy from each source. For this, let us write the posterior measures $P_{t+1}(A = A^*)$ resulting from the expert and learner's data,

$$P^e_{t+1}(A = A^*) = \int_\Theta \mathbb{1}[A^*(\theta) = A] P_t(\theta \mid Y^*_t) d\theta \propto \int_\Theta \mathbb{1}[A^*(\theta) = A] p_{\theta, A^*(\theta)}(Y^*_t) P_t(\theta) d\theta,$$
$$P^s_{t+1}(A = A^*) = \int_\Theta \mathbb{1}[A^*(\theta) = A] P_t(\theta \mid Y_t, A_t) d\theta \propto \int_\Theta \mathbb{1}[A^*(\theta) = A] p_{\theta, A_t}(Y_t) P_t(\theta) d\theta. \tag{6}$$

Observe that, although computing posterior densities can be computationally complex, sampling from the posteriors in (6) can be done relatively fast through particle samples.

Remark 1. *Agents following Algorithm 1 to decide where to learn from estimate MI between data and A^* one step at a time. This is bound to introduce bias and variance issues. First, from Jensen's inequality, computing the entropy of an estimated distribution will have a bias. Second, the conditional entropy in* (5) *is taken in expectation over data, but one step ahead the agents only have one sample to compute this estimate. A better approach, if the agents can sustain a buffer, would be to collect a number $N' >> 0$ samples from both the expert and their own experience, and estimate the information gain. We showcase in Section 5 how this is indeed the case, and one-step ahead MI estimation results in learning collapse.*

Algorithm 1 Information Choice: Who To Learn From

Sample $\{\theta^{(n)}\}_{1 \leq n \leq K}$ from $P_t(\theta)$.
for $a \in \mathcal{A}$ **do**
  Approximate $\hat{P}^e_{t+1}(a = A^*) = \frac{1}{K} \sum_n \mathbb{1}[A^*(\theta^{(n)}) = A] p_{\theta^{(n)}, A^*(\theta^{(n)})}(Y^*_t)$.
  Approximate $\hat{P}^s_{t+1}(a = A^*) = \frac{1}{K} \sum_n \mathbb{1}[A^*(\theta^{(n)}) = A] p_{\theta^{(n)}, A_t}(Y_t)$.
end for
Renormalise $\hat{P}^e_{t+1}(a = A^*), \hat{P}^s_{t+1}(a = A^*)$.
Estimate $\hat{\mathbb{H}}^e_t(A^*)$, $\hat{\mathbb{H}}^s_t(A^*)$.
Select $\arg\max_{d \in \{e,s\}} \{\hat{\mathbb{H}}^d_t(A^*)\}$.

4.1 Exploiting Naive Expert Trust

Until now, the analysis has focused on the setting in which the learning agent *fully trusts the expert*; there is an implicit assumption that expert samples are drawn (with full confidence) from the optimal arm distribution p_{A^*}. A natural question that follows is how this can be affected by misaligned, imperfect, or adversarial experts. This can introduce robustness failure modes in agent learning, some of which can be more severe than others. To formalise this, consider the expert is sampling and providing outcomes from some (possibly adversarial) distribution $q \in \Delta(\mathcal{A})$[2]. Take N samples from q, $\{Y^q_n\}_{1 \leq N}$. Recall that since the learner is naive, it still updates its posterior based on the data:

$$P^q_1(\theta) \propto P_0(\theta) \prod_{n=1}^N p_{A^*,\theta}(Y^q_n) = P_0(\theta) \exp\big(\sum_{n=1}^N \log p_{A^*,\theta}(Y^q_n)\big). \tag{7}$$

Observe that this is a specific form of a misspecified Bayesian inference problem; the agent is trying to infer a posterior thinking the data is coming from $p_{A^*,\theta}$, and uses a corresponding likelihood, while the data is in fact sampled from a different q [Nott et al., 2023]. Let us use $l^q_N(\theta) := \frac{1}{N} \sum_{n=1}^N \log p_{A^*,\theta}(Y^q_n)$, and observe that again $l^q_N(\theta) = \mathbb{H}(q) - D_{KL}(q \| p_{A^*,\theta}) + \delta^q_N(\theta)$. The optimal action distribution under the misspecified posterior P^q_1 is[3]

$$P^q_1(a = A^*) = \frac{\int_{\theta \in \Theta^*_a} P_0(\theta) e^{N l^q_N(\theta)} d\theta}{\sum_{b \in \mathcal{A}} \int_{\theta \in \Theta^*_b} P_0(\theta) e^{N l^q_N(\theta)} d\theta} = \frac{\int_{\theta \in \Theta^*_a} P_0(\theta) e^{N(-D_{KL}(q \| p_{A^*,\theta}) + \delta^q_N(\theta))} d\theta}{\sum_{b \in \mathcal{A}} \int_{\theta \in \Theta^*_b} P_0(\theta) e^{N(-D_{KL}(q \| p_{A^*,\theta}) + \delta^q_N(\theta))} d\theta}.$$

For $N \to \infty$, from established misspecified Bayes results [Berk, 1966, Bochkina, 2019] and under mild assumptions (measurability, compact Θ^*_a, $P_0(\theta) > 0$...) the posterior $P^q_1(a = A^*)$ will concentrate probability mass around the set $\Theta_q := \{\theta \in \Theta : \min_\theta D_{KL}(q \| p_{A^*(\theta),\theta})\}$; in other words, the set of parameters that result in an optimal action distribution that is as close as possible to q. We discuss therefore two possible scenarios.

[2] This is a generalisation over previous sections; take $q = p^*_{A^*}$ and we recover the *benign* expert.

[3] The derivation follows the same step as in the proof of Proposition 1.

The expert agent makes mistakes The simplest example of robustness failure is the case where the expert agent provides samples from a mixture $\tilde{Y}_t^* \sim \sum_{a \in \mathcal{A}} w_a p_a$, where $w \in \Delta(\mathcal{A})$ is some mixture vector indicating how often the expert samples from each action. The asymptotic effect on the offline posterior P_1^q will depend on the specific problem instance. For example, take $w_{A^*} = 1 - \epsilon$, and $\sum_{a \in \mathcal{A} \setminus A^*} w_a = \epsilon \approx 0$. If ϵ is small, then θ^* will still be the minimiser $\theta^* = \min_\theta D_{KL}(q \| p_{A^*(\theta),\theta})$. In this case, the posterior will still concentrate around θ^* asymptotically and the agent will learn in the limit, but at a slower rate. For an empirical example on this, see Appendix B.

The expert agent is adversarial A more aggressive example is one where the expert is adversarial (and possibly deceptive), and samples with probability $\epsilon \in [0, 1]$ a true optimal outcome from p_A^*, and with probability $1 - \epsilon$ an *adversarial outcome* that steers the agent's beliefs over θ to the *worse possible parameter* (this is know as the Huber contamination model [Huber, 1992]). In other words, the parameter $\theta^{adv} \in \Theta$ such that $\theta^{adv} := \min_{\theta \in \Theta} \mu(A^*(\theta), \theta^*)$. In this case, depending on the problem instance, there is a threshold ϵ^* after which the agent will inevitably incur linear regret; whenever $D_{KL}((1-\epsilon)p_{A^*(\theta^*)} + \epsilon p_{A^*(\theta^*)} \| p_{A^*(\theta^*)}) \leq D_{KL}((1-\epsilon)p_{A^*(\theta^*)} + \epsilon p_{A^*(\theta^*)} \| p_{A^*(\theta^{adv})})$, the agent will end up being *confidently wrong*. For an empirical example on this, see Appendix B.

5 Experiments

We present now a set of bandit experiments to showcase the results presented in previous sections.

Symmetric Countable Worlds: Countable $\Theta = \{\theta_1, \theta_2, ..., \theta_M\}$ where all bandits have the same set of actions $\mathcal{A}$ with finite supports, but *shuffled*. That is, each bandit will have the same optimal action distribution assigned to a different action. In this case, there is no information gain from expert data.

Asymmetric Countable Worlds: Countable $\Theta = \{\theta_1, \theta_2, ..., \theta_M\}$ where all bandits have the same number of actions with equal support, but the probability distributions $p_{\theta,a}$ are generated at random for each θ, a by adding normally distributed noise to a uniform distribution. That is, every bandit has (similar) but numerically different action distributions. In this case, using expert data should asymptotically lead to zero regret.

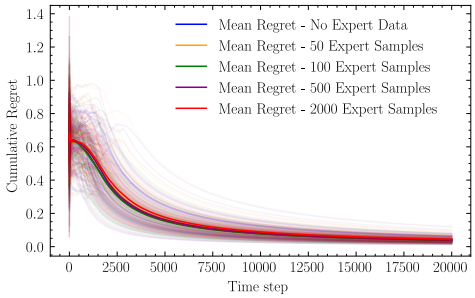
Strongly Asymmetric Countable Worlds: Countable $\Theta = \{\theta_1, \theta_2, ..., \theta_M\}$ where all bandits have the same number of actions with equal support, the probability distributions $p_{\theta,a}$ are generated at random for each θ, a, but we fix the true bandit θ^* to have $p_{A^*,\theta^*}(y^*) = 1$ for some fixed y^* with positive reward. On average, this problem is similarly hard to a traditional Thompson Sampling agent, but an agent learning from expert data should infer with few samples the true θ^*

We fix all experiments to $M = 500$, $|\mathcal{A}| = 50$, $\mathcal{Y} = \{-50, -49, ..., 49, 50\}$, $R(Y) = Y$ is the identity map and unless specifically stated, $\text{supp}(p_{\theta,a}) = \mathcal{Y}$ for all θ, a. We restrict the experiments to countable worlds and finite actions since this allows us to express priors and posteriors with categorical distributions and compute Bayesian updates exactly.

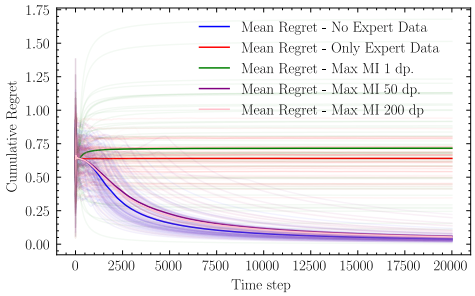
5.1 Symmetric Bandits

We present first the learning results on the symmetric bandits with countable parameter set. We generate $M = 500$ bandit models (parameters) by generating 50 arms from adding random noise to a uniform distribution over $\mathcal{Y}$ and normalizing. Then, we select one model at random from the 500 parameters to be the true model. The prior is $P_0(\theta) = \text{uniform}(\Theta)$ in all cases. We run each scenario with 50 different random seeds and present all runs in transparent color, and the means in thicker opaque lines. In all cases, we plot the cumulated regret rate $Reg(T)/T$.

Results in Symmetric Bandits The results are presented in Figure 1. First, we can see how offline learning with expert samples does not improve the Thompson Sampling regret at all in the symmetric bandit case. Having information over the optimal action distribution does not help when all bandits for any θ have the same optimal action distribution. Second, the fastest learning rate is obtained for the case where the agent only considers their own data at every time-step. Learning from expert data only results in linear regret (no learning). Additionally, as mentioned in Remark 1, we can see how



(a) Regret with priors P_0 computed with 50, 100, 500 and 2000 expert samples.



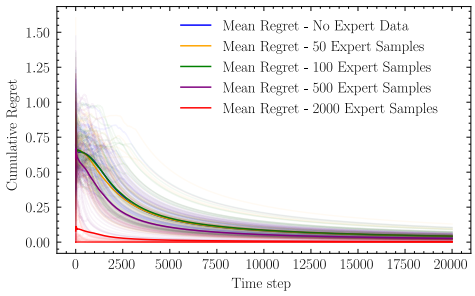
(b) Regret of agents running Algorithm 1, estimating MI from 1, 50 or 200 samples (*dp*).

Figure 1: Regret obtained by TS agents with expert data in symmetric bandits.

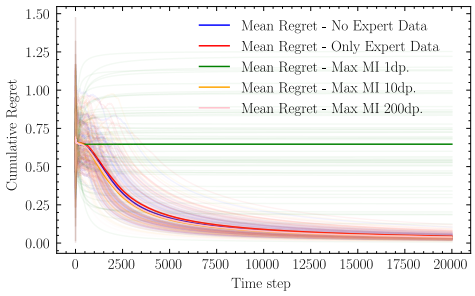
estimating MI and attempting to learn from the maximum information source results in catastrophic performance when the MI is estimated from just 1 sample. As the number of samples increase, the performance gets closer to the no expert data baseline.

5.2 Asymmetric Bandits

We simulated agents with $M = 500$ bandits, where for each θ the distributions $p_{a,\theta}$ are generated as a (renormalised) uniform distribution over $\mathcal{Y}$ with gaussian zero mean noise in each entry. This results in bandit instances that are hard to distinguish, but that have different distributions for each a, θ. In principle, in this case the agent would be able to solve the bandit problem in one step if it had access to infinite expert data.



(a) Regret with priors P_0 computed with 50, 100, 500 and 2000 expert samples.



(b) Regret of agents running Algorithm 1, estimating MI from 1, 50 or 200 samples (*dp*).

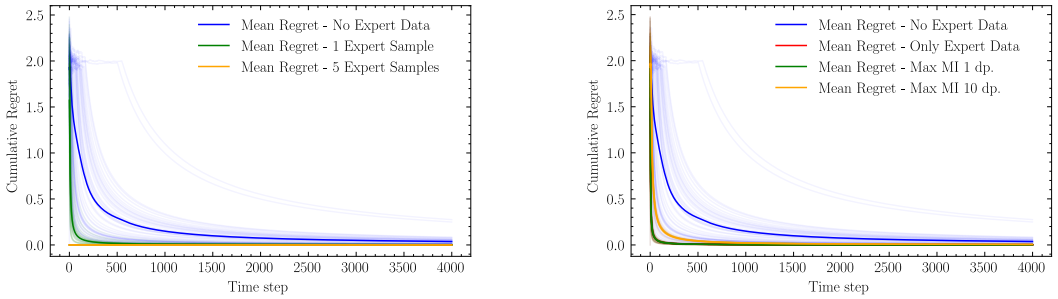
Figure 2: Regret obtained by TS agents with expert data in asymmetric bandits.

Results in Asymmetric Bandits We present the corresponding results in Figure 2. In this case, we can observe how having access to an expert dataset offline yields heavy improvements in total regret when running Thompson Sampling with the resulting posteriors. In the case with 2000 expert samples, the resulting agents achieve almost zero regret from the start of the Thompson Sampling phase. Interestingly, in this case the selection of information source does result in an overall improvement in learning speed. In particular, when comparing the regret rate at low time-steps, the agents running Algorithm 1 with 10 and 200 samples get an improvement of -7% and -4.7% correspondingly with respect to the fastest learning single source agent (No Expert Data)[4].

[4]These values may seem moderate, but they are in fact quite significant considering the overall setting. It means that, across a wide range of randomly generated problem instances, selecting information sources based on past data results in a $\approx 7\%$ learning rate improvement over an (already efficient) Thompson Sampling agent at no additional sampling cost.

5.3 Strongly Asymmetric Bandits

To test the cases where having expert data solves the bandit problem almost immediately, we simulated agents with $M = 500$ bandits, with distributions generated identically to the previous asymmetric experiments, but with one change. Once the true parameter θ^* is selected (at random), one of the action distributions a' is replaced by a (Dirac delta) distribution $p_{a',\theta^*}(2) = 1$. This results in all cases in $a' = A^*$. Since the agent knows the problem class, solving the bandit in a traditional Thompson Sampling approach will still require a (relatively) large amount of steps, but having expert samples would allow the agent to immediately infer θ^*.



(a) Regret with priors P_0 computed with 1 and 5 expert samples.

(b) Regret of agents running Algorithm 1, estimating MI from 1 or 10 samples (*dp*).

Figure 3: Regret obtained by TS agents with expert data in strongly asymmetric bandits.

Results in Strongly Asymmetric Bandits The results are presented in Figure 3. Observe that, in the left hand plot, having a single expert sample to compute an offline prior causes the regret rate to drop almost immediately after a few Thompson Sampling steps. For only 5 expert samples, the resulting offline posterior yields a zero regret Thompson Sampling algorithm for all times in all instances computed. In this case, the improvements in regret rates are dramatic for agents running Algorithm 1. In particular, for agents using a single sample to estimate the MI, after 500 steps the improvement in regret is of -99% when compared to regular Thompson Sampling. This means the agents are successfully able to estimate that the gains in mutual information from the expert source are very beneficial (even with a single sample) and choose to learn from this source at all times[5].

6 Discussion

In this work we studied the problem of Bayesian online learning when agents have access to expert *outcomes*, and are able to use these outcomes to improve their learning. We first propose and formally justify how to use offline expert data to update a Bayesian prior and second, we propose an information-directed algorithm that adaptively chooses between self-generated and expert-provided data and study its implications empirically. We showed empirically how expert information is most valuable in asymmetric worlds where it provides discriminative evidence to prune the parameter space. In such settings, both offline pre-training and our simultaneous learning algorithm dramatically reduce regret. We also highlight a critical limitation: the one-step, one-sample MI estimate is high-variance and can mislead the agent, suggesting that more sophisticated density estimation techniques are necessary for robust performance.

Limitations and Future Work Parts of our analysis were conducted in countable parameter and action spaces, which enabled exact posterior updates. Extending this framework to continuous spaces with function approximation is a significant next step, likely requiring variational or particle-based approximations of mutual information. Additionally, the adversarial expert case deserves a deeper analysis, both theoretical and empirical. Furthermore, our agent model assumes full trust in the expert's optimality. Future work could explore models where the agent also maintains a belief over the expert's reliability, allowing it to explicitly reason about when to trust the provided information.

[5]The agent seems to learn slower for the case where the MI is estimated from 10 samples, but this is an artifact given that the agent chooses their own data until 10 samples have been collected, and only switches then.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Shubhada Agrawal, Sandeep Juneja, Karthikeyan Shanmugam, and Arun Sai Suggala. Optimal best-arm identification in bandits with access to offline data. *arXiv preprint arXiv:2306.09048*, 2023.

Dilip Arumugam and Benjamin Van Roy. Deciding what to learn: A rate-distortion approach. In *International Conference on Machine Learning*, pages 373–382. PMLR, 2021.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Robert H Berk. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

Natalia Bochkina. Bernstein–von mises theorem and misspecified models: A review. *Foundations of modern statistics*, pages 355–380, 2019.

Simina Brânzei and Yuval Peres. Multiplayer bandit learning, from competition to cooperation. In *Conference on Learning Theory*, pages 679–723. PMLR, 2021.

Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, May 1997. ISSN 0004-5411. doi: 10.1145/258128.258179. URL `https://doi.org/10.1145/258128.258179`.

William Chang and Yuanhao Lu. Multiplayer information asymmetric contextual bandits. *arXiv preprint arXiv:2503.08961*, 2025.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Wang Chi Cheung and Lixing Lyu. Leveraging (biased) information: Multi-armed bandits with offline data. *arXiv preprint arXiv:2405.02594*, 2024.

John D Geanakoplos and Heraklis M Polemarchakis. We can't disagree forever. *Journal of Economic theory*, 28(1):192–200, 1982.

Michael Gimelfarb, Scott Sanner, and Chi-Guhn Lee. Reinforcement learning with multiple experts: A bayesian model combination approach. *Advances in neural information processing systems*, 31, 2018.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Yoram Moses and Gal Nachum. Agreeing to disagree after all. In *Proceedings of the 3rd conference on Theoretical aspects of reasoning about knowledge*, pages 151–168, 1990.

David J Nott, Christopher Drovandi, and David T Frazier. Bayesian inference for misspecified generative models. *Annual Review of Statistics and Its Application*, 11, 2023.

Benjamin Plaut, Juan LiÃŠvano-Karim, and Stuart Russell. Asking for help enables safety guarantees without sacrificing effectiveness. *arXiv preprint arXiv:2502.14043*, 2025a.

Benjamin Plaut, Hanlin Zhu, and Stuart Russell. Avoiding catastrophe in online learning by asking for help. In *Forty-second International Conference on Machine Learning*, 2025b.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in neural information processing systems*, 27, 2014.

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

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444. URL `http://www.jstor.org/stable/2332286`.

Le Yang, Vincent YF Tan, and Wang Chi Cheung. Best arm identification with possibly biased offline data. *arXiv preprint arXiv:2505.23165*, 2025.

A Mathematical Proofs

Proposition 1. First, let us write $\sum_{i=1}^{N} \log p_{A^*,\theta}(Y_i^*) = N\big(\frac{1}{N}\sum_{i=1}^{N} \log p_{A^*,\theta}(Y_i^*)\big)$. Now we have

$$N\left(\frac{1}{N}\sum_{i=1}^{N} \log p_{A^*,\theta}(Y_i^*)\right) = N\left(\mathbb{E}_{Y\sim p^*_{A^*}}[\log p_{A^*,\theta}(Y)] + \delta_N(\theta)\right),$$

and $\delta_N(\theta) := \frac{1}{N}\sum_{i=1}^{N} \log p_{A^*,\theta}(Y_i^*) - \mathbb{E}_{Y\sim p^*_{A^*}}[\log p_{A^*,\theta}(Y)]$, which goes to zero *almost surely* as $N \to \infty$ by the law of large numbers. Observe that $\mathbb{E}_{Y\sim p^*_{A^*}}[\log p_{A^*,\theta}(Y)]$ is the cross entropy between $p^*_{A^*}$ and $p_{A^*,\theta}$, and thus

$$\mathbb{E}_{Y\sim p^*_{A^*}}[\log p_{A^*,\theta}(Y)] = -\mathbb{H}(p^*_{A^*}) - D_{KL}(p^*_{A^*}||p_{A^*,\theta}).$$

Then, substituting back in the posterior update,

$$\begin{aligned}
P_1(\theta \mid D_N^*) = \frac{\prod_{i=1}^{N} p_{A^*,\theta}(Y_i^*)P_0(\theta)}{\sum_{\nu\in\Theta}\prod_{i=1}^{N} p_{A^*,\nu}(Y_i^*)P_0(\nu)} = \frac{\exp\big(\sum_{i=1}^{N}\log p_{A^*,\theta}(Y_i^*)\big)P_0(\theta)}{\sum_{\nu\in\Theta}\exp\big(\sum_{i=1}^{N}\log p_{A^*,\nu}(Y_i^*)\big)P_0(\nu)} = \\
= \frac{\exp\big(N\left(\mathbb{E}_{Y\sim p^*_{A^*}}[\log p_{A^*,\theta}(Y)] + \delta_N(\theta)\right)\big)P_0(\theta)}{\sum_{\nu\in\Theta}\exp\left(N\left(\mathbb{E}_{Y\sim p^*_{A^*}}[\log p_{A^*,\nu}(Y)] + \delta_N(\nu)\right)\right)P_0(\nu)} = \\
= \frac{\exp\big(N\left(-\mathbb{H}(p^*_{A^*}) - D_{KL}(p^*_{A^*}||p_{A^*,\theta}) + \delta_N(\theta)\right)\big)P_0(\theta)}{\sum_{\nu\in\Theta}\exp\Big(N\left(-\mathbb{H}(p^*_{A^*}) - D_{KL}(p^*_{A^*}||p_{A^*,\nu}) + \delta_N(\nu)\right)\Big)P_0(\nu)} = \\
= \frac{\exp\big(N\left(-D_{KL}(p^*_{A^*}||p_{A^*,\theta}) + \delta_N(\theta)\right)\big)P_0(\theta)}{\sum_{\nu\in\Theta}\exp\Big(N\left(-D_{KL}(p^*_{A^*}||p_{A^*,\nu}) + \delta_N(\nu)\right)\Big)P_0(\nu)},
\end{aligned}$$

where the last step holds since $\exp(-\mathbb{H}(p^*_{A^*}))^N$ does not depend on θ and it cancels out with the normalization constant. Observe that, from the definition of almost sure convergence, for any $\epsilon > 0$ there exists a $N' < \infty$ such that $\delta_N < \epsilon$ almost surely for any $N > N'$. Then, taking $\epsilon_0 \in (0,1)$ and $\epsilon = \epsilon_0 \min_{\theta\in\Theta\setminus\tilde{\Theta}} D_{KL}(p^*_{A^*}||p_{A^*,\theta})$,

$$\begin{aligned}
&\lim_{N\to\infty}\exp\Big(N\left(-D_{KL}(p^*_{A^*}||p_{A^*,\theta}) + \delta_N(\theta)\right)\Big) = \lim_{N\to\infty}\exp\Big((N+N')\left(-D_{KL}(p^*_{A^*}||p_{A^*,\theta}) + \delta_{N+N'}(\theta)\right)\Big) \leq \\
&\leq \lim_{N\to\infty}\exp\Big((N+N')\left(-(1-\epsilon_0)D_{KL}(p^*_{A^*}||p_{A^*,\theta})\right)\Big) = \exp(-\infty) = 0.
\end{aligned} \tag{8}$$

Now, let us consider the subsets $\tilde{\Theta}$ and $\Theta \setminus \tilde{\Theta}$. First, take $\theta \in \tilde{\Theta}$. For any such theta, the posterior update is

$$\lim_{N\to\infty} P_1(\theta \mid D_N^*) = \lim_{N\to\infty} \frac{\exp\big(N\delta_N(\theta)\big)P_0(\theta)}{\sum_{\nu\in\Theta}\exp\Big(N\left(-D_{KL}(p^*_{A^*}||p_{A^*,\nu}) + \delta_N(\nu)\right)\Big)P_0(\nu)} \quad \forall\,\theta\in\tilde{\Theta}.$$

Dividing the numerator and denominator by $\exp\big(N\delta_N(\theta)\big)$,

$$\lim_{N\to\infty} P_1(\theta \mid D_N^*) = \lim_{N\to\infty} \frac{P_0(\theta)}{\sum_{\nu\in\Theta}\exp\Big(N\left(-D_{KL}(p^*_{A^*}||p_{A^*,\nu}) + \delta_N(\nu) - \delta_N(\theta)\right)\Big)P_0(\nu)} \quad \forall\,\theta\in\tilde{\Theta}.$$

First, any term in the denominator with $\nu \in \tilde{\Theta}$ has the same likelihood function for the optimal action. Therefore, $\delta_N(\nu) - \delta_N(\theta) = 0\ a.s.$ for any $\nu, \theta \in \tilde{\Theta}$. Second, by the same argument as (8), any term $\nu \notin \tilde{\Theta}$ goes to zero. Therefore,

$$\begin{aligned}
&\lim_{N\to\infty} \frac{P_0(\theta)}{\sum_{\nu\in\Theta}\exp\Big(N\left(-D_{KL}(p^*_{A^*}||p_{A^*,\nu}) + \delta_N(\nu) - \delta_N(\theta)\right)\Big)P_0(\nu)} = \\
&= \frac{1}{\sum_{\nu\in\tilde{\Theta}} P_0(\nu)} P_0(\theta) \quad \forall\,\theta\in\tilde{\Theta}.
\end{aligned}$$

Now consider $\theta \in \Theta \setminus \tilde{\Theta}$. Pick an arbitrary reference $\nu_0 \in \tilde{\Theta}$. We can bound the limit fraction as:

$$\lim_{N\to\infty} \frac{\exp\left(N\left(-D_{KL}(p^*_{A^*}||p_{A^*,\theta}) + \delta_N(\theta)\right)\right)P_0(\theta)}{\sum_{\nu\in\Theta}\exp\Big(N\left(-D_{KL}(p^*_{A^*}||p_{A^*,\nu}) + \delta_N(\nu)\right)\Big)P_0(\nu)} \leq$$

$$\leq \lim_{N\to\infty} \frac{\exp\left(N\left(-D_{KL}(p^*_{A^*}||p_{A^*,\theta}) + \delta_N(\theta)\right)\right)P_0(\theta)}{\exp\Big(N\delta_N(\nu_0)\Big)P_0(\nu_0)} \quad \forall\, \theta \in \Theta \setminus \tilde{\Theta}.$$

Now, re-arranging terms,

$$\lim_{N\to\infty} \frac{\exp\left(N\left(-D_{KL}(p^*_{A^*}||p_{A^*,\theta}) + \delta_N(\theta)\right)\right)P_0(\theta)}{\exp\Big(N\delta_N(\nu_0)\Big)P_0(\nu_0)} =$$

$$= \lim_{N\to\infty} \exp\left(N\left(-D_{KL}(p^*_{A^*}||p_{A^*,\theta}) - \delta_N(\nu_0) + \delta_N(\theta)\right)\right)\frac{P_0(\theta)}{P_0(\nu_0)}.$$

By the same argument as (8), the exponent limit goes to zero, and thus

$$\lim_{N\to\infty} \frac{\exp\left(N\left(-D_{KL}(p^*_{A^*}||p_{A^*,\theta}) + \delta_N(\theta)\right)\right)P_0(\theta)}{\sum_{\nu\in\Theta}\exp\Big(N\left(-D_{KL}(p^*_{A^*}||p_{A^*,\nu}) + \delta_N(\nu)\right)\Big)P_0(\nu)} \leq 0 \quad \forall\, \theta \in \Theta \setminus \tilde{\Theta}.$$

This completes the proof, and we have

$$\lim_{N\to\infty} P_1(\theta \mid D^*_N) = \frac{\mathbb{I}[p^*_{A^*} \mid \theta]}{\sum_{\nu\in\tilde{\Theta}} P_0(\nu)} P_0(\theta).$$

□

Proposition 2. The event $A = A^*$ under the posterior density can be written as

$$P_1(A = A^*) = \int_{\Theta^*_A} P_0(\theta \mid D^*_N)d\theta = \int_{\Theta^*_A} \frac{p_{A,\theta}(D^*_N)P_0(\theta)}{\int_\Theta p_{A,\nu}(D^*_N)P_0(\nu)d\nu}d\theta = \frac{\int_{\Theta^*_A} p_{A,\theta}(D^*_N)P_0(\theta)d\theta}{\int_\Theta p_{A,\nu}(D^*_N)P_0(\nu)d\nu}. \tag{9}$$

Now observe, by Bayes' Theorem,

$$P_0(A = A^* \mid D^*_N) = \frac{P(D^*_N \mid A = A^*)P_0(A = A^*)}{P_0(D^*_N)}. \tag{10}$$

The likelihood $P(D^*_N \mid A = A^*)$ represents the likelihood of the expert samples given that A is optimal. This can be computed by marginalising over all parameters that make A optimal,

$$P(D^*_N \mid A = A^*) = \int_\Theta p(D^*_N \mid \theta)P_0(\theta \mid A = A^*)d\theta = \int_{\Theta^*_A} p_{A,\theta}(D^*_N)\frac{P_0(\theta)}{P_0(A = A^*)}d\theta \implies$$

$$\implies P(D^*_N \mid A = A^*)P_0(A = A^*) = \int_{\Theta^*_A} p_{A,\theta}(D^*_N)P_0(\theta)d\theta, \tag{11}$$

where the last equality is simply a conditional probability relation, and the last step holds since $P(A = A^*)$ does not depend on θ. Therefore, substituting (10) and (11) in (9):

$$P_1(A = A^*) = \frac{\int_{\Theta^*_A} p_{A,\theta}(D^*_N)P_0(\theta)d\theta}{\int_\Theta p_{A,\nu}(D^*_N)P_0(\nu)d\nu} = \frac{P(D^*_N \mid A = A^*)P_0(A = A^*)}{\int_\Theta p_{A,\nu}(D^*_N)P_0(\nu)d\nu} = P_0(A = A^* \mid D^*_N).$$

□

Theorem 1 (Regret Reduction from Offline Expert Data). The result follows directly from the information-theoretic analysis of Russo and Van Roy [2016], which bounds the Bayesian regret of

a Thompson Sampling agent by the entropy of the optimal arm under its current belief distribution. Agent TS_1 has belief P_1. The regret, conditioned on a specific realization of D_N^*, is bounded by:

$$\mathbb{E}[Reg_{TS_1}(T) \mid D_N^*] \leq C\sqrt{T \cdot \mathbb{H}_1(A^*)}$$

To find the unconditional expected regret, we take the expectation over the expert data $D_N^* \sim p_{A^*}(\theta^*)$, where the uncertainty about θ^* is captured by the prior P_0:

$$\mathbb{E}[Reg_{TS_1}(T)] = \mathbb{E}_{D_N^*}\left[\mathbb{E}[Reg_{TS_1}(T) \mid D_N^*]\right] \leq \mathbb{E}_{D_N^*}\left[C\sqrt{T \cdot \mathbb{H}_1(A^*)}\right].$$

Applying Jensen's inequality we have $\mathbb{E}[\sqrt{X}] \leq \sqrt{\mathbb{E}[X]}$, which gives:

$$\mathbb{E}[Reg_{TS_1}(T)] \leq C\sqrt{T \cdot \mathbb{E}_{D_N^*}[\mathbb{H}_1(A^*)]}.$$

From Proposition 2, we have $P_1(A = A^*) = P_0(A = A^* \mid D_N^*)$. Therefore, the entropy $\mathbb{H}_1(A^*)$ is precisely the conditional entropy of the optimal arm given the expert data, under the original measure P_0:

$$\mathbb{H}_1(A^*) = -\sum_{a \in \mathcal{A}} P_1(A = A^*) \log P_1(A = A^*) = -\sum_{a \in \mathcal{A}} P_0(A = A^* \mid D_N^*) \log P_0(A = A^* \mid D_N^*) =: \mathbb{H}_0(A^* \mid D_N^*).$$

Substituting this into the bound, we get:

$$\mathbb{E}[Reg_{TS_1}(T)] \leq C\sqrt{T \cdot \mathbb{E}_{D_N^*}[\mathbb{H}_0(A^* \mid D_N^*)]}.$$

Finally, from the definition of mutual information: $\mathbb{E}_{D_N^*}[\mathbb{H}_0(A^* \mid D_N^*)] = \mathbb{H}_0(A^*) - \mathbb{I}_0(A^*; D_N^*)$. This yields the main result:

$$\mathbb{E}[Reg_{TS_1}(T)] \leq C\sqrt{T\left(\mathbb{H}_0(A^*) - \mathbb{I}_0(A^*; D_N^*)\right)}.$$

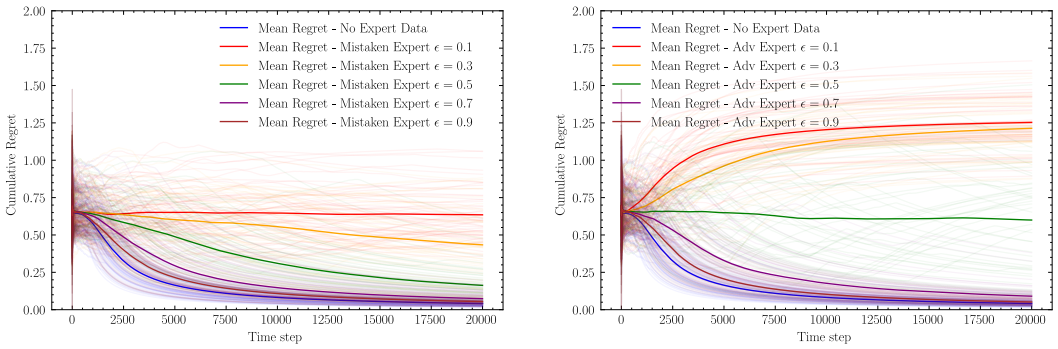
This completes the proof. □

B Adversary Experiments

We include here empirical results on the adversarial cases described in Section 4.1. We use the same asymmetric countable world setting as in Section 5. We compute results for the following:

- A scenario with a 'mistaken' expert, where the expert samples with probability ϵ a true optimal outcome and samples with probability $1 - \epsilon$ an outcome from a uniform action distribution over $\mathcal{A} \setminus A^*$.
- A scenario with an 'adversarial' expert, where the expert samples with probability ϵ a true optimal outcome and samples with probability $1 - \epsilon$ an outcome from an optimal action in an adversarial world θ^{adv}.

Results for Adversary Experiments As discussed in Section 4.1, we can see how the mistaken expert, in the worst case, induces no improvement of regret, which is reasonable since it samples from all actions uniformly. As ϵ increases, the only minimiser in Θ_q becomes θ^* since this is an asymmetric bandit class. Then, the cumulated regret still converges to zero, but at a much slower rate. For the adversarial expert results, we can see how for low ϵ the regret actually increases away from the mean 'uninformed' initial value; the expert forces the agent to believe it lives in a completely different world θ^{adv}. Similarly, as ϵ increases, the set Θ_q becomes a singleton (θ^*) and the agent still manages to achieve zero regret.



(a) Regret when learning from mistaken expert data. (b) Regret when learning from adversarial expert data.

Figure 4: Regret obtained by TS agents with mistaken or adversarial expert data.