Bayesian Design Principles for Frequentist Sequential Learning

Yunbei Xu[1] **Assaf Zeevi**[1]

Abstract

We develop a general theory to optimize the frequentist regret for sequential learning problems, where efficient bandit and reinforcement learning algorithms can be derived from unified Bayesian principles. We propose a novel optimization approach to create "algorithmic beliefs" at each round, and use Bayesian posteriors to make decisions. This is the first approach to make Bayesian-type algorithms prior-free and applicable to adversarial settings, in a generic and optimal manner. Moreover, the algorithms are simple and often efficient to implement. As a major application, we present a novel algorithm for multi-armed bandits that achieves the "best-of-all-worlds" empirical performance in the stochastic, adversarial, and non-stationary environments. And we illustrate how these principles can be used in linear bandits, convex bandits, and reinforcement learning.

1. Introduction

1.1. Background

We address a broad class of sequential learning problems in the presence of *partial feedback*, which arise in numerous application areas including personalized recommendation (Li et al., 2010), game playing (Silver et al., 2016) and control (Mnih et al., 2015). An agent sequentially chooses among a set of possible decisions to maximize the cumulative reward. By "partial feedback" we mean the agent is only able to observe the feedback of her chosen decision, but does not generally observe what the feedback would be if she had chosen a different decision. For example, in multi-armed bandits (MAB), the agent can only observe the reward of her chosen action, but does not observe the rewards of other actions. In reinforcement learning (RL), the agent is only able to observe her state insofar as the chosen action is concerned, while other possible outcomes are not observed and the underlying state transition dynamics are unknown. In this paper, we present a unified approach that applies to bandit problems, RL, and beyond.

The central challenge for sequential learning with partial feedback is to determine the optimal trade-off between *exploration* and *exploitation*. That is, the agent needs to try different decisions to learn the environment; at the same time, she wants to focus on "good" decisions that maximize her payoff. There are two basic approaches to study such exploration-exploitation trade-off: frequentist and Bayesian. One of the most celebrated examples of the frequentist approach is the family of Upper Confidence Bound (UCB) algorithms (Lai et al., 1985; Auer et al., 2002a). Here, the agent typically uses sample average or regression to estimate the mean rewards; and she optimizes the upper confidence bounds of the mean rewards to make decisions. Another widely used frequentist algorithm is EXP3 (Auer et al., 2002b) which was designed for adversarial bandits; it uses inverse probability weighting (IPW) to estimate the rewards, and then applies exponential weighting to construct decisions. One of the most celebrated examples of the Bayesian approach is Thompson Sampling (TS) with a pre-specifed, fixed prior (Thompson, 1933). Here, the agent updates the Bayesian posterior at each round to learn the environment, and she uses draws from that posterior to optimize decisions.

The advantage of the frequentist approach is that it does not require a priori knowledge of the environment. However, it heavily depends on a case-by-case analysis exploiting special structure of a particular problem. For example, regression-based approaches can not be easily extended to adversarial problems; and IPW-type estimators are only known for simple rewards such as discrete and linear. The advantage of Bayesian approach is that Bayesian posterior is a generic and often optimal estimator if the prior is known. However, the Bayesian approach requires knowing the prior at the inception, which may not be accessible in complex or adversarial environments. Moreover, maintaining posteriors is computationally expensive for most priors.

In essence, frequentist approach requires less information, but is less principled, or more bottom-up. On the other hand, the Bayesian approach is more principled, or top-down, but

[1] Graduate School of Business, Columbia University, New York, New York, USA. Correspondence to: Yunbei Xu <yunbei.xu@gsb.columbia.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

requires stronger assumptions. In this paper we focus on the following research:

Can we design principled Bayesian-type algorithms, that are prior-free, computationally efficient, and work well in both stochastic and adversarial/non-stationary environments?

1.2. Contributions

In this paper, we synergize frequentist and Bayesian approaches to successfully answer the above question, through a novel idea that creates "algorithmic beliefs" that are generated sequentially in each round, and uses Bayesian posteriors to make decisions. Our contributions encompass over theoretical discoveries, novel methodology, and applications thereby. We summarize the main contributions as follows.

Making Bayesian-type algorithms prior-free and applicable to adversarial settings. To the best of our knowledge, we provide the first approach that allows Bayesian-type algorithms to operate without prior assumptions and be applicable in adversarial settings, in a generic, optimal, and often computationally efficient manner. The regret bounds of our algorithms are no worse than the best theoretical guarantees known in the literature. It is worth noting that the main ideas underlying our methodology and proofs are quite insightful and can be explained in a succinct manner.

General theory of "Algorithmic Information Ratio" (AIR). We introduce an objective function that depends on an "algorithmic belief" and round-dependent information, which we refer to as "Algorithmic Information Ratio" (AIR). Our approach always selects algorithmic beliefs by (approximately) maximizing AIR, and the regret of our algorithms can always be bounded by the cumulative sum of the values of AIR at each round. We then show that AIR can be upper bounded by previously known complexity measures such as information ratio (IR) (Russo & Van Roy, 2016) and decision-estimation coefficient (DEC) (Foster et al., 2021). As an immediate consequence, our machinery converts existing non-constructive results using information ratio and DEC, into concrete frequentist algorithms. And we provide methods and guarantees to approximately maximize AIR.

"Best of all worlds" empirical performance for MAB As a major illustration, we propose a novel algorithm for Bernoulli multi-armed bandits (MAB) that achieves the "best-of-all-worlds" empirical performance in stochastic, adversarial, and non-stationary environments. This algorithm is quite different from and performs much better than the traditional EXP3 algorithm, which has been the default choice for adversarial MAB for decades. At the same time, the algorithm outperforms UCB and is comparable to Thompson Sampling in the stochastic environment. Moreover, it outperforms traditional Thompson Sampling and "clairvoyant" restarted algorithms in non-stationary environments.

Applications to linear bandits, convex bandits, and RL. For adversarial linear bandits, we derive a modified version of EXP2 based on our framework, which establishes a novel connection between inverse propensity weighting (IPW) and Bayesian posteriors. For bandit convex optimization, we propose the first algorithm that attains the best-known $\tilde{O}(d^{2.5}\sqrt{T})$ regret with a finite $\text{poly}(e^d \cdot T)$ running time. Lastly, we provide a generic closed-form algorithm that is near-optimal for a broad class of reinforcement learning problems in the stochastic setting. We will briefly overview these applications in Section 6 and leave most of the details into Appendix due to space considerations.

Combining estimation and decision-making. Our approach is the first efficient approach to jointly optimize the belief of an environment and probability of decision. Most existing algorithms including UCB, EXP3, Estimation-to-Decision (E2D) (Foster et al., 2021), TS, and Information-Directed Sampling (IDS) (Russo & Van Roy, 2018) maintain a different viewpoint that separates algorithm design into a black-box estimation method (sample average, linear regression, IPW, Bayesian posterior...) and a decision-making rule that makes the estimate as input to an optimization problem. In contrast, by optimizing AIR to generate new beliefs, our algorithm *simultaneously* deals with estimation and optimization. This viewpoint is quite powerful and broadens the general scope of bandit algorithms.

1.3. Related literature

(Russo & Van Roy, 2016; 2018) propose the concept of "information ratio" to analyze and design Bayesian bandit algorithms. Their work studies Bayesian regret with a known prior rather than the frequentist regret. (Lattimore & Gyorgy, 2021) proposes an algorithm called "Exploration by Optimization (EBO)," which is the first general frequentist algorithm that optimally bounds the frequentist regret of bandit problems using information ratio. However, the EBO algorithm is more of a conceptual construct as it requires intractable optimization over the complete class of "functional estimators," and hence is not implementable in most settings of interest. Our algorithms are inspired by EBO, but are simpler in structure and run in decision and model spaces (rather than intractable functional spaces). In particular, our approach advances EBO by employing explicit construction and randomization of estimators, offering flexibility in selecting updating rules, and providing computation guidelines that come with provable guarantees. The recent work (Foster et al., 2021) proposes the concept of "decision-estimation coefficient" (DEC) as a general complexity measure for bandit and reinforcement learning problems. Algorithms in this work typically separate black-box estimation method and decision-making rule, and for this reason the proposed E2D algorithm do not generally achieve optimal regret for bandit problems. The subsequent work

(Foster et al., 2022b) extends the theory of DEC to adversarial environments. However, the algorithm is an adaptation of EBO in (Lattimore & Gyorgy, 2021), which, as discussed, may present computational challenges.

2. Preliminaries and Definition of AIR

2.1. Problem formulation

To state our results in the broadest manner, we adopt the general formulation of Adversarial Decision Making with Structured Observation (Adversarial DMSO) (Foster et al., 2022b), which covers broad problems including bandit problems, reinforcement learning, and partial monitoring. For a locally compact metric space we denote by $\Delta(\cdot)$ the set of Borel probability measures on that space. Let Π be a compact decision space. Let $\mathcal{M}$ be a compact model class where each model $M : \Pi \to \mathcal{O}$ is a mapping from the decision space to a locally compact observation space $\mathcal{O}$. A problem instance in this protocol can be described by the decision space Π and the model class $\mathcal{M}$. We define the mean reward function associated with model M by f_M.

Consider a $T-$round game played by a randomized player in an adversarial environment. At each round $t = 1, \dots, T$, the agent determines a probability p_t over the decisions, and the environment selects a model $M_t \in \mathcal{M}$. Then the decision $\pi_t \sim p_t$ is sampled and an observation $o_t \sim M_t(\pi_t)$ is revealed to the agent. An *admissible algorithm* ALG can be described by a sequence of mappings where the $t-$th mapping maps the past decision and observation sequence $\{\pi_i, o_i\}_{i=1}^{t-1}$ to a probability p_t over decisions. The *frequentist regret* of the algorithm ALG against the usual target of single best decision in hindsight is defined as

$$\mathfrak{R}_T = \sup_{\pi^* \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} f_{M_t}(\pi^*) - \sum_{t=1}^{T} f_{M_t}(\pi_t)\right],$$

where the expectation is taken with respect to the randomness in decisions and observations. There is a large literature that focuses on the so-called stochastic environment, where $M_t = M^* \in \mathcal{M}$ for all rounds, and the single best decision $\pi^* \in \arg\min f_{M^*}(\pi)$ is the natural oracle. Regret bounds for adversarial sequential learning problems naturally apply to stochastic problems. We illustrate how the general formulation covers bandit problems, and leave the discussion of reinforcement learning to Section C.

Example 2.1 (Bernoulli multi-armed bandits (MAB))**.** We illustrate how the general formulation reduces to the basic MAB problem with Bernoulli reward. Let $\Pi = [K] = \{1, \cdots, K\}$ be a finite set of K actions, and $\mathcal{F}$ be the set of all possible mappings from $[K]$ to $[0, 1]$. Take $\mathcal{M} = \{M_f : f \in \mathcal{F}\}$ as the induced model class, where each M_f maps π into the Bernoulli distribution $\text{Bern}(f(\pi))$. The mean reward function for model M_f is f itself. At each round t, the environment selects a mean reward function f_t, and the observation o_t is the incurred reward $r_t \sim \text{Bern}(f_t(\pi_t))$.

Example 2.2 (Structured bandits)**.** We consider bandit problems with general structure of the mean reward function. Let Π be a $d-$dimensional action set, and $\mathcal{F} \subseteq \{f : \Pi \to [0, 1]\}$ be a function class that encodes the structure of the mean reward function. Take $\mathcal{M} = \{M_f : f \in \mathcal{F}\}$ as the induced model class, where each M_f maps π to the Bernoulli distribution $\text{Bern}(f(\pi))$. The mean reward function for model M_f is f itself. For example, in $d-$dimensional linear bandits, the mean reward function f is parametrized by some $\theta \in \Theta \subseteq \mathbb{R}^d$ such that $f(\pi) = \theta^T \pi, \forall \pi \in \Pi$. And in bandit convex optimization, the mean reward (or loss) function class $\mathcal{F}$ is the set of all concave (or convex) mappings from Π to $[0, 1]$.

2.2. Algorithmic Information Ratio

Let ν be a probability measure of the joint random variable $(M, \pi^*) \in \mathcal{M} \times \Pi$, and p be a distribution of another independent random variable $\pi \in \Pi$. Given a probability measure ν, let

$$\nu_{\pi^*}(\cdot) = \int_{\mathcal{M}} \nu(M, \cdot) dM$$

be the marginal distribution of $\pi^* \in \Pi$. Viewing ν as a prior belief over (M, π^*), drawing decision π independently and observation $o \sim M(\pi)$, we define $\nu_{\pi^*|\pi,o}(\cdot)$ as the marginal posterior belief of π^* conditioned on decision π and observation o. Denote $\text{KL}(\mathbb{P}, \mathbb{Q}) = \int \log \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{P}$ as the KL divergence between two probability measures.

Now we introduce a central definition in this paper—Algorithmic Information Ratio.

Definition 2.3 (Algorithmic Information Ratio)**.** Given a reference probability $q \in \text{int}(\Delta(\Pi))$ and learning rate $\eta > 0$, we define the "Algorithmic Information Ratio" (AIR) for decision p and distribution ν as

$$\texttt{AIR}_{q,\eta}(p, \nu) = \mathbb{E}_{p,\nu}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{\eta}\text{KL}(\nu_{\pi^*|\pi,o}, q)\right],$$

where the expectation is taken with $\pi \sim p, (\mathcal{M}, \pi^*) \sim \nu$.

A key aspect of AIR is the reference probability distribution q in its definition. Note that AIR is linear with respect to p and concave with respect to ν, as conditional entropy is always concave with respect to the joint probability measure (see Lemma H.1). It will be illustrative to write AIR as the sum of three items:

$$\texttt{AIR}_{q,\eta}(p, \nu) = \underbrace{\mathbb{E}_{p,\nu}\left[f_M(\pi^*) - f_M(\pi)\right]}_{\text{expected regret}}$$
$$-\frac{1}{\eta}\underbrace{\mathbb{E}_{p,\nu}\left[\text{KL}(\nu_{\pi^*|\pi,o}, \nu_{\pi^*})\right]}_{\text{information gain}} - \frac{1}{\eta}\underbrace{\text{KL}(\nu_{\pi^*}, q)}_{\text{regularization by } q},$$

where: the “expected regret” measures the difficulty of exploitation; “information gain” is the amount of information gained about π^* by observing π and o, and this in fact measures the degree of exploration; and the last “regularization” term forces the marginal distribution of π^* to be “close” to the reference probability distribution q. By maximizing AIR, we generate an “algorithmic belief” that simulates the worst-case environment. This algorithmic belief will automatically balance exploration and exploitation, as well as being close to the chosen reference belief (e.g., a standard reference is the posterior from previous round, as used in traditional Thompson Sampling).

2.3. Relation to IR and DEC.

We present here the traditional definition of Bayesian information ratio (IR) (Russo & Van Roy, 2016).

Definition 2.4 (Information ratio)**.** Given belief ν of (M, π^*) and decision probability p of π, the information ratio is defined as

$$\texttt{IR}(\nu, p) = \frac{\left(\mathbb{E}_{\nu,p}\left[f_M(\pi^*) - f_M(\pi)\right]\right)^2}{\mathbb{E}_{\nu,p}\left[\text{KL}(\nu_{\pi^*|\pi,o}, \nu_{\pi^*})\right]}. \qquad (1)$$

Note that the traditional information ratio (1) does not involve any reference probability distribution q (unlike AIR). By completing the square, it is easy to show that AIR can always be bounded by IR as follows.

Lemma 2.5 (Bounding AIR by IR)**.** *For any $q \in int(\Delta(\Pi))$, $p \in \Delta(\Pi)$, belief $\nu \in \Delta(\mathcal{M} \times \Pi)$, and $\eta > 0$, we have*

$$\texttt{AIR}_{q,\eta}(p, \nu) \le \frac{\eta}{4} \cdot \texttt{IR}(\nu, p).$$

The recent paper (Foster et al., 2021) introduced DEC as a novel complexity measure, aiming to unify bandits and various reinforcement learning problems. We demonstrate that AIR can be bounded by DEC in a manner, see Appendix A for detailed arguments and discussion. Notably, our framework allows for the utilization of nearly all existing upper bounds for IR and DEC in practical applications, enabling the derivation of the sharpest regret bounds known, along with the development of constructive algorithms.

3. Algorithms

3.1. A generic regret bound leveraging AIR

Given an arbitrary admissible algorithm ALG (defined in Section 2.1), we can generate a sequence of *algorithmic beliefs* $\{\nu_t\}_{t=1}^T$ and a corresponding sequence of *reference probabilities* $\{q_t\}_{t=1}^T$ in a sequential manner as shown in Algorithm 1. Maximizing AIR to create algorithmic beliefs is an alternative approach to traditional estimation procedures, as the resulting algorithmic beliefs will simulate the true or worst-case environment. In particular, this approach only stores a single distribution $(\nu_t)_{\pi^*|\pi_t,o_t}$ at round t, which is the Bayesian posterior obtained from belief ν_t and observations π_t, o_t, and it is made to forget all the rest information from the past.

Algorithm 1 Maximizing AIR to create algorithmic beliefs

```
Input algorithm ALG and learning rate $\eta > 0$.
Initialize $q_1$ to be the uniform distribution over $\Pi$.
for round $t = 1, 2, \cdots, T$ do
    Obtain $p_t$ from ALG. Find a distribution $\nu_t$ of $(M, \pi^*)$ that solves
        $\sup_{\nu \in \Delta(\mathcal{M}\times\Pi)} \texttt{AIR}_{q_t,\eta}(p_t, \nu)$.
    The algorithm ALG samples decision $\pi_t \sim p_t$ and observes the feedback $o_t \sim M_t(\pi_t)$.
    Update $q_{t+1} = (\nu_t)_{\pi^*|\pi_t,o_t}$.
end for
```

Based on these algorithmic beliefs, we can provide regret bound for an arbitrary algorithm. Here we assume Π to be finite (but potentially large) for simplicity; this assumption can be relaxed using standard discretization and covering arguments.

Theorem 3.1 (Generic regret bound for arbitrary learning algorithm)**.** *Given a finite decision space Π, a compact model class $\mathcal{M}$, the regret of an arbitrary learning algorithm* ALG *is bounded as follows, for all $T \in \mathbb{N}_+$,*

$$\mathfrak{R}_T \le \frac{\log|\Pi|}{\eta} + \sum_{t=1}^{T} \texttt{AIR}_{q_t,\eta}(p_t, \nu_t). \qquad (2)$$

Note that Theorem 3.1 provides a powerful tool to study the regret of an arbitrary algorithm using the concept of AIR. More importantly, it suggests that the algorithm should choose decision with probability p_{t+1} according to the posterior $((\nu_t)_{\pi^*|\pi_t,o_t}$. Building on this principle to generate algorithmic beliefs, we provide two concrete algorithms: “Adaptive Posterior Sampling” (APS) and “Adaptive Minimax Sampling” (AMS). Surprisingly, their regret bounds are as sharp as the best known regret bounds of existing Bayesian algorithms that *require* knowledge of a well-specified prior.

3.2. Adaptive Posterior Sampling (APS)

When the agent always selects p_{t+1} to be equal to the posterior $q_{t+1} = (\nu_t)_{\pi^*|\pi_t,o_t}$, and optimizes for algorithmic beliefs as in Algorithm 1, we call the resulting algorithm “Adaptive Posterior Sampling” (APS).

At round t, APS inputs p_t to the objective $\texttt{AIR}_{p_t,\eta}(p_t, \nu)$

Algorithm 2 Adaptive Posterior Sampling (APS)

```
Input learning rate $\eta > 0$.
Initialize $p_1 = \mathrm{Unif}(\Pi)$.
for round $t = 1, 2, \cdots, T$ do
    Find a distribution $\nu_t$ of $(M, \pi^*)$ that solves
        $$\sup_{\nu \in \Delta(\mathcal{M} \times \Pi)} \mathtt{AIR}_{p_t,\eta}(p_t, \nu).$$
    Sample decision $\pi_t \sim p_t$ and observe $o_t \sim M_t(\pi_t)$.
    Update $p_{t+1} = (\nu_t)_{\pi^*|\pi_t,o_t}$.
end for
```

to optimize for the algorithmic belief ν_t; and it sets p_{t+1} to be the Bayesian posterior obtained from belief ν_t and observations π_t, o_t. Unlike traditional TS, APS does not require knowing the prior or stochastic environment; instead, APS creates algorithmic beliefs "on the fly" to simulate the worst-case environment. We can prove the following theorem using the regret bound (2) in Theorem 3.1 and the relationship between AIR and IR established in Lemma 2.5.

Theorem 3.2 (Regret of APS)**.** *Assume that $f_M(\pi) \in [0, 1]$ for all $M \in \mathcal{M}$ and $\pi \in \Pi$. The regret of Algorithm 2 with $\eta = \sqrt{2 \log |\Pi| / (\mathtt{IR}_{\mathrm{H}}(\mathrm{TS}) \cdot T + 4T)}$ is bounded as follows, for all $T \geq 2 \log |\Pi| \mathtt{IR}_{\mathrm{H}}(\mathrm{TS}) + 4$,*

$$\mathfrak{R}_T \leq \sqrt{\log |\Pi| \left(\mathtt{IR}_{\mathrm{H}}(\mathrm{TS})/2 + 2\right) T},$$

where $\mathtt{IR}_{\mathrm{H}}(\mathrm{TS}) := \sup_\nu \mathtt{IR}_{\mathrm{H}}(\nu, \nu_{\pi^*})$[1] is the maximal value of information ratio for Thompson Sampling.

For $K-$armed bandits, APS achieves the near-optimal regret $O(\sqrt{KT \log K})$ because $\mathtt{IR}_{\mathrm{H}}(\mathrm{TS}) \leq K$; for $d-$dimensional linear bandits, APS recovers the optimal regret $O(\sqrt{d^2 T})$ because $\mathtt{IR}_{\mathrm{H}}(\mathrm{TS}) \leq d$.

The main messages about APS and Theorem 3.2 are: 1) the regret bound of APS is no worse than the standard regret bound of TS (Russo & Van Roy, 2016), but in contrast to the latter, does not rely on any knowledge needed to specify a prior! 2) Because APS only keeps the marginal beliefs of π^* but forgets beliefs of the models, it is robust to adversarial and non-stationary environments. And 3) Experimental results in Section 4 show that APS achieves "best-of-all-worlds" empirical performance for Bernoulli MAB in different environments.

To the best of our knowledge, Theorem 3.2 is the first generic result to make TS prior-free and applicable to adversarial environment. To that end, we note that Corollary 19 in (Lattimore & Gyorgy, 2021) only applies to $K-$armed bandits because of their truncation procedure.

3.3. Adaptive Minimax Sampling (AMS)

When the agent selects decision p_t by solving the minimax problem

$$\inf_{p_t} \sup_{\nu} \mathtt{AIR}_{q_t,\eta}(p, \nu),$$

and optimizes for algorithmic beliefs as in Algorithm 1, we call the resulting algorithm "Adaptive Minimax Sampling" (AMS). By the regret bound (2) in Theorem 3.1 and the relationship between between AIR and IR established in Lemma 2.5, it is straightforward to prove the following.

Algorithm 3 Adaptive Minimax Sampling (AMS)

```
Input learning rate $\eta > 0$.
Initialize $q_1 = \mathrm{Unif}(\Pi)$.
for round $t = 1, 2, \cdots, T$ do
    Find a distribution $p$ of $\pi$ and a distribution $\nu_t$ of $(M, \pi^*)$ that solves the saddle point of
        $$\inf_{p \in \Delta(\Pi)} \sup_{\nu \in \Delta(\mathcal{M} \times \Pi)} \mathtt{AIR}_{q_t,\eta}(p, \nu).$$
    Sample decision $\pi_t \sim p_t$ and observe $o_t \sim M_t(\pi_t)$.
    Update $q_{t+1} = (\nu_t)_{\pi^*|\pi_t,o_t}$.
end for
```

Theorem 3.3 (Regret of AMS)**.** *For a finite decision space Π and a compact model class $\mathcal{M}$, the regret of Algorithm 3 with $\eta = 2\sqrt{\log |\Pi| / (\mathtt{IR}(\mathrm{IDS}) \cdot T)}$ is always bounded by*

$$\mathfrak{R}_T \leq \frac{1}{2} \sqrt{\log |\Pi| \cdot \mathtt{IR}(\mathrm{IDS}) \cdot T},$$

where $\mathtt{IR}(\mathrm{IDS}) := \sup_\nu \inf_p \mathtt{IR}(\nu, p)$ is the maximal information ratio of Information-Directed Sampling.

Theorem 3.3 shows that the regret bound of AMS is always no worse than that of IDS (Russo & Van Roy, 2018). By showing implicit equivalence and making clean-ups, Algorithm 3 can also be explained as a much simplified implementation of the key ideas in the EBO algorithm from (Lattimore & Gyorgy, 2021), but AMS runs in computationally tractable spaces (rather than intractable functional spaces) and does not require unnecessary truncation.

3.4. Using approximate maximizers

In Algorithm 1, we ask for the algorithmic beliefs to maximize AIR. In order to give computationally efficient algorithms in practical applications (MAB, linear bandits, RL, ...), we will require the algorithmic beliefs to approximately maximize AIR. This argument is made rigorous in the following theorem, which uses the first-order optimization error of AIR to represent the regret bound.

[1] For technical reason we use the Hellinger distance to define $\mathtt{IR}_{\mathrm{H}}$ (instead of KL as in $\mathtt{IR}$), but there is no distinction between $\mathtt{IR}_{\mathrm{H}}$ and $\mathtt{IR}$ in all known applications within their current bounds.

Theorem 3.4 (Generic regret bound using approximate maximizers)**.** *Given a finite* Π*, a compact* $\mathcal{M}$*, an arbitrary algorithm* $\texttt{ALG}$ *that produces decision probability* $p_1, \dots, p_T$*, and a sequence of beliefs* $\nu_1, \dots, \nu_T$ *where* $q_t = (\nu_{t-1})_{\pi^*|\pi,o} \in \text{int}(\Delta(\Pi))$ *for all rounds, we have*

$$\mathfrak{R}_T \le \frac{\log|\Pi|}{\eta} + \sum_{t=1}^{T} \Bigg(\texttt{AIR}_{q_t,\eta}(p_t, \nu_t) + \sup_{\nu^*} \left(\frac{\partial \texttt{AIR}_{q_t,\eta}(p_t, \nu)}{\partial \nu}\bigg|_{\nu=\nu_t} \right)^{\top} (\nu^* - \nu_t) \Bigg).$$

Thus we give a concrete approach towards computationally efficient algorithms with rigorous guarantees—making the gradient of AIR small to approximately maximize AIR.

4. Application to Bernoulli MAB

Our Bayesian design principles give rise to a novel algorithm for the Bernoulli multi-armed bandits (MAB) problem. It is well-known that every bounded-reward MAB problem can equivalently be reduced to the Bernoulli MAB problem, so our algorithm and experimental results actually apply to all bounded-reward MAB problems. The reduction is very simple: assuming the rewards are always bounded in $[a, b]$, then after receiving r_t at each round, the agent re-samples a binary reward $\tilde{r}_t \sim \text{Bern}((r_t - a)/b - a)$ so that $\tilde{r}_t \in \{0, 1\}$.

4.1. Simplified APS for Bernoulli MAB

In Example 2.1, $\Pi = [K] = \{1, \cdots, K\}$ is a set of K actions, and each model $\mathcal{M}$ is a mapping from actions to Bernoulli distributions. Given belief $\nu \in \Delta(\mathcal{M} \times [K])$, we introduce the following parameterization: $\forall i, j \in [K]$,

$$\begin{aligned} \theta_i(j) &:= \mathbb{E}\left[r(j) | \pi^* = i\right], && \text{(conditional mean reward)} \\ \alpha(i) &:= \nu_{\pi^*}(i), && \text{(marginal belief)} \\ \beta_i(j) &:= \alpha(i) \cdot \theta_i(j). && \text{(guarantees concavity)} \end{aligned}$$

Then we have a concave parameterization of AIR by the $K(K+1)-$dimensional vector $(\alpha, \boldsymbol{\beta}) = (\alpha, \beta_1, \cdots, \beta_K)$:

$$\texttt{AIR}_{q,\eta}(p, \nu) = \sum_{i \in [K]} \beta_i(i) - \sum_{i,j \in [K]} p(j)\beta_i(j) - \frac{1}{\eta} \sum_{i,j \in [K]} p(j)\alpha(i) \text{kl}\left(\frac{\beta_i(j)}{\alpha(j)}, \sum_{i \in [K]} \beta_i(j) \right) - \frac{1}{\eta} \text{KL}(\alpha, q),$$

where $\text{kl}(x, y) := x \log \frac{x}{y} + (1 - x) \log \frac{1-x}{1-y}$ for all $x, y \in (0, 1)$. By setting the gradients of AIR with respect to all K^2 coordinates in β to be exactly zero, and choosing $\alpha = p$ (which results in the gradient of AIR with respect to α being suitbly bounded), we are able to write down a simplified APS algorithm in closed form (see Algorithm 4). We apply Theorem 3.4 to show that the algorithm achieves near-optimal $O(\sqrt{KT \log K})$ regret in the general adversarial setting. We leave the detailed derivation and analysis of the Algorithm 4 to Appendix F.2.

Algorithm 4 Simplified APS for Bernoulli MAB

```
Input learning rate $\eta > 0$.
Initialize $p_1 = \text{Unif}(\Pi)$.
for round $t = 1, 2, \cdots, T$ do
    Sample action $\pi_t \sim p_t$ and receives $r_t$.
    Update $p_{t+1}$ by
        $p_{t+1}(\pi_t) = \begin{cases} \frac{1-\exp(-\eta)}{1-\exp(-\eta/p_t(\pi_t))}, & \text{if } r_t = 1 \\ \frac{1-\exp(\eta)}{1-\exp(\eta/p_t(\pi_t))}, & \text{if } r_t = 0 \end{cases}$, and
        $p_{t+1}(\pi) = p_t(\pi) \cdot \frac{1 - p_{t+1}(\pi_t)}{1 - p_t(\pi_t)}, \quad \forall \pi \neq \pi_t.$
end for
```

At each round, Algorithm 4 increases the weight of the selected action π_t if $r_t = 1$, and decreases the weight if $r_t = 0$. The algorithm also maintains the "relative weight" between all unchosen actions $\pi \neq \pi_t$, allocating probabilities to these actions proportionally to p_t. Algorithm 4 is clearly very different from the well-known EXP3 algorithm, which instead updates p_{t+1} by the formula

$$p_{t+1}(\pi) = p_t(\pi) \exp\left(\eta \cdot \frac{r_t \mathbb{1}\{\pi = \pi_t\}}{p_t(\pi_t)} \right), \quad \forall \pi \in \Pi.$$

In Section B.2 we recover a modified version of EXP3 by Bayesian principle assuming Gaussian reward. We conclude that Algorithm 2 uses a precise posterior for Bernoulli reward, while EXP3 estimates worst-case Gaussian reward. This may explain why Algorithm 4 performs much better in all of our experiments.

4.2. Numerical experiments

We implement Algorithm 4 (with the legend "APS" in the figures) in the stochastic, adversarial and non-stationary environments. We plot expected regret (average of 100 runs) for different choices of η, and set $\gamma = 0.001$ in all experiments. We find APS 1) outperforms UCB and matches TS in the stochastic environment; 2) outperforms EXP3 in the adversarial environment; and 3) outperforms EXP3 and is comparable to the "clairvoyant" benchmarks (that have prior knowledge of the changes) in the non-stationary environment. For this reason we say Algorithm 4 (APS) achieves the "best-of-all-worlds" performance. We note that the optimized choice of η in APS differ instance by instance, but by an initial tuning we typically see good results, whether we tune η optimally or not optimally.

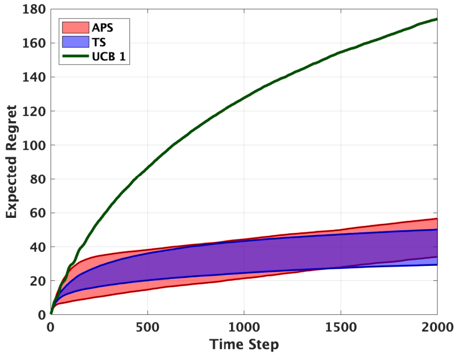

Figure 1. Sensitivity analysis in a stochastic bandit problem.

4.2.1. Stochastic Bernoulli MAB

In Figure 1 we report the expected regret for APS with different choices of η, TS with different Beta priors, and the UCB 1 algorithm, in a stochastic 16-armed Bernoulli bandit problem. We refer to this as “sensitivity analysis” because the red, semi-transparent, area reports the regret of APS when learning rates η are chosen across a range of values drawn from the interval $[0.05, 0.5]$ (the interval is specified by an initial tuning); and the priors of TS are chosen from Beta$(c, 1)$ where $c \in [0.5, 5]$. In particular, the bottom curve of the red (or blue) area is the regret curve of APS (or TS) using optimally tuned η (respectively, prior). The conclusion is that APS outperforms UCB 1, and is comparable to TS in this stochastic environment.

4.2.2. Adversarial Bernoulli MAB

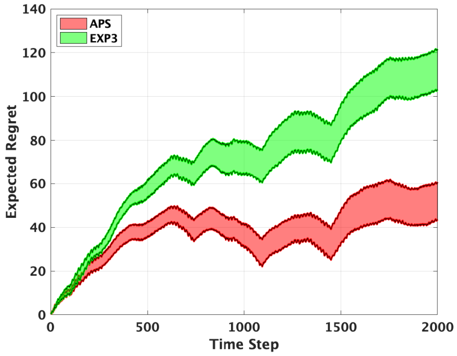

Figure 2. Sensitivity analysis in an adversarial bandit problem.

We equidistantly take 16 horizontal lines from an abstract art piece by Jackson Pollock to simulate the rewards (prespecified) in an adversarial environment, and study this via a 16-armed bandit problem. Figure 2 shows the sensitivity analysis for APS and EXP3 when both the learning rates are chosen from $[0.1, 5]$ (the interval is specified by an initial tuning). In particular, the red and green lower curves compare the optimally tuned versions of APS and EXP3. The conclusion is that APS outperforms EXP3 whether η is tuned optimally or not.

4.2.3. Non-stationary Bernoulli MAB (with change points)

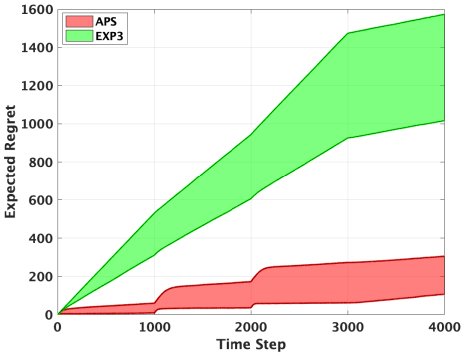

Figure 3. Sensitivity analysis in a “change points” environment.

We study a 16-armed Bernoulli bandit problem in a non-stationary environment. We generate 4 batches of i.i.d. sequences, where the changes in the environment occur after round 1000, round 2000, and round 3000. We consider a stronger notion of regret known as the dynamic regret (Besbes et al., 2014), which compares the cumulative reward of an algorithm to the cumulative reward of the best non-stationary policy (rather than a single arm) in hindsight. In this particular setting, the benchmark is to select the best arm in all the 4 batches. In Figure 3 we perform sensitivity analysis for APS and EXP3, where the learning rates are chosen across $[0, 05, 5]$. Since the agent will not know when and how the adversarial environment changes in general, it is most reasonable to compare APS with EXP3 without any knowledge of the environment as in Figure 3. We observe that APS dramatically improves the dynamic regret by several times.

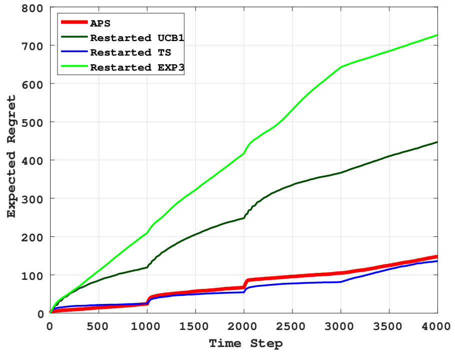

Figure 4. Comparing APS to “clairvoyant” restarted algorithms.

In Figure 4, we compare APS to three “clairvoyant” restarted algorithms, which require knowing that the environment consists of 4 batches of i.i.d. sequences, as well as knowing the exact change points. We tune the parameters in these algorithms optimally. Without knowledge of the environment, APS performs better than restarted EXP3 and restarted UCB 1, and is comparable to restarted TS. (It is important to emphasize again that the latter algorithms are restarted based on foreknowldge of the change points.)

4.2.4. Non-stationary Bernoulli MAB (with "sine curve" reward sequences)

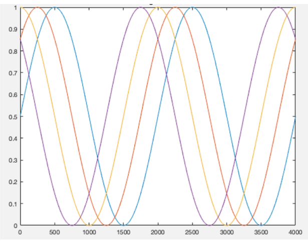

Figure 5. "Sine curve" rewards.

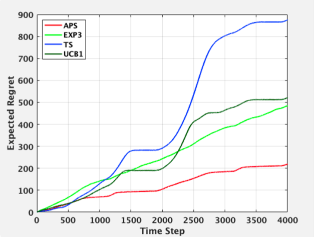

Figure 6. Regrets.

We generate a 4-armed bandit problem with the mean-reward structure shown in Figure 5. The four sine curves (with different colors) in Figure 5 represent the mean reward sequences of the 4 arms. We tune the parameters in all the algorithms to optimal and report their regret curves in Figure 6. As shown in Figure 6, APS achieves the best performance, while TS fails in this non-stationary environment. This experiment shows the vulnerability of TS if the environment is not stationary, such as the sine curve structure shown here. To better illustrate the smartness of APS compared with TS in the non-stationary environment, we track the selected arms and the best arms throughout the process. In Figure 7 and Figure 8, the horizontal line represents the 4000 rounds, and the vertical lines represent the 4 arms (indexed as 1, 2, 3, and 4). In Figure 7, the red points show the selected arms of APS, and the black points represent the best arms at each round in this "sine curve" non-stationary environment. In Figure 8, the blue points show the selected arms of TS. The more consistent the selected arms are with the best arms (black points), the better choices an algorithm makes. Comparing Figure 7 and Figure 8, we can see that APS is highly responsive to changes in the best arm, whereas TS is relatively sluggish in this regard. The implication of this experiment is that creating a new algorithmic belief at each round has the potential to significantly improve performance and be a game changer in many problem settings.

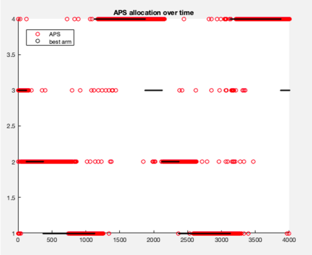

Figure 7. Selected arms of APS.

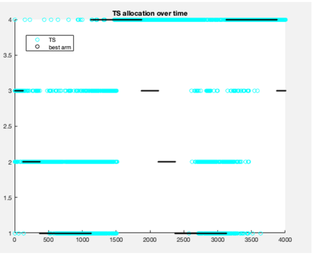

Figure 8. Selected arms of TS.

These experiments provide some numerical evidence indicating that APS achieves the "best-of-all-worlds" across stochastic, adversarial, and non-stationary environments.

5. Key Intuitions of the Proof

It is worth noting that the proof of Theorem 3.1 is quite insightful and parsimonious. The two major steps in the proof may be interesting on its own. The first step is a succinct analysis to bound the cumulative regret by sum of AIR (see Section 5.1); and the second step is to extend the classical minimax theory of "exchanging values" into a constructive approach to design minimax decisions (estimators, algorithms, etc.), which will be presented in Section 5.2.

5.1. Bounding regret by sum of AIR

For every $\bar\pi \in \Pi$, we have

$$\sum_{t=1}^{T}\left[\log\frac{q_{t+1}(\bar\pi)}{q_t(\bar\pi)}\right] = \log\frac{q_T(\bar\pi)}{q_1(\bar\pi)} \le \log|\Pi|. \tag{3}$$

Taking $q_{t+1} = (\nu_t)_{\pi^*|\pi_t,o_t}$ as in Algorithm 1, and taking expectation on the left hand side of (3), we have

$$\sum_{t=1}^{T}\mathbb{E}_{\pi_t,o_t}\left[\log\frac{(\nu_t)_{\pi^*}(\bar\pi|\pi_t,o_t)}{q_t(\bar\pi)}\right] \le \log|\Pi|. \tag{4}$$

By substracting the addition elements in the left hand side of (4) (divided by η) from per-round regret, we have

$$\begin{aligned}
&\mathfrak{R}_T - \frac{1}{\eta}\cdot\sum_{t=1}^{T}\mathbb{E}_{\pi_t,o_t}\left[\log\frac{(\nu_t)_{\pi^*}(\bar\pi|\pi_t,o_t)}{q_t(\bar\pi)}\right]\\
&=\sum_{t=1}^{T}\mathbb{E}\left[f_{M_t}(\bar\pi) - f_{M_t}(\pi_t) - \frac{1}{\eta}\log\frac{(\nu_t)_{\pi^*|\pi_t,o_t}(\bar\pi)}{q_t(\bar\pi)}\right]\\
&\le\sum_{t=1}^{T}\sup_{M,\bar\pi}\mathbb{E}\left[f_{M}(\bar\pi) - f_{M}(\pi_t) - \frac{1}{\eta}\log\frac{(\nu_t)_{\pi^*|\pi_t,o_t}(\bar\pi)}{q_t(\pi^*)}\right]\\
&\overset{(*)}{=}\sum_{t=1}^{T}\mathtt{AIR}_{q_t,\eta}(p_t,\nu_t)
\end{aligned} \tag{5}$$

where the inequality is by taking supremum at each rounds; and the last equality (5) is by Lemma 5.2, an important identity to be explained in Section 5.2, which is derived from the fact that the pair of maximizer ν_t and posterior functional is a Nash equilibrium of a convex-concave function.

5.2. Mimimax theory: from value to construction

Consider a decision space $\mathcal{X}$, a space $\mathcal{Y}$ of the adversary's outcome, and a convex-concave function $\psi(x, y)$ defined in $\mathcal{X}\times\mathcal{Y}$. The classical minimax theorem (Sion, 1958) says that, under regularity conditions, the minimax and maximin values of $\psi(x, y)$ are equal:

$$\min_{\mathcal{X}} \max_{\mathcal{Y}} \psi(x, y) = \max_{\mathcal{Y}} \min_{\mathcal{X}} \psi(x, y).$$

We refer to $\arg\min_{\mathcal{X}} \max_{\mathcal{Y}} \psi(x, y)$ as the set of "minimax decisions," as they are optimal in the worst-case scenario. And we say $\tilde{x} \in \arg\min_{\mathcal{X}} \psi(x, \bar{y})$ is "maximin decision" if $\bar{y} \in \arg\max_{\mathcal{Y}} \min_{\mathcal{X}} \psi(x, y)$ is "maximin adversary's outcome." One natural and important question is, *when will the "maximin decision" $\tilde{x}$ also be a "minimax decision?"* The study to this question may provide a constructive way to design frequentist estimators and algorithms through worst-case Bayesian posteriors and regularization. Making use of strong convexity, we extends the classical minimax theorem for values into the following minimax theorem for decisions:

Lemma 5.1 (Constructing minimax decisions)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be convex and compact sets, and $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ a function which for all y is strongly convex and continuous in x and for all x is concave and continuous in y. For each $y \in \mathcal{Y}$, let $x_y = \min_{x \in \mathcal{X}} \psi(x, y)$ be the corresponding unique minimizer. Then by maximizing the concave objective*

$$\bar{y} \in \max_{y \in \mathcal{Y}} \psi(x_y, y),$$

the pair $(x_{\bar{y}}, \bar{y})$ will be a Nash equilibrium that solves the minimax optimization problem $\min_{\mathcal{X}} \max_{\mathcal{Y}} \psi(x, y)$.

Applying Lemma 5.1 to our framework, we can show: 1) Bayesian posterior $\nu_{\pi^*|\pi,o}$ is the optimal functional to make decision under belief ν; and 2) by choosing worst-case belief $\bar{\nu}$, we construct a Nash equilibrium. As a result, we can prove the following per-round identity which is the last step in proving Theorem 3.1 (the key identity $(*)$ in (5)).

Lemma 5.2 (Identity by Nash equilibrium)**.** *Given $q \in \operatorname{int}(\Delta(\Pi))$, $\eta > 0$ and $p \in \Delta(\Pi)$, denote $\bar{\nu} \in \arg\max \mathtt{AIR}_{q,\eta}(p, \nu)$. Then we have*

$$\sup_{M, \bar{\pi}} \mathbb{E}\left[f_M(\bar{\pi}) - f_M(\pi) - \frac{1}{\eta} \log \frac{\bar{\nu}_{\pi^*|\pi,o}(\bar{\pi})}{q(\bar{\pi})} \right] = \mathtt{AIR}_{q,\eta}(p, \bar{\nu}).$$

6. Extensions

Our design principles can be applied in many sequential learning and decision making environments. In order to maximize AIR in practical applications, we parameterize the belief ν, and make the gradient of AIR with respect to such parameter small. Going beyond multi-amred bandits (MAB), we often need to constrain the search of algorithmic belief within a tractable subspace; and we study useful concave relaxations of AIR towards efficient algorithm design. We provide a high-level overview of the applications in linear bandits, bandit convex optimization, and reinforcement learning (RL) to showcase the broad applicability of our approach. Detailed results and explanations can be found in Appendix B and C.

Application to linear bandits. An established algorithm for tackling adversarial linear bandits (described in Example 2.2) is the EXP2 algorithm (Dani et al., 2007). It employs the inverse probability weighting (IPW) technique as a black-box estimation method for linear rewards, combined with an exponential weight updating rule. By the principle of optimizing AIR, we derive a modified version of EXP2 within our framework by assuming Gaussian distributions for the rewards. The resulting algorithm is computationally efficient and achieves the optimal $O(\sqrt{d^2 T})$ regret bound for all adversarial linear bandits with sub-Gaussian rewards. This outcome reveals an intriguing connection between IPW and Bayesian posteriors involving Gaussian rewards.

Application to bandit convex optimization. Bandit convex optimization, as described in Example 2.2, poses a well-known and challenging problem that has received significant attention in terms of understanding its minimax regret and designing algorithms. The current best-known result, derived through non-constructive information-ratio analysis in (Lattimore, 2020), achieves a regret bound of approximately $\tilde{O}(d^{2.5}\sqrt{T})$. In Corollary of Theorem 3.3, we demonstrate that Adaptive Minimax Sampling (AMS) achieves this same regret bound with a constructive algorithm that can be computed in polynomial time, specifically poly$(e^d \cdot T)$. To the best of our knowledge, this is the first algorithm with a finite running time that achieves the optimal $\tilde{O}(d^{2.5}\sqrt{T})$ regret bound. While the EBO algorithm in (Lattimore & Gyorgy, 2021) also achieves the same regret bound, it operates in an abstract functional space, making the computation less straightforward.

Application to RL. In the stochastic environment, where $M_t = M^* \in \mathcal{M}$ for all rounds, we want to find the optimal decision π_{M^*} that minimizes the mean reward function $f_{M^*}(\pi)$. Unlike the adversarial setting, where algorithmic beliefs are formed over pairs of models and optimal decisions, in the stochastic setting, we only need to search for algorithmic beliefs regarding the underlying model. This distinction allows us to develop a strengthened version of AIR, which we call "Model-index AIR" (MAIR), particularly suited for studying reinforcement learning problems.

Crucially, we can construct a generic and closed-form sequence of algorithmic beliefs that approximate the maximization of MAIR at each round. By leveraging these beliefs, we develop a model-based APS algorithm that achieves the sharpest known bounds for RL problems within the bilinear class (Du et al., 2021; Foster et al., 2021). Our algorithm features a generic and closed-form updating rule, making it potentially well-suited for efficient implementation through efficient sampling oracles. As a point of comparison, it is worth noting that the E2D algorithm introduced in (Foster et al., 2021) is not expressed in closed form and requires minimax optimization.

References

Abernethy, J., Hazan, E., and Rakhlin, A. Competing in the dark: An efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory, COLT 2008*, 2008.

Agarwal, A. and Zhang, T. Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity. *arXiv preprint arXiv:2206.07659*, 2022.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002a.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in Neural Information Processing Systems*, 27, 2014.

Bogachev, V. I. and Ruas, M. A. S. *Measure theory*, volume 1. Springer, 2007.

Dani, V., Kakade, S. M., and Hayes, T. The price of bandit information for online optimization. *Advances in Neural Information Processing Systems*, 20, 2007.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pp. 2826–2836. PMLR, 2021.

Foster, D. J. and Rakhlin, A. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *arXiv preprint arXiv:2002.04926*, 2020.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Foster, D. J., Golowich, N., Qian, J., Rakhlin, A., and Sekhari, A. A note on model-free reinforcement learning with the decision-estimation coefficient. *arXiv preprint arXiv:2211.14250*, 2022a.

Foster, D. J., Rakhlin, A., Sekhari, A., and Sridharan, K. On the complexity of adversarial decision making. *arXiv preprint arXiv:2206.13063*, 2022b.

Foster, D. J., Golowich, N., and Han, Y. Tight guarantees for interactive decision making with the decision-estimation coefficient. *arXiv preprint arXiv:2301.08215*, 2023.

Hao, B. and Lattimore, T. Regret bounds for information-directed reinforcement learning. *arXiv preprint arXiv:2206.04640*, 2022.

Hazan, E. and Karnin, Z. Volumetric spanners: an efficient exploration basis for learning. *The Journal of Machine Learning Research*, 17(1):4062–4095, 2016.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34, 2021.

Lai, T. L., Robbins, H., et al. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Lattimore, T. Improved regret for zeroth-order adversarial bandit convex optimisation. *Mathematical Statistics and Learning*, 2(3):311–334, 2020.

Lattimore, T. and Gyorgy, A. Mirror descent and the information ratio. In *Conference on Learning Theory*, pp. 2965–2992. PMLR, 2021.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, 2010.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Rockafellar, R. T. Convex analysis. In *Convex Analysis*. Princeton university press, 2015.

Russo, D. and Van Roy, B. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

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

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Sion, M. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Song, M. Proving that the conditional entropy of a probability measure is concave. *Mathematics Stack Exchange*, https://math.stackexchange.com/q/3080334, 2019.

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

Xu, Y. and Zeevi, A. Towards optimal problem dependent generalization error bounds in statistical learning theory. *arXiv preprint arXiv:2011.06186*, 2020.

Zhang, T. Feel-good thompson sampling for contextual bandits and reinforcement learning. *arXiv preprint arXiv:2110.00871*, 2021.

Part I

Applications to structured bandits and RL

A. Relation between AIR and DEC

We present here the definition of DEC (Foster et al., 2021).

Definition A.1 (Decision-estimation coefficient)**.** Given a model class $\mathcal{M}$, a nominal model $\bar{M}$ and $\eta > 0$, we define the decision-estimation coefficient by

$$\mathtt{DEC}_\eta\left(\mathcal{M}, \bar{M}\right) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\nu,p}\Big[f_M(\pi_M) - f_M(\pi) - \frac{1}{\eta} D_{\mathrm{H}}^2\left(M(\pi), \bar{M}(\pi)\right)\Big],$$

where $D_{\mathrm{H}}^2(\mathbb{P}, \mathbb{Q}) = \int(\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}})^2$ is the squared Hellinger distance between two probability measures.

DEC provides a tighter complexity measure compared to several existing measures in the literature, such as the bilinear dimension (Du et al., 2021) and the Eluder Bellman dimension (Jin et al., 2021), for reinforcement learning (RL) problems. Moreover, a slightly strengthened version of DEC, defined through the KL divergence instead of the Hellinger divergence, can be bounded by the traditional information ratio. This result follows from Proposition 9.1 in (Foster et al., 2021).

We establish that the worst-case value of AIR under a "maximin" strategy for selecting p is bounded by the worst-case value of the decision-estimation coefficient (DEC) for the convex hull of the model class in the following lemma.

Lemma A.2 (Bounding AIR by DEC)**.** *Given model class $\mathcal{M}$ and $\eta > 0$, we have*

$$\sup_{q \in \mathrm{int}(\Delta(\Pi))} \sup_{\nu} \inf_{p} \mathtt{AIR}_{q,\eta}(p, \nu) \leq \sup_{\bar{M} \in \mathrm{conv}(\mathcal{M})} \mathtt{DEC}_\eta(\Delta(\mathcal{M}), \bar{M}). \tag{6}$$

To prove Lemma A.2, we can start by noting that the left-hand side of (6) is equivalent to the "parametric information ratio," defined as

$$\max_{\nu} \min_{p} \mathbb{E}_{\nu,p}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}(\nu_{\pi^*|\pi,o}, \nu_{\pi^*})\right], \tag{7}$$

which was introduced in (Foster et al., 2022b). This equivalence can be shown by using the concavity of AIR to exchange $\sup$ over q and $\min$ over p. Furthermore, the inequality between (7) and the right-hand side of (6) has been established by Theorem 3.1 in (Foster et al., 2022b). Therefore, we obtain a proof of Lemma A.2.

We highlight that AIR is the tightest complexity measure in the adversarial setting. However, for reinforcement learning problems in the stochastic setting, it is often desirable to remove the convex hull in the right-hand side of (6). To this end, we introduce a tighter version of AIR, called "Model-index AIR" (MAIR), which allows us to apply most existing regret upper bounds using DEC to our framework. In Section C, we discuss our theory about MAIR and its application to RL in the stochastic setting.

B. Applications to Infinite-armed Bandits

Our design principles can be applied in many sequential learning and decision making environments. In order to maximize AIR in practical applications, we parameterize the belief ν, and make the gradient of AIR with respect to such parameter small. Going beyond multi-armed bandits (MAB), we often need to constrain the search of algorithmic belief within a tractable subspace; and we study useful concave relaxations of AIR towards efficient algorithm design. We will present our results for linear bandits and bandit convex optimization in this section and present our results for reinforcement learning in Section C. We give a high-level overview of the applications to linear bandits and bandit convex optimization here.

Application to linear bandits. A classical algorithm for adversarial linear bandits (described in Example 2.2) is the EXP2 algorithm (Dani et al., 2007), which uses IPW for linear loss as a black-box estimation method, and combines it with continuous exponential weight. We derive a modified version of EXP2 from our framework, establishing interesting connection between IPW and Bayesian posteriors.

Application to bandit convex optimization. Bandit convex optimization (described in Example 2.2) is a notoriously challenging problem, and much effort has been put to understanding its minimax regret and algorithm design. The best known result, which is of order $\tilde{O}(d^{2.5}\sqrt{T})$, is derived through the non-constructive information-ratio analysis in (Lattimore, 2020). As a corollary of Theorem 3.3, Adaptive Minimax Sampling (AMS) recovers the best known regret bound with a constructive algorithm, which can be computed in $\text{poly}(e^d \cdot T)$ time. To the best of our knowledge, this is the first finite-running-time algorithm that attains the best known $\tilde{O}(d^{2.5}\sqrt{T})$ regret.

B.1. Maximization of AIR for structured bandits

Consider the structured bandit problems described in Example 2.2. We consider the computation complexity of the optimization problem

$$\sup_{\nu\in\Delta(\mathcal{M}\times\Pi)} \texttt{AIR}_{q,\eta}(p,\nu). \tag{8}$$

The computational complexity of (8) may be $O(\text{poly}(\exp(\exp(d))))$ in the worst case as the size of $\mathcal{M}\times\Pi$. However, when the mean reward function class $\mathcal{F}$ is a convex function class, the computational complexity will be $O(\text{poly}(|\Pi|))$ which is efficient for $K-$armed bandits and is no more than $O(\text{poly}(e^d))$ in general (by standard discretization and covering arguments, we may assume $\Pi\subset\mathbb{R}^d$ to have finite cardinality $O(e^d)$ for the simplicity of theoretical analysis). Moreover, we also give efficient algorithm for linear bandits with exponential-many actions. We refer to Appendix F.1 for the detailed discussion on the parameterization method and computational complexity.

B.2. Application to Gaussian linear bandits

We consider the adversarial linear bandit problem with Gaussian reward. In such a MAB problem, $\Pi=\mathcal{A}\subseteq\mathbb{R}^d$ is an action set with dimension d. The model class $\mathcal{M}$ can be parameterized by a $d-$dimensional vector $\theta\in\mathbb{R}^d$ that satisfies $\theta^\top a\in[0,1]$ for all $a\in\mathcal{A}$. The reward $r(a)$ for each action $a\in\mathcal{A}$ is independently drawn from a Gaussian distribution that has mean $\theta^\top a$ and variance σ^2, and we assume that $\sigma\le 1$. Here we use the notations $\mathcal{A}$ (as action set), a (as action) and a^* (as optimal action) to follow the tradition of literature about linear bandits.

As discussed in Section B.1, we restrict our attention to sparse ν where for each $\pi^*\in\Pi$ there is only one model M, which corresponds to the Gaussian distribution $r(\pi)\sim N(\theta_{\pi^*}(\pi),1)$. We parameterize the prior ν by vectors $\{\beta_a^*\}_{a^*\in\mathcal{A}}$ and $\alpha\in\Delta(\Pi)$, where $\alpha=\mathbb{P}_\nu(a^*)$ and $\beta_{a^*}=\alpha(a^*)\cdot\theta_{a^*}$. As discussed in (25) in Appendix F.3, we propose to define a surrogate version of AIR by

$$\begin{aligned}\overline{\texttt{AIR}}_{q,\eta}(p,\nu)=&\int_{\mathcal{A}}\beta_{a^*}^\top a^* da^* - \int_{\mathcal{A}}\int_{\mathcal{A}} p(a)\beta_{a^*}^\top a da^* da\\ &-\frac{1}{2\eta}\int_{\mathcal{A}}\int_{\mathcal{A}} p(a)\alpha(a^*)\left(\frac{\beta_{a^*}^\top a}{\alpha(a^*)}-\int_{\mathcal{A}}\beta_{a^*}^\top a da^*\right)^2 da - \frac{1}{\eta}\text{KL}(\alpha,q).\end{aligned} \tag{9}$$

As discussed in Section B.1, It can be shown that approximate maximizers of this surrogate lead to rigorous regret bounds. Note that the surrogate defined in (9) can be bounded by the classical information ratio bounds defined by square loss (see, e.g., (Russo & Van Roy, 2016; Lattimore, 2020)).

By making all the gradients of (9) with respect to $\{\beta_{a^*}\}_{a^*\in\mathcal{A}}$ to be exactly zero, and taking $\alpha=p$, we obtain an approximate maximizer of AIR in (9). We calculate the Bayesian posterior, and find that the resulting algorithm is an exponential weight algorithm with a modified IPW estimator: at each round t, the agent update p_{t+1} by

$$\tilde{p}_{t+1}(a)\propto p_t(a)\exp\left(\eta\hat{r}_t(a)\right),$$

where $\hat{r}_t$ is the modified IPW estimator for linear loss,

$$\hat{r}_t(a)=a^\top(\mathbb{E}_{a\sim p_t}[aa^\top])^{-1}a_t r_t(a_t)-\frac{\eta}{2}(a^\top(\mathbb{E}_{a\sim p_t}[aa^\top])^{-1}a_t)^2. \tag{10}$$

Note that in order to avoid boundary conditions in our derivation, we require forced exploration to ensure $\lambda_{\min}(\mathbb{E}_{a\sim p}[aa^\top])\ge\eta$. This can be done with the help of the volumetric spanners constructed in (Hazan & Karnin, 2016). The use of volumetric spanner makes our final proposed algorithm (Algorithm 5) to be slightly more involved, but

Algorithm 5 Simplified APS for Gaussian linear bandits

```
Input learning rate $\eta > 0$, forced exploration rate $\gamma$, and action set $\mathcal{A}$.
Initialize $p_1 = \mathrm{Unif}(\Pi)$.
for round $t = 1, 2, \cdots, T$ do
    Let $S_t'$ be a $(p_t, \exp(-(4\sqrt{d} + \log(2T))))-$exp-volumetic spanner of $\mathcal{A}$,
    Let $S_t''$ be a $2\sqrt{d}-$ratio-volumetric spanner of $\mathcal{A}$.
    Set $S_t$ as the union of $S_t'$ and $S_t''$.
    Sample action $a_t \sim p_t$ and receives $r_t$.
    Calculate $\tilde{p}_{t+1}$ by
        $$\tilde{p}_{t+1}(a) \propto p_t(a) \exp\left(\eta \hat{r}_t(a)\right),$$
    where $\hat{r}_t$ is the modified IPW estimator for linear loss,
        $$\hat{r}_t(a) = a^\top (\mathbb{E}_{a\sim p_t}[aa^\top])^{-1} a_t r_t(a_t) - \frac{\eta}{2}(a^\top (\mathbb{E}_{a\sim p_t}[aa^\top])^{-1} a_t)^2.$$
    Update $p_{t+1}$ by $p_{t+1}(a) = (1-\gamma)\tilde{p}_t(a) + \frac{\gamma}{|S_t|}\mathbb{1}\{a \in S_t\}$
end for
```

we only use the volumetric spanner in a "black-box" manner. We highlight that the algorithm is computationally efficient as the reward estimator (10) is concave, so one can apply log-concave sampling when executing exponential weighting. The additional term in (10) is ignorable from a regret analysis perspective, so the standard analysis for exponential weight algorithms applies to Algorithm 5 to establish the optimal $O(\sqrt{d^2 T}$ regret bound. One may also analyze Algorithm 5 within our algorithmic belief framework through Theorem 3.4, as we did for Algorithm 4 in Section F.2; we omit the analysis here. Finally, we note that the algorithm reduces to a modified version of EXP3 for finite armed bandits, a connection we mentioned at the end of Section 4.1.

B.3. Application to bandit convex optimization

We consider the bandit convex optimization problem described in Example 2.2. In bandit convex optimization, $\Pi \subseteq \mathbb{R}^d$ is a $d-$dimensional action set whose diameter is bounded by $\mathrm{diam}(\Pi)$, and the mean reward (or loss) function is required to be concave (respectively, convex) with respect to actions:

$$\mathcal{F} = \{f : \Pi \to [0,1] : f \text{ is concave w.r.t. } \pi \in \Pi\}.$$

The problem is often formed with finite (but exponentially large) action set by standard discretization arguments (Lattimore, 2020). Bandit convex optimization is a notoriously challenging problem, and much effort has been put to understanding its minimax regret and algorithm design. The best known result, which is of order $\tilde{O}(d^{2.5}\sqrt{T})$, is derived through the non-constructive information-ratio analysis in (Lattimore, 2020). By the information ratio upper bound for the non-constructive Bayesian IDS algorithm in (Lattimore, 2020), Lemma 2.5 that bounds AIR by IR, and Theorem 3.3 (regret of AMS), we immediately have that Algorithm 3 (AMS) with optimally tuned η achieves

$$\mathfrak{R}_T \leq O\left(d^{2.5}\sqrt{T} \cdot \mathrm{polylog}(d, \mathrm{diam}(\mathcal{A}), T)\right)$$

As a result, AMS recovers the best known $\tilde{O}(d^{2.5}\sqrt{T})$ regret with a constructive algorithm. By our discussion on the computational complexity in Appendix F.1, AMS solves convex optimization in a $\mathrm{poly}(|\Pi|)$-dimensional space, so it can be computed in $\mathrm{poly}(e^d \cdot T)$ time for bandit convex optimization. To the best of our knowledge, this is the first algorithm with a finite running time that attains the best known $\tilde{O}(d^{2.5}\sqrt{T})$ regret. We note that the EBO algorithm in (Lattimore & Gyorgy, 2021) has given a constructive algorithm that achieves the same $\tilde{O}(d^{2.5}\sqrt{T})$ regret derived by Bayesian non-constructive analysis. However, EBO operates in an abstract functional space, so it is less clear how to execute the computation.

C. Model-index AIR and Application to RL

In the stochastic environment, where $M_t = M^* \in \mathcal{M}$ for all rounds, we want to find the optimal decision π_{M^*} that minimizes the mean reward function $f_{M^*}(\pi)$. Unlike the adversarial setting, where algorithmic beliefs are formed over pairs of models and optimal decisions, in the stochastic setting, we only need to search for algorithmic beliefs regarding the underlying model. This distinction allows us to develop a strengthened version of AIR, which we call "Model-index AIR" (MAIR), particularly suited for studying reinforcement learning problems.

Crucially, we can construct a generic and closed-form sequence of algorithmic beliefs that approximate the maximization of MAIR at each round. By leveraging these beliefs, we develop a model-based APS algorithm that achieves the sharpest known bounds for RL problems within the bilinear class (Du et al., 2021; Foster et al., 2021). Our algorithm features a generic and closed-form updating rule, making it potentially well-suited for efficient implementation through efficient sampling oracles.

C.1. Model-index Algorithmic Information Ratio

We denote decision $\pi_M \in \arg\min_{\Pi} f_M(\pi)$ be the induced optimal decision of model M. We introduce the following definition of Model-index AIR (MAIR).

Definition C.1 (Model-index AIR)**.** Denote $\rho \in \text{int}(\Delta(\mathcal{M}))$ be a reference distribution of models, and $\mu \in \text{int}(\Delta(\mathcal{M}))$ be a prior belief of models, we define the "Model-index Algorithmic Information Ratio" as

$$\texttt{MAIR}_{\rho,\eta}(p,\mu) = \mathbb{E}_{\mu,p}\left[f_M(\pi_M) - f_M(\pi) - \frac{1}{\eta}\text{KL}(\mu(\cdot|\pi,o),\rho)\right],$$

where $\mu(\cdot|\pi,o)$ is the Bayesian posterior belief of models induced by the prior belief μ.

It can be seen from the definition that KL divergence between two model distributions will be no smaller than KL divergence between the two induced decision distributions. Thus we have the following Lemma.

Lemma C.2 (MAIR smaller than AIR)**.** *When q is the decision distribution of π_M induced by the model distribution ρ, and ν is the distribution of (M, π_M) induced by the model distribution μ, we have*

$$\texttt{MAIR}_{\rho,\eta}(p,\mu) \leq \texttt{AIR}_{q,\eta}(p,\nu).$$

Lemma A.2 has shown that the worst-case value of AIR under the "maximin" strategy is smaller than DEC of the convex hull of $\mathcal{M}$. Now we demonstrate that the worst-case value of MAIR under a "maximin" strategy is smaller than the worst-case value of DEC, which does not uses the convex hull of model class in its first argument.

Lemma C.3 (Bounding MAIR by DEC)**.** *Given model class $\mathcal{M}$ and $\eta > 0$, we have*

$$\sup_{\rho\in\text{int}(\Delta(\mathcal{M}))} \sup_{\mu} \inf_{p} \texttt{MAIR}_{\rho,\eta}(p,\nu) \leq \sup_{\bar{M}\in\text{conv}(\mathcal{M})} \texttt{DEC}_{\eta}(\mathcal{M},\bar{M}).$$

Moreover, when the reference distribution ρ is centered at M^* and has "small" variance, we may completely removes the convex hull in the expression of DEC (unlike Lemma C.3 still leaving a convex hull restriction in the subscribe). This enable us to match the tightest possible version of DEC, and is discussed in Section C.3.

Comparing AIR and MAIR. We have seen that 1) Maximin AIR can be bounded by DEC of the convex hull $\Delta(\mathcal{M})$; 2) Maximin MAIR can be bounded by DEC of the original class $\mathcal{M}$; and 3) MAIR is "smaller" than AIR as illustrated in Lemma C.2. However, as we will later show in Theorem 3.1 and Theorem C.4, the regret bound using AIR will scale with a $\log|\Pi|$ factor (estimation complexity of decision space), while the regret bound using MAIR will scale with a bigger $\log|\mathcal{M}|$ factor (estimation complexity of model class). We explain their difference as follows.

When to use AIR *versus* MAIR*?* First, AIR is useful for both stochastic and adversarial bandit learning problems, while MAIR may only be useful for stochastic environments. Second, using AIR will result in a $\log|\Pi|$ factor along with information ratio (or DEC), while MAIR will result in a bigger $\log|\mathcal{M}|$ factor, so AIR is often the tighter option for bandit problems. For example, AIR provides optimal regret for K-armed bandits and $\sqrt{T}-$type regret bound for the challenging problem bandit convex optimization, while MAIR may not. On the other hand, MAIR can achieve optimal regret for stochastic linear bandits and stochastic model-based contextual bandits (Foster & Rakhlin, 2020), and it is more useful than AIR for reinforcement learning problems where taking convex hull to the model class may greatly increase the richness of model class. For example, the model class (especially the state transition dynamic) in reinforcement learning problems may not satisfy convexity. In general, AIR is more useful for “infinite divisible” problems where taking convex hull does not greatly increase the complexity of model class; while MAIR is more useful for stochastic model-based bandit and reinforcement learning problems where one wants to avoid taking convex hull.

C.2. Near-optimal algorithmic beliefs in closed form

For any fixed decision probability p, it is illustrative to write MAIR as

$$\begin{aligned}\mathtt{MAIR}_{\rho,\eta}(p,\mu) &= \mathbb{E}\left[f_M(\pi_M) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}(\mu(M|\pi,o),\rho)\right]\\ &= \mathbb{E}\left[f_M(\pi_M) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}(\mu(M|\pi,o),\mu) - \frac{1}{\eta}\mathrm{KL}(\mu,\rho)\right]\\ &= \mathbb{E}\left[f_M(\pi_M) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}(M(\pi),\mu_{o|\pi}) - \frac{1}{\eta}\mathrm{KL}(\mu,\rho)\right],\end{aligned} \tag{11}$$

where $\mu_{o|\pi} = \mathbb{E}_{M\sim\mu}[M(\pi)]$ is the induced distribution of o conditioned on π, and the third equality is by property of mutual information. We would like to give a sequence of algorithmic beliefs that approximately maximize MAIR at each rounds, as well as have closed-form expression.

We consider the following algorithmic priors at each round:

$$\mu_t(M) \propto \rho_t(M)\cdot\exp(\eta(f_M(\pi_M) - \mathbb{E}_{p_t}[f_M(\pi)])),$$

and use their corresponding posteriors to update the sequence of reference probabilities:

$$\mu_{t+1} = \mu_t(M|\pi_t,o_t) \propto \mu_t(M)[M(\pi_t)](o_t).$$

This results in the following update of ρ:

$$\rho_{t+1}(M) = \exp\left(\sum_{s=1}^{t}\left(\underbrace{\log[M(\pi_s)](o_s)}_{\text{log likelihood}} + \underbrace{\eta\left(f_M(\pi_M) - \mathbb{E}_{p_s}\left[f_M(\pi)\right]\right)}_{\text{adaptive algorithmic belief}}\right)\right). \tag{12}$$

Our algorithm (12) updates both the log likelihood term and an adaptive algorithmic belief term at each iteration, whereas traditional (fixed-prior) Thompson Sampling only updates the log likelihood term and relies on a fixed prior term.

Algorithm 6 Model-index AIR generation

```
Input algorithm ALG and learning rate $\eta > 0$.
Initialize $\rho_1$ to be the uniform distribution over $\mathcal{M}$.
for round $t = 1, 2, \cdots, T$ do
    Obtain $p_t$ from ALG. The algorithm ALG samples $\pi_t \sim p_t$ and observe the feedback $o_t \sim M_t(\pi_t)$.
    Update
        $\rho_{t+1}(M) \propto \exp\left(\sum_{s=1}^{t}\left(\log[M(\pi_s)](o_s) + \eta\left(f_M(\pi_M) - \mathbb{E}_{p_s}\left[f_M(\pi)\right]\right)\right)\right).$
end for
```

In Lemma C.3, we demonstrate how an upper bound on DEC can automatically translate into an upper bound on the maximin value of MAIR. However, the variable $\bar{M}$ in DEC is maximized within the convex hull $\Delta(\mathcal{M})$ rather than $\mathcal{M}$. Therefore, to directly apply the upper bounds on DEC proved in (Foster et al., 2021), we need to establish a stronger regret bound that completely eliminates convex hull from the expression of DEC. To achieve this, we prove that when the prior distribution μ is sufficiently "centered," we can bound MAIR using a DEC-type quantity that does not involve taking convex hull at all. Specifically, $\bar{M}$ takes values from $\mathcal{M}$ instead of $\Delta(\mathcal{M})$. We are motivated to prove this result by the fact that the update (12) will converge to M^* over time.

Theorem C.4 (Generic regret bound in the stochastic setting)**.** *Given a finite model class $\mathcal{M}$ where the underlying true model is $M^* \in \mathcal{M}$, and $f_M(\pi) \in [0,1]$ for every $M \in \mathcal{M}$ and $\pi \in \Pi$. For an arbitrary algorithm* `ALG`*, the regret of algorithm* `ALG` *is bounded by*

$$\mathfrak{R}_T \le \frac{\log(|\mathcal{M}|T)+1}{\eta} + 2 + \sum_{t=1}^{T} \mathbb{E}_{\mu_t,p_t}\left[5\left(f_M(\pi_M) - f_M(\pi)\right) - \frac{1}{\eta} D^2_{\mathrm{H}}(M(\pi), M^*(\pi)) - \frac{1}{\eta}\mathrm{KL}(\mu_t,\rho_t)\right],$$

where $\mu_t(M) \propto \exp\left(\sum_{s=1}^{t} \left(\log[M(\pi_s)](o_s) + \eta\left(f_M(\pi_M) - \mathbb{E}_{p_s}\left[f_M(\pi)\right]\right)\right)\right)$.

C.3. Model-index APS

In our applications, we often use a simple posterior sampling strategy for which we always induce the distribution of optimal decisions from the posterior distribution of models. We refer to the resulting algorithm, Algorithm 7, as "Model-index Adaptive Posterior Sampling."

Algorithm 7 Model-index Adaptive Posterior Sampling

```
Input learning rate $\eta$ and forced exploration rate $\gamma$.
Initialize $\rho_1$ to be the uniform distribution over $\mathcal{M}$.
for round $t = 1, 2, \cdots, T$ do
    Sample $\pi_t \sim p_t$ where $p_t(\pi) = \sum_{\pi=\pi_M} \rho_t(M)$, and observe the feedback $o_t \sim M_t(\pi_t)$.
    Update
        $\rho_{t+1}(M) \propto \exp\left(\sum_{s=1}^{t} \left(\log[M(\pi_s)](o_s) + \eta\left(f_M(\pi_M) - \mathbb{E}_{p_s}\left[f_M(\pi)\right]\right)\right)\right).$
end for
```

Algorithm 7 is inspired by and closely related to the optimistic posterior sampling algorithm proposed in (Agarwal & Zhang, 2022) (also termed as feel-good Thompson sampling in (Zhang, 2021)). Our analysis of sequential estimation (see Appendix G.2) is built on the analysis in (Agarwal & Zhang, 2022; Zhang, 2021). However, our approach has adaptive terms in our algorithmic beliefs rather than using a pre-specified optimistic prior. Moreover, our regret bounds can be applied to both on-policy bilinear class as well as the general bilinear class (as we will explain shortly in Theorem C.5 and Section C.4), while the theoretical results of optimistic posterior sampling in (Agarwal & Zhang, 2022) are only proved for on-policy bilinear class.

For model class $\mathcal{M}$, a nominal model $\bar{M}$, and the posterior sampling strategy $p(\pi) = \mu(\{M : \pi_M = \pi\})$, we can define the Bayesian decision-estimation coefficient of Thompson Sampling by

$$\mathtt{DEC}^{\mathrm{TS}}_{\eta}(\mathcal{M}, \bar{M}) = \sup_{\mu \in \Delta(\mathcal{M})} \mathbb{E}_{\nu,p}\left[f_M(\pi_M) - f_M(\pi) - \frac{1}{\eta} D^2_{\mathrm{H}}\left(M(\pi), \bar{M}(\pi)\right)\right]. \tag{13}$$

This value is bigger than the minimax DEC in Definition A.1, but often easier to use in model-based RL problems.

Theorem C.5 (Regret of Model-index Adaptive Posterior Sampling)**.** *Given a finite model class $\mathcal{M}$ where $f_M(\pi) \in [0,1]$ for every $M \in \mathcal{M}$ and $\pi \in \Pi$. The regret of Algorithm 7 with $\eta \le 1/10$ is bounded by*

$$\mathfrak{R}_T \le \frac{\log(|\mathcal{M}|T)+1}{\eta} + 5 \cdot \sup_{\bar{M} \in \mathcal{M}} \mathtt{DEC}^{\mathrm{TS}}_{2\eta}(\mathcal{M}, \bar{M}) \cdot T + 2.$$

C.4. Application to reinforcement learning

By using Algorithm 6 and Algorithm 7, we are able to recover several results in (Foster et al., 2021) that bound the regret of RL by DEC and the estimation complexity $\log|\mathcal{M}|$ of the model class. Note that we are able to prove such results for the

Model-index APS (Algorithm 2), which has the potential to be efficiently implemented through efficient sampling oracles. In contrast, the E2D algorithm in (Foster et al., 2021) is not in closed form and requires minimax optimization, and the sharp regret bounds are proved through the non-constructive Bayesian Thompson Sampling. The paper also presents regret bounds for a constructive algorithm using the so-called "inverse gap weighting" updating rules, but that algorithm has worse regret bounds than those proved through the non-constructive approach (by a factor of the bilinear dimension). As a result, Algorithm 7 makes an improvement because its simplicity and achieving the sharpest regret bound proved in (Foster et al., 2021) for RL problems in the bilinear class.

We discuss how our general problem formulation in Section 2.1 covers RL problems as follows.

Example C.6 (Reinforcement learning)**.** An episodic finite-horizon reinforcement learning problems is defined as follows. Let H be the horizon and $\mathcal{A}$ be a finite action set. Each model $M \in \mathcal{M}$ specifies a non-stationary Markov decision process (MDP) $\{\{S^{(h)}\}_{h=1}^{H}, \mathcal{A}, \{P_M^{(h)}\}_{h=1}^{H}, \{R_M^{(h)}\}_{h=1}^{H}, \mu\}$, where μ is the initial distribution over states; and for each layer h, $S^{(h)}$ is a finite state space, $P_M^{(h)} : S^{(h)} \times \mathcal{A} \to (S^{(h+1)})$ is the probability transition kernel, and $R_M^{(h)} : S^{(h)} \times \mathcal{A} \to \Delta([0,1])$ is the reward distribution. We allow the transition kernel and loss distribution to be different for different $M \in \mathcal{M}$ but assume μ to be fixed for simplicity. Let Π_{NS} be the space of all deterministic non-stationary policies $\pi = (u^{(1)}, \ldots, u^{(H)})$, where $u^{(h)} : S^{(}h) \to \mathcal{A}$. Given an MDP M and policy π, the MDP evolves as follows: beginning from $s^{(1)} \sim \mu$, at each layer $h = 1, \ldots, H$, the action $a^{(h)}$ is sampled from $u^{(h)}(s^{(h)})$, the loss $r^{(h)}(a^{(h)})$ is sampled from $R_M(s^{(h)}, a^{(h)})$ and the state $s^{(h+1)}$ is sampled from $P_M(\cdot|s^{(h)}, a^{(h)})$. Define $f_M(\pi) = \mathbb{E}[\Sigma_{h=1}^{H} r^{(h)}(a^{(h)})]$ to be the expected reward under MDP M and policy π. The general framework covers episodic reinforcement learning problems by taking the observation o_t to be the trajectory $(s_t^{(1)}, a_t^{(1)}, r_t^{(1)}), \ldots, (s_t^{(H)}, a_t^{(H)}, r_t^{(H)})$ and Π be a subspace of Π_{NS}. While our framework and complexity measures allow for agnostic policy classes, recovering existing results often requires us to make realizability-type assumptions.

We now focus on a broad class of structured reinforcement learning problems called "bilinear class" (Du et al., 2021; Foster et al., 2021). The following definition of the bilinear class is from (Foster et al., 2021).

Definition C.7 (Bilinear class)**.** A model class $\mathcal{M}$ is said to be bilinear relative to reference model $\bar{M}$ if:

1. There exist functions $W_h(\cdot; \bar{M}) : \mathcal{M} \to \mathbb{R}^d$, $X_h(\cdot; \bar{M}) : \mathcal{M} \times \mathbb{R}^d$ such that for all $M \in \mathcal{M}$ and $h \in [H]$,

$$|\mathbb{E}^{\bar{M}, \pi_M}[Q_h^{M,*}(s_h, a_h) - r_h - V_h^{M,*}(s_{h+1})]| \le |\langle W_h(M; \bar{M}), X_h(M; \bar{M})\rangle|.$$

We assume that $W_h(M : \bar{M}) = 0$.

2. Let $z_h = (s_h, a_h, r_h, s_{h+1})$. There exists a collection of estimation policies $\{\pi_M^{\text{est}}\}_{M \in \mathcal{M}}$ and estimation functions $\{\ell_M^{\text{est}}(\cdot; \cdot)\}_{M \in \mathcal{M}}$ such that for all $M, M' \in \mathcal{M}$ and $h \in [H]$,

$$\langle X_h(M; \bar{M}), W_h(M' : \bar{M})\rangle = \mathbb{E}^{\bar{M}, \pi_M \circ h \pi_M^{\text{est}}}[\ell_M^{\text{est}}(M'; z_h)].$$

If $\pi_M^{\text{est}} = \pi_M$, we say that estimation is on-policy.

If M is bilinear relative to all $\bar{M} \in \mathcal{M}$, we say that $\mathcal{M}$ is a bilinear class. We let $d_{\text{bi}}(\mathcal{M}, \bar{M})$ denote the minimal dimension d for which the bilinear class property holds relative to $\bar{M}$, and define $d_{\text{bi}}(\mathcal{M}) = \sup_{\bar{M} \in \mathcal{M}} d_{\text{bi}}(\mathcal{M}, \bar{M})$. We let $L_{\text{bi}}(\mathcal{M}; \bar{M}) \ge 1$ denote any almost sure upper bound on $|\ell_M^{\text{est}}(M'; z_h)|$ under $\bar{M}$, and let $L_{\text{bi}}(\mathcal{M}) = \sup_{\bar{M} \in \mathcal{M}} L_{\text{bi}}(\mathcal{M}; \bar{M})$.

For $\gamma \in [0,1]$, let π_M^{γ} be the randomized policy that—for each h—plays $\pi_{M,h}$ with probability $1 - \gamma/H$ and $\pi_{M,h}^{\text{est}}$ with probability γ/H. As an application of Theorem 7.1 in (Foster et al., 2021), we have upper bounds for $\texttt{DEC}_\eta^{\text{TS}}$ as follows.

Proposition C.8 (Upper bounds for bilinear class reinforcement learning)**.** *Let $\mathcal{M}$ be a bilinear class and let $\bar{M} \in \mathcal{M}$. Let $\mu \in \Delta(\mathcal{M})$ be given, and consider the modified Bayesian posterior sampling strategy that samples $M \sim \mu$ and plays π_M^{α}, where $\alpha \in [0,1]$ is a parameter.*

1. If $\pi_M^{\text{est}} = \pi_M$ (i.e., estimation is on-policy), this strategy with $\alpha = 0$ certifies that

$$\text{DEC}_\eta^{\text{TS}}(\mathcal{M}, \bar{M}) \le 4\eta H^2 L_{\text{bi}}^2(\mathcal{M}) d_{\text{bi}}(\mathcal{M}; \bar{M})$$

for all $\eta > 0$.

2. For general estimation policies, this strategy with $\gamma = \left(8\eta H^3 L_{\text{bi}}^2(\mathcal{M}) d_{\text{bi}}(\mathcal{M}, \bar{M})\right)^{1/2}$ certifies that

$$\text{DEC}_\eta^{\text{TS}_\gamma}(\mathcal{M}, \bar{M}) \le \left(32\eta H^3 L_{\text{bi}}^2(\mathcal{M}) d_{\text{bi}}(\mathcal{M}; \bar{M})\right)^{1/2}.$$

whenever $\gamma \geq 32H^3 L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \bar{M})$.

By applying the upper bounds on $\mathrm{DEC}^{\mathrm{TS}}\eta$ from Proposition C.8 to Theorem C.5, we can immediately obtain regret guarantees for RL problems in the bilinear class. In the on-policy case, Algorithm 7 with optimally tuned η achieves regret

$$\mathfrak{R}_T \leq O\big(H^2 Lbi^2 d_{\mathrm{bi}}(\mathcal{M}) \cdot T \cdot \log|\mathcal{M}|\big).$$

In the general case, Algorithm 7 with forced exploration rate $\gamma = \big(8\eta H^3 L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \bar{M})\big)^{1/2}$ and optimally tuned η achieves regret

$$\mathfrak{R}_T \leq O\Big(\big(H^3 L_{\mathrm{bi}}^2 d_{\mathrm{bi}}(\mathcal{M}) \log|\mathcal{M}|\big)^{1/3} \cdot T^{2/3}\Big).$$

As a closed-form algorithm that may be computed through sampling techniques, Algorithm 7 matches the sharp results for the non-constructive Bayesian Posterior Sampling algorithm proved in (Foster et al., 2021), and it achieves better regret bounds than the closed-form "inverse gap weighting" algorithm provided in the same paper. Its regret bound for RL problems in the bilinear class also match the E2D algorithms in (Foster et al., 2021; 2022a) that are not in closed-form and require more challenging minimax optimization.

Our results in this section apply to reinforcement learning problems where the DEC is easy to upper bound, but bounding the information ratio may be more challenging, particularly for complex RL problems where the model class $\mathcal{M}$ may not be convex and the average of two MDPs may not belong to the model class. Specifically, we propose MAIR and provide a generic algorithm that uses DEC and the estimation complexity of the model class ($\log|\mathcal{M}|$) to bound the regret. Another promising research direction is to extend our general results for AIR and the tools from Section B.2 to reinforcement learning problems with suitably bounded information ratios, such as tabular MDPs and linear MDPs, as suggested in (Hao & Lattimore, 2022). We anticipate that our tools can pave the way for developing constructive algorithms that provide regret bounds scaling solely with the estimation complexity of the value function class, which is typically smaller than that of the model class.

D. Conclusion and Future Directions

In this work, we propose a novel approach to solve sequential learning problems by generating "algorithmic beliefs." We optimize the Algorithmic Information Ratio (AIR) to generate these beliefs. Surprisingly, our algorithms achieve regret bounds that are as good as those assuming prior knowledge, even in the absence of such knowledge, which is often the case in adversarial or complex environments. Our approach results in simple and often efficient algorithms for various problems, such as multi-armed bandits, linear and convex bandits, and reinforcement learning.

Our work provides a new perspective on designing and analyzing bandit and reinforcement learning algorithms. Our theory applies to any algorithm through the notions of AIR and algorithmic beliefs, and it provides a simple and constructive understanding of the duality between frequentist regret and Bayesian regret in sequential learning. Optimizing AIR is a key principle to design effective and efficient bandit and RL algorithms. We demonstrate the effectiveness of our framework empirically via experiments on Bernoulli MAB and show that our derived algorithm achieves "best-of-all-worlds" empirical performance. Specifically, our algorithm outperforms UCB and is comparable to TS in stochastic bandits, outperforms EXP3 in adversarial bandits, and outperforms TS as well as clairvoyant restarted algorithms in non-stationary bandits.

Our study suggests several future research directions. First, we aim to provide computational guidelines for optimizing algorithmic beliefs, including techniques for selecting belief subspaces, parameterization, and surrogate objective functions. Second, we plan to develop efficient algorithm designs for infinite-armed bandit and reinforcement learning problems. As a first step, we aim to explore the Bayesian interpretation of frequentist approaches, such as gaining a deeper understanding of the inverse probability weighting (IPW) estimators and existing computationally-efficient algorithms for infinite-armed bandits (such as SCRiBLe (Abernethy et al., 2008)). Third, we aim to simulate average-case or non-stationary environments through constraint optimization for algorithmic beliefs. Fourth, we plan to investigate the essential features of the offset and constraint formulations in the algorithmic belief approach and explore possible connections with localized complexity in statistical learning theory (Xu & Zeevi, 2020) (offset formulation of DEC has been recently studied in (Foster et al., 2023)). Lastly, we aim to study instance-dependent bounds by leveraging AIR and algorithmic beliefs, which, to the best of our knowledge, is currently lacking in the context of information ratio.

Part II

Proofs and details

E. Extensions and Proofs for AIR

E.1. Extensions of AIR

Extension to general Bregman divergence We can generaliza AIR from using KL divergence to using general Bregman divergence. And all the results in Section 3 can be extended as well. This generalization is inspired by (Lattimore & Gyorgy, 2021), which defines information ratio and studies algorithm design using general Bregman divergence.

Let $\Psi : \Delta(\Pi) \to \mathbb{R} \cup \infty$ be a convex, Legendre, and second-order differentiable function. Denote D_Ψ to be the Bregman divergence of Ψ, and $\text{diam}(\Psi)$ to be the diameter of Ψ (see Appendix H.3) for the background). Given a reference probability $q \in \text{int}(\Delta(\Pi))$ in the interior of the simplex and learning rate $\eta > 0$, we define the generalized Algorithmic Information Ratio with potential function Ψ for decision p and distribution ν by

$$\texttt{AIR}^{\Psi}_{q,\eta}(p,\nu) = \mathbb{E}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{\eta} D_\Psi(\nu_{\pi^*|\pi,o}, \nu_{\pi^*}) - \frac{1}{\eta} D_\Psi(\nu_{\pi^*}, q)\right]. \tag{14}$$

We generalize Theorem 3.1 and Theorem 3.4 to second-order differentiable Bregman divergence for general convex Legendre function, with the $\frac{\log|\Pi|}{\eta}$ term in the regret bound be replaced by $\frac{\text{diam}(\Psi)}{\eta}$. Using the extension, we can generalize Theorem 3.2 (regret of APS) and Theorem 3.3 (regret of AMS) to generalized Bregman divergence as well, where the definition of information ratio will also use the corresponding Bregman divergence as in (Lattimore & Gyorgy, 2021). We state the extension of Theorem 3.4 here.

Theorem E.1 (Using general Bregman divergence)**.** *Assume $\Psi : \Delta(\Pi) \to \mathbb{R} \cup \infty$ is convex, Legendre, second-order differentiable, and has bounded diameter. Given a compact $\mathcal{M}$, an arbitrary algorithm $\texttt{ALG}$ that produces decision probability $p_1, \ldots, p_T$, and a sequence of beliefs $\nu_1, \ldots, \nu_T$ where $(\nu_t)_{\pi^*|\pi,o} \in \text{int}(\Delta(\Pi))$ for all rounds, we have*

$$\begin{aligned}\mathfrak{R}_T \leq \frac{\text{diam}(\Psi)}{\eta} + \sum_{t=1}^{T} \Bigg(& \texttt{AIR}^{\Psi}_{q_t,\eta}(p_t,\nu_t) \\ & + \sup_{\nu^*} \left(\left.\frac{\partial \texttt{AIR}^{\Psi}_{q_t,\eta}(p_t,\nu)}{\partial \nu}\right|_{\nu=\nu_t} \right)^{\top} (\nu^* - \nu_t) \Bigg).\end{aligned}$$

Extension to high probability bound We conjecture that the results in Section 3 may be able to be extended to high probability bounds, with some modification in our algorithm and complexity measure. We refer to (Foster et al., 2022b) for a possible approach to achieve this goal.

E.2. Proof of Theorem 3.1:

By the discussion in Section 5.1 and 5.2, we only need to prove Lemma 5.2 in order to prove Theorem 3.1.

Proof of Lemma 5.2: let $\mathcal{Q}$ the space of all mappings from $\Pi \times \mathcal{O}$ to $\Delta(\Pi)$. For a mapping $Q \in \mathcal{Q}$, denote $Q[\pi_t, o_t] \in \Delta(\Pi)$ as the image of (π, o). Define $B : \Delta(\mathcal{M} \times \Pi) \times \mathcal{Q} \to \mathbb{R}$ by

$$B(\nu, Q) = \mathbb{E}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{\eta} \log \frac{Q[\pi,o](\pi^*)}{q(\pi^*)}\right]. \tag{15}$$

$B(\nu, Q)$ is linear with respect to ν, convex with respect to Q. In order to apply minimax theorem to the concave-convex objective function B, we need to verify that the sets $\mathcal{Q}$ and $\Delta(\mathcal{M} \times \Pi)$ are convex and compact sets, and B is continuous with respect to both $Q \in \mathcal{Q}$ and $\nu \in \Delta(\mathcal{M} \times \Pi)$. This verification step assumes a basic understanding of general topology, as it involves infinite sets (compactness and continuity for finite sets are trivial); and eager readers may choose to skip this step. For this reason we put the verification step to the end of the proof. By applying Sion's minimax theorem (Lemma H.2), we have

$$\sup_{v} \inf_{Q} B(\nu, Q) = \inf_{Q} \sup_{\nu} B(\nu, Q). \tag{16}$$

From the definition of AIR and first-order optimality condition, we have

$$\begin{aligned}
&\texttt{AIR}_{q,\eta}(p, \nu) \\
=&\mathbb{E}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}(\mathbb{P}_\nu(\pi^*|\pi, o), q)\right] \\
=&\inf_{Q \in \mathcal{Q}} B(\nu, Q),
\end{aligned}$$

so $\bar{\nu}$ is the maximizer of $\sup_\nu \inf_Q B(\nu, Q)$. Define $Q_{\bar{\nu}}$ as the mapping that maps each (π, o) to the conditional probability $\mathbb{P}_{\bar{\nu}}(\pi^*|\pi, o)$, then $Q_{\bar{\nu}}$ is the unique minimizer of $B(\bar{\nu}, Q)$. Denote $\bar{Q}$ to be the minimizer of $\inf_Q \sup_\nu B(\nu, Q)$. From the equality (16), $(\bar{\nu}, \bar{Q})$ must be a Nash equilibrium of B, i.e.

$$\texttt{AIR}_{q,\eta}(p, \bar{\nu}) = \sup_{\nu} \inf_{Q} B(\nu, Q) = B(\bar{\nu}, \bar{Q}) = \inf_{Q} \sup_{\nu} B(\nu, Q).$$

Then $\bar{Q}$ is a minimizer of $B(\nu, Q)$, which implies $Q_{\bar{\nu}} = \bar{Q}$ as the minimizer is unique. As a result, we have

$$\begin{aligned}
&\texttt{AIR}_{q,\eta}(p, \bar{\nu}) \\
=&\sup_{\nu} B(\nu, Q_{\bar{\nu}}) \\
=&\sup_{M, \pi^*} \mathbb{E}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{\eta}\log\frac{\mathbb{P}_{\bar{\nu}}(\pi^*|\pi, o)}{q(\pi^*)}\right].
\end{aligned}$$

Verification of the conditions of Sion's minimax theorem: It is straightforward to see convexity of the sets $\mathcal{Q}$ and $\Delta(\mathcal{M} \times \Pi)$. As a collection of mappings, $\mathcal{Q}$ is compact with respect to the product topology by Tychonoff's theorem, and B is continuous with respect to Q by the definition of product topology. Because the probability measure on the compact set is compact with respect to the weak*-topology, $\Delta(\mathcal{M} \times \Pi)$ is a compact set. We refer to the book (Bogachev & Ruas, 2007) for the basic background of general topology. Finally, B is continuous with respect to ν because B is linear in ν. □

E.3. Proof of Theorem 3.3

Combining Theorem 3.1 and Lemma 2.5, we prove Theorem 3.3.

E.4. Proof of Theorem 3.2

We prove the following lemma that upper bounds $\texttt{AIR}_{q_t,\eta}(q_t, \nu_t)$ by DEC and information ratio for Thompson Sampling. Theorem 3.2 will be a straightforward consequence of the regret bound (2) in Theorem 3.1 and this lemma.

Lemma E.2 (Bounding AIR by DEC and IR for TS)**.** *Assume that $f_M(\pi)$ is bounded in $[0,1]$ for all M, π. Then for $\eta \in (0, 1/2]$ and all $q \in \mathrm{int}(\Delta(\Pi))$, we have*

$$\texttt{AIR}_{q,\eta}(q, \nu) \leq \sup_{M \in \Delta(\mathcal{M})} \texttt{DEC}_{2\eta}(\Delta(\mathcal{M}), M) + 2\eta \leq \frac{\eta}{2} \cdot \texttt{IR}^{\mathrm{TS}} + 2\eta.$$

Proof of Lemma E.2: Given a probability measure ν, Denote $\nu_{o|\pi^*,\pi} = \mathbb{E}_{M \sim \nu_{M|\pi^*}}[M(\pi)]$ to be the posterior belief of observation o conditioned on π^* and π, and $\nu_{o|\pi} = \mathbb{E}_{(M,\pi^*) \sim \nu}[M(\pi)]$ to be posterior belief of o conditioned solely on π.

Denote the $|\Pi|-$dimensional vector X, Y by

$$\begin{aligned}
X(\pi) &= \mathbb{E}_{(M,\pi^*) \sim \nu}\left[f_M(\pi^*) - f_M(\pi)\right], \\
Y(\pi) &= \mathbb{E}_{(M,\pi^*) \sim \nu}\left[D_{\mathrm{H}}^2(\nu_{o|\pi^*,\pi}, \nu_{o|\pi})\right].
\end{aligned}$$

Note that the Algorithmic Information Ratio can always be written as

$$\begin{aligned}\mathtt{AIR}_{q,\eta}(p,\nu) &= \mathbb{E}_{\nu,p}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}(\nu_{\pi^*|\pi,o}, q)\right]\\ &= \mathbb{E}_{\nu,p}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}(\nu_{\pi^*|\pi,o}, \nu^*_\pi) - \frac{1}{\eta}\mathrm{KL}(\nu_{\pi^*}, q)\right]\\ &= \mathbb{E}_{\nu,p}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}(\nu_{o|\pi^*,\pi}, \nu_{o|\pi}) - \frac{1}{\eta}\mathrm{KL}(\nu_{\pi^*}, q)\right],\end{aligned} \tag{17}$$

where the first equality is the definition of AIR; the second equality is because the expectation of posterior is equal to prior; and the third equality is due to the symmetry property of mutual information. By (17) we have that

$$\begin{aligned}&\mathtt{AIR}_{q,\eta}(q,\nu)\\ \leq&\mathbb{E}_{\nu,q}\left[f_M(\pi_M) - f_M(\pi) - \frac{1}{\eta}D^2_{\mathrm{H}}(\nu_{o|\pi^*,\pi}, \nu_{o|\pi})\right] - \frac{1}{\eta}\mathrm{KL}(\nu^*_\pi, \rho_t)\\ =&\langle q, X\rangle - \frac{1}{2\eta}\mathrm{KL}(\nu_{\pi^*}, q) - \frac{1}{\eta}\langle q, Y\rangle - \frac{1}{2\eta}\mathrm{KL}(\nu_{\pi^*}, q)\\ \leq&\langle \nu_{\pi^*}, X\rangle + 2\eta - \frac{1}{\eta}\langle q, Y\rangle - \frac{1}{2\eta}\mathrm{KL}(\nu_{\pi^*}, q)\\ \leq&\langle \nu_{\pi^*}, X\rangle + 2\eta - \frac{1}{\eta}\langle q, Y\rangle - \frac{1}{2\eta}D^2_{\mathrm{H}}(\nu_{\pi^*}, q)\\ \leq&\langle \nu_{\pi^*}, X\rangle - \frac{(1-\eta)}{(1+\eta)\eta}\langle \nu_{\pi^*}, Y\rangle + 2\eta\\ \leq&\mathbb{E}_{\nu,\nu_{\pi^*}}\left[f_M(\pi_M) - f_M(\pi) - \frac{1}{2\eta}D^2_{\mathrm{H}}(\nu_{o|\pi^*,\pi}, \nu_{o|\pi})\right] + 2\eta,\end{aligned} \tag{18}$$

where the first inequality is by Lemma H.6; the second inequality is by Lemma H.8 and the fact $f_M(\pi) \in [0,1]$ for all $M \in \mathcal{M}$ and $\pi \in \Pi$; the third inequality is by Lemma H.6; the fourth inequality is a consequence of Lemma H.7 and the AM-GM inequality; and the last inequality uses the condition $\eta \leq \frac{1}{2}$.

Finally, by combining (18) and Lemma 2.5, we have that

$$\mathtt{AIR}_{q,\eta}(q,\nu) \leq \sup_{M\in\Delta(\mathcal{M})} \mathtt{DEC}_{2\eta}(\Delta(\mathcal{M}), M) + 2\eta \leq \frac{\eta}{2}\cdot \mathtt{IR} + 2\eta.$$

□

E.5. Proof of Theorem 3.4 and Theorem E.1

In this section we prove Theorem E.1, which is a more general extension to Theorem 3.4 (Theorem E.1 applies to general Bregman divergence with second-order differentiable Ψ while Theorem 3.4 is stated with the KL divergence). Theorem 3.4 and E.1 are consequences of the following "envelop theorem," which shows that gradients of AIR with respect to ν is equal to the gradient of the adversary when one uses the posterior mapping as the decision rule.

Lemma E.3 (Envelop theorem)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be convex sets, and $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ a function such that for all $y \in \mathcal{Y}$, $\phi(\cdot, y)$ is a Legendre function in x; and for all $x \in \mathcal{X}$, $-\phi(x,\cdot)$ is a Legendre function in y. For each $y \in \mathcal{Y}$, let x_y be the a minimizer of the convex optimization problem*

$$\min_{x\in\mathcal{X}} \phi(x,y),$$

and assume that x_y is differentiable with respect to y. Then for all $y \in$ int(dom(ψ)), we have

$$\frac{\partial\phi(x_y,y)}{\partial y} = \left.\frac{\partial\phi(x,y)}{\partial y}\right|_{x=x_y}.$$

We consider generalized AIR, where Ψ is a Legendre and second-order differentiable function. Recall in (14) we define

$$\texttt{AIR}_{q,\eta}^{\Psi}(p,\nu) = \mathbb{E}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{\eta}D_\Psi(\nu_{\pi^*|\pi,o},\nu_{\pi^*}) - \frac{1}{\eta}D_\Psi(\nu_{\pi^*},q)\right].$$

And similar to (15), for all $\nu \in \Delta(\mathcal{M}\times\Pi)$ such that $\mathbb{P}_\nu(\pi^*) \in \text{int}(\Delta(\Pi))$ and $R : \Pi \times \mathcal{O} \to \mathbb{R}^{|\Pi|}$, we define

$$B(\nu, R) = \mathbb{E}\left[f_M(\pi^*) - f_M(\pi) + \frac{1}{\eta}\langle \nabla\Psi(q) - R[\pi,o]), \mathbb{1}(\pi^*) - q\rangle + \frac{1}{\eta}D_{\Psi^*}(R[\pi,o], \nabla\Psi(q))\right],$$

where $\mathbb{1}(\pi^*)$ is the vector whose π^*-coordinate is 1 but all other coordinates are 0. Note that $B(\nu, R)$ is a convex function with respect to R. By Lemma H.4 and the property (b) in Lemma H.3, we know that

$$\texttt{AIR}_{q,\eta}^{\Psi}(p,\nu) = B(\nu, \nabla\Psi(\nu_{\pi^*|\cdot,\cdot})) = \inf_R B(\nu, R),$$

where the last equality is by the first-order optimal condition. By Lemma E.3, when Ψ is second-order differentiable, we further have

$$\frac{\partial \texttt{AIR}_{q,\eta}^{\Psi}(p,\nu)}{\partial \nu} = \left.\frac{\partial B(\nu,R)}{\partial\nu}\right|_{R=\nabla\Psi(\nu_{\pi^*|\cdot,\cdot})}.$$

By the above identity and the linearity of $B(\nu, Q)$ with respect to ν, we have

$$B(\nu^*, \nu_{\pi^*|\cdot,\cdot}) = \texttt{AIR}_{q,\eta}^{\Psi}(p,\nu) + \left\langle \frac{\partial \texttt{AIR}_{q,\eta}^{\Psi}(p,\nu)}{\partial\nu}, \nu^* - \nu\right\rangle, \quad \forall \nu^*.$$

Following the same steps in proving Theorem 3.1, we prove Theorem E.1 (and consequently Theorem 3.4).

F. Details and Proofs for Bandit Problems

F.1. Concave parameterization with Bernoulli reward

We consider Bernoulli structured bandit with an action set $\Pi \subset \mathbb{R}^d$ and a mean reward function class $\mathcal{F} \subset (\Pi :\mapsto [0,1])$ that is convex. (As discussed in the beginning of Section 4, every bounded-reward bandit problem can equivalently be reduced to a Bernoulli bandit problem. For simplicity we make the standard assumption that Π is finite, which can be removed using standard descretization and covering argument. The goal here is to make the computation complexity to be independent of the size of model class $\mathcal{M}$, but only depends on $|\Pi|$. The general principle to achieve this goal is as follows. For each possible value of π^*, we assign an "effective model" M_{π^*} to π^* so that the optimization problem (8) reduces to selecting those $|\Pi|$ "effective models," as well as the probability distribution over them.

We introduce the following parametrization: $\forall a, a^* \in \Pi$ (we use notation a^*, a as the index sometimes to avoid repetition of notation π^*, π),

$$\theta_{a^*}(a) = \mathbb{E}\left[r(a)|\pi^* = a^*\right],$$
$$\alpha(a^*) = \nu_{a^*|\pi,o}(a^*),$$
$$\beta_{a^*}(a) = \alpha(a^*)\cdot\theta_{a^*}(a).$$

Then we have represent AIR by $(\alpha, \beta_{\pi^*}{}_{\pi^*\in\Pi})$:

$$\begin{aligned}\texttt{AIR}_{q,\eta}(p,\nu) = &\sum_{\pi^*\in\Pi}\beta_{\pi^*}(\pi^*) - \sum_{\pi^*,\pi\in[K]}p(\pi)\beta_{\pi^*}(\pi)\\ &-\frac{1}{\eta}\sum_{\pi^*,\pi\in[K]}p(\pi)\alpha(\pi^*)\text{kl}\left(\frac{\beta_{\pi^*}(\pi)}{\alpha(\pi)}, \sum_{\pi^*\in\Pi}\beta_{\pi^*}(\pi)\right) - \frac{1}{\eta}\text{KL}(\alpha,q),\end{aligned} \tag{19}$$

and the constraint of $(\alpha, \beta_{\pi^*}{}_{\pi^*\in\Pi})$ is that the functions parameterized by θ_{a^*} belong to the mean reward function class $\mathcal{F}$. We know the constraint set of $(\alpha, \beta_{\pi^*}{}_{\pi^*\in\Pi})$ to be convex because the convexity of perspective function.

Now we want to prove that the AIR objective in the maximization problem (8) is concave. We have

$$\begin{aligned} B(\nu, Q) &= \mathbb{E}_{p,\nu}\left[f_M(\pi^*) - f_M(\pi) + \frac{1}{\eta}\int_{\mathcal{O}} \mathbb{P}_\nu(\pi, o, \pi^*) \log \frac{q(\pi^*)}{Q[\pi, o](\pi^*)} do\right] \\ &= \sum_{\pi^*} \beta_{\pi^*}(\pi^*) - \sum_{\pi,\pi^*} p(\pi)\beta_{\pi^*}(\pi) + \frac{1}{\eta}\sum_{\pi,\pi^*} p(\pi)\beta_{\pi^*}(\pi) \log \frac{q(\pi^*)}{Q[\pi,1](\pi^*)} \\ &\qquad + \frac{1}{\eta}\sum_{\pi,\pi^*} p(\pi)(\alpha(\pi^*) - \beta_{\pi^*}(\pi^*)) \log \frac{q(\pi^*)}{Q[\pi,0](\pi^*)}. \end{aligned} \tag{20}$$

This means that after parameterizing ν with α and $\{\beta_{\pi^*}\}_{\pi^* \in \Pi}$, $B(\nu, Q)$ will be a linear function of $(\alpha, \{\beta_{\pi^*}\}_{\pi^* \in \Pi})$. As a result,

$$\texttt{AIR}_{q,\eta}(p, \nu) = \inf_Q B(\nu, Q)$$

will be a concave function of $(\alpha, \{\beta_{\pi^*}\}_{\pi^* \in \Pi})$. So the optimization problem to maximize AIR is a convex optimization problem, whose computational complexity will be poly-logarithmic to the cardinality of $(\alpha, \{\beta_{\pi^*}\}_{\pi^* \in \Pi})$. As a result, the computational complexity to maximize AIR is polynomial in $|\Pi|$ and does not depends on cardinality of the model class. This discussion shows that we give finite-running-time algorithm with computational complexity poly(e^d) even when the cardinality of model class is double-exponential. Still, the computation is only efficient for simple problems such as $K-$armed bandits, but we also give efficient algorithm for linear bandits in Appendix B.2.

F.2. Simplified APS for Bernoulli MAB

For Bernolli $K-$armed bandits discussed in in Section 4.1, we give the details about how to use first-order optimality conditions to derive Algorithm 4.

We denote $\nu_{\pi^*}(i|j, 1)$ as the shorthand for $\nu_{\pi^*}(i|\pi = j, o = 1)$, the conditional probability $\mathbb{P}(\pi^* = i|\pi = j, o = 1)$ when the underlying probability measure is ν.

By (20) and Lemma E.3 (using the envelop theorem and the bivariate function (20) to calculate the derivatives is easier than directly calculating the derivatives of the AIR parameterization (19)), we have for each $i \in [K]$,

$$\frac{\partial \texttt{AIR}_{q,\eta}(p, \nu)}{\partial \beta_i(i)} = (1 - p(i)) - \frac{1}{\eta} p(i) \left(\log \nu_{\pi^*}(i|i, 1) - \log \nu_{\pi^*}(i|i, 0)\right). \tag{21}$$

And for every $i \neq j \in [K]$,

$$\frac{\partial \texttt{AIR}_{q,\eta}(p, \nu)}{\partial \beta_i(j)} = -p(j) - \frac{1}{\eta} p(j) \left(\log \nu_{\pi^*}(i|j, 1) - \log \nu_{\pi^*}(i|j, 0)\right). \tag{22}$$

Lastly, for each $i \in [K]$,

$$\frac{\partial \texttt{AIR}_{q,\eta}(p, \nu)}{\partial \alpha(i)} = \frac{1}{\eta} \sum_{j \in [K]} p(j) \left(\log q(i) - \log \nu_{\pi^*}(i|j, 0)\right). \tag{23}$$

We let the derivatives in (21) and (22) be zero, which means that the derivatives with respect to all coordinates of β are zero. We have for all $i \neq j \in [K]$

$$\begin{aligned} \log \frac{\nu_{\pi^*}(j|j, 1)}{\nu_{\pi^*}(j|j, 0)} &= \frac{\eta}{p(j)} - \eta, \\ \log \frac{1 - \nu_{\pi^*}(i|j, 1)}{1 - \nu_{\pi^*}(i|j, 0)} &= -\eta, \quad i \neq j. \end{aligned}$$

Solving the above two equation we obtain

$$\begin{aligned}
\nu_{\pi^*}(j|j,1) &= \frac{1-\exp(-\eta)}{1-\exp(-\eta/p(j))}, \quad \forall j \in [K],\\
\nu_{\pi^*}(i|j,1) &= \frac{\exp(-\eta)-\exp(-\eta/p(j))}{1-\exp(-\eta/p(j))}\cdot\frac{\alpha(i)}{1-\alpha(j)}, \quad \forall i \neq j \in [K],\\
\nu_{\pi^*}(j|j,0) &= \frac{\exp(\eta)-1}{\exp(\eta/p(j))-1}, \quad \forall j \in [K],\\
\nu_{\pi^*}(i|j,0) &= \frac{\exp(\eta/p(j))-\exp(\eta)}{\exp(\eta/p(j))-1}\cdot\frac{\alpha(i)}{1-\alpha(j)}, \quad \forall i \neq j \in [K].
\end{aligned} \tag{24}$$

Now we set $\alpha = p = q$ so that the posterior updates (24) all have closed forms. Now we want to prove that (23) (the derivatives with respect to coordinates of α) are bounded by constants. As can be seen from (24), when the observed reward at the chosen action j is $r_t = 0$, the posterior $\nu_{\pi^*}(j|j,0)$ for the chosen action will be smaller than its prior belief $q(j)$; and the posteriors $\nu_{\pi^*}(i|j,0)$ will be larger than the prior beliefs $q(i)$ for all unchosen actions $i \neq j$. As a result, we have for every $i \in [K]$,

$$\begin{aligned}
\frac{\partial \texttt{AIR}_{q,\eta}(p,\nu)}{\partial \alpha(i)} =& \frac{1}{\eta}\sum_{j\in[K]} p(j)\log\frac{q(i)}{\nu_{\pi^*}(i|j,0)}\\
\leq& \frac{1}{\eta}p(i)\log\frac{q(i)}{\nu_{\pi^*}(i|i,0)}\\
=& \frac{1}{\eta}p(i)\log\frac{p(i)(\exp(\eta/p(i))-1)}{\exp(\eta)-1}\\
\leq& 1,
\end{aligned}$$

where the first equality is by (23); the first inquality is because $q(i) < \nu_{\pi^*}(i|j,0)$ for all $j \neq i$; the second equality because of (23) and $p = q$; and the last inequality is a consequence of the following application of Jensen's inequality:

$$\frac{1}{1+p(i)}\exp(\eta) + \frac{p(i)}{1+p(i)}\exp\left(-\frac{\eta}{p(i)}\right) \geq 1.$$

Now we have shown that the derivatives of AIR with respect to all $\{\beta_i(j)\}_{i,j\in[K]}$ are zeros, and the derivatives of AIR with respect to all $\{\alpha(i)\}_{i\in[K]}$. We note that AIR is $\frac{1}{\eta}$−strongly convex with respect to α when the gradient with respect to $\{\beta_i(j)\}_{i,j\in[K]}$ are all zeros. Then by Theorem 3.4 and Theorem 3.2 we can prove that

Theorem F.1 (Regret of Simplified APS for Bernoulli MAB)**.** *The regret of Algorithm 4 with $\eta = \gamma = \sqrt{2\log K/(KT+4T)}$ is bounded as follows, for all $T \geq 2K\log K + 4$,*

$$\mathfrak{R}_T \leq \sqrt{(5K+4)T\log K}.$$

F.3. Surrogate concave objective with Gaussian reward

For structured bandit with Gaussian reward structure, we can formulate the optimization problem as a surrogate optimization problem, where all classical upper bounds about information ratio in practical applications apply (e.g., see the square-loss formulation of IR in (Russo & Van Roy, 2016; Lattimore, 2020)). For the simplicity of presentation, we restrict our attention to Gaussian reward with mean bounded in $[0,1]$ and variance $\sigma^2 \leq 1$.

Denote $\alpha = \mathbb{P}_\nu(\pi^*)$ and $\beta_{\pi^*} = \mathbb{P}_\nu(\pi^*)\cdot\theta_{\pi^*}$. Define a variant of AIR as

$$\overline{\texttt{AIR}} = \mathbb{E}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}(N(\theta_{\pi^*}(\pi),1), N(\theta_{\text{avg}}(\pi),1)) - \frac{1}{\eta}\mathrm{KL}(\alpha,q)\right], \tag{25}$$

where we denote

$$\theta_{\text{avg}} = \sum_{\pi^*\in\Pi}\alpha(\pi^*)\theta_{\pi^*}.$$

And we define

$$\begin{aligned}\overline{B}(\nu,\omega) &= \mathbb{E}_{(M,\pi^*)\sim\nu,\pi\sim p}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{2\eta\sigma^2}\left((\theta^\omega_{\pi^*}(\pi)-o)^2 - (\theta^\omega_{\text{avg}}(\pi)-o)^2\right) - \frac{1}{\eta}\log\frac{\alpha^\omega(\pi^*)}{q(\pi^*)}\right]\\ &= \mathbb{E}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{2\eta\sigma^2}\left(\theta^\omega_{\pi^*}(\pi)^2 - \theta^\omega_{\text{avg}}(\pi)^2\right) - \frac{1}{\eta}\log\frac{\alpha^\omega(\pi^*)}{q(\pi^*)}\right] + \frac{1}{\eta\sigma^2}\mathbb{E}\left[o(\theta_{\pi^*} - \theta^\omega_{\text{avg}})\right]. \qquad (26)\end{aligned}$$

Then we have

$$\overline{\texttt{AIR}}_{q,\eta}(p,\nu) = \inf_{\omega}\overline{B}(\nu,\omega).$$

This means that $\overline{\texttt{AIR}}_{q,\eta}(p,\nu)$ will be a concave function of $(\alpha,\{\beta_{\pi^*}\}_{\pi^*\in\Pi})$. And we can develop a parallel theory for approximately optimizing $\overline{\texttt{AIR}}$ as we have done for AIR in Section 3.4. In particular, we need to verify that $\overline{B}(\nu,\omega)$ is always a upper bound of $B(\nu,\mathbb{P}_\omega(\pi^*|\cdot,\cdot))$ in order to derive regret bounds. This is true if we assume the variance $\sigma^2\le 1$ as the normal probability density function is locally concave round its mean.

F.4. Simplified APS for Gaussian linear bandits, relationship with IPW

In this subsection we derive Algorithm 5 for adversarial linear bandits with Gaussian reward. As the decision space is an $d-$dimensional action set $\Pi = \mathcal{A}\subseteq\mathbb{R}^d$, we will use the notations $\mathcal{A}$ (as action set), a (as action) and a^* (as optimal action) to follow the tradition of literature about linear bandits.

By (26) and Lemma E.3, we have for each $a^*\in\mathcal{A}$,

$$\frac{\partial\overline{\texttt{AIR}}_{q,\eta}(p,\nu)}{\partial\beta_{a^*}} = a^* - \mathbb{E}_{a\sim p}[a] - \frac{1}{\eta}\mathbb{E}_{a\sim p}[aa^\top]\left(\theta_{a^*} - \theta_{\text{avg}}\right). \qquad (27)$$

And for each $a^*\in\mathcal{A}$,

$$\frac{\partial\overline{\texttt{AIR}}_{q,\eta}(p,\nu)}{\partial\alpha(a^*)} = -\frac{1}{2\eta}\int_{\mathcal{A}} p(a)\left(\theta_{a^*}^\top a - \theta_{\text{avg}}^\top a\right)^2 da - \frac{1}{\eta}\log\frac{\alpha(a^*)}{q(a^*)}. \qquad (28)$$

Let the derivatives in (27) be zero. If the matrix $\mathbb{E}_{a\sim p}[aa^\top]$ have full rank, then we have

$$\begin{aligned}\theta_{a^*} - \theta_{\text{avg}} &= \eta(\mathbb{E}_{a\sim p}[aa^\top])^{-1}(a^* - \mathbb{E}_{a\sim p}[a]), \quad \forall a^*\in\mathcal{A},\\ \alpha &= p.\end{aligned}$$

Taking the above relationship into (28), we have

$$\frac{\partial\overline{\texttt{AIR}}_{q,\eta}(p,\nu)}{\partial\alpha(a^*)} = -\frac{\eta}{2}(a^* - \mathbb{E}_{a\sim p}[a])(\mathbb{E}_{a\sim p}[aa^\top])^{-1}(a^* - \mathbb{E}_{a\sim p}[a]) - \frac{1}{\eta}\log\frac{\alpha(a^*)}{q(a^*)}.$$

Assume the minimal eigenvalue of $\mathbb{E}_{a\sim p}[aa^\top]$ satisfies $\lambda_{\min}(\mathbb{E}_{a\sim p}[aa^\top])\ge\eta$, then one can verify that the following solution is approximately optimal to the problem (8) (with controllable precision):

$$\begin{aligned}\alpha &= p,\\ \theta_{a^*} &= \eta(\mathbb{E}[aa^\top])^{-1}a^*, \quad \forall a^*\in\mathcal{A}. \qquad (29)\end{aligned}$$

Note that this solution satisfies $\theta_i\in[0,1]^K$ for all $i\in[K]$.

By Bayes' rule and (29), the posterior update $\mathbb{P}_\nu(\pi^*|\pi,r(\pi))$ can be expressed as follows. Given $a^*\in\mathcal{A}$, we have

$$\begin{aligned}\mathbb{P}_\nu(a^* = \bar{a}|a,r(a)) &= \frac{\alpha(\bar{a})\exp(-\frac{1}{2}(r(a) - \theta_{\bar{a}}^\top a)^2)}{\int_{\mathcal{A}} p(a^*)\exp(-\frac{1}{2}(r(a) - \theta_{a^*}(a))^2)da^*}\\ &= \frac{\alpha(\bar{a})\exp\left(r(a)\theta_{\bar{a}}^\top a - \frac{1}{2}(\theta_{\bar{a}}^\top a)^2\right)}{\int_{\mathcal{A}}\alpha(a^*)\exp\left(r(a)\theta_{a^*}^\top a - \frac{1}{2}(\theta_{a^*}^\top a)^2\right)da^*}.\end{aligned}$$

The resulting algorithm is an exponential weight algorithm with a modified importance weight estimator

$$\hat{r}_t(a) = a^\top (\mathbb{E}_{a\sim p}[aa^\top])^{-1} a_t r_t(a_t) - \frac{\eta}{2}(a^\top (\mathbb{E}_{a\sim p}[aa^\top])^{-1} a_t)^2.$$

The forced exploration to ensure $\lambda_{\min}(\mathbb{E}_{a\sim p}[aa^\top]) \geq \eta$ can be done with the help of the volumetric spanners constructed in (Hazan & Karnin, 2016).

G. Proofs for MAIR

G.1. Proof of Lemma C.3

By Definition C.1 we have

$$\begin{aligned}
&\sup_{\rho\in\mathrm{int}(\Delta(\mathcal{M})))} \sup_{\mu\in\Delta(\mathcal{M})} \inf_{p\in\Delta(\Pi)} \mathtt{MAIR}_{\rho_t,\eta}(p,\mu) \\
=& \sup_{\rho\in\mathrm{int}(\Delta(\mathcal{M})))} \sup_{\mu\in\Delta(\mathcal{M})} \inf_{p\in\Delta(\Pi)} \mathbb{E}\left[f_M(\pi_M) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}(M(\pi), \mu_{o|\pi}) - \frac{1}{\eta}\mathrm{KL}(\mu,\rho)\right] \\
=& \sup_{\mu\in\Delta(\mathcal{M})} \inf_{p\in\Delta(\Pi)} \mathbb{E}\left[f_M(\pi_M) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}(M(\pi), \mu_{o|\pi})\right] \\
\leq& \sup_{\bar{M}\in\mathrm{conv}(M)} \mathtt{DEC}_\eta^{\mathrm{KL}}(\mathcal{M},\bar{M}),
\end{aligned}$$

where the second equality is by (11) and the inequality is because Hellinger distance is bounded by KL divergence (Lemma H.6). □

G.2. Analysis of sequential estimation

Consider the optimistic Bayesian posterior update

$$\rho_{t+1}(M) \propto \exp\left(\sum_{s=1}^{t} (\log[M(\pi_s)](o_s) + \eta W_s(M))\right), \tag{30}$$

where $\{W_s\}_{s=1}^T$ is a series of non-negative weights in $[0,1]^M$. When all $W_s(M)=0$ for all M and s, the update reduces to the update of Bayesian posterior. We want to upper bound the cumulative estimation error of updating rule (30). We present the following theorem, whose proof is inspired by (Agarwal & Zhang, 2022; Zhang, 2021).

Theorem G.1. *Applying the updating rule* (30) *with* $W_s(M)\in[0,1]$ *for all* $M\in\mathcal{M}$ *and* $s=1,\ldots,T$*, we have*

$$\sum_{t=1}^T \mathbb{E}_{\mu_t,p_t}[D_H^2(M(\pi), M^*(\pi))] \leq 2\eta + 1 + 2\log(|\mathcal{M}|T) + 4\sum_{t=1}^T \mathbb{E}_{\mu_t,p_t}\left[W_t(M)\right].$$

Proof of Theorem G.1: denote $\mathbb{E}_t[\cdot]$ be the conditional expectation conditioned on the filtration from round 1 to round t. Denote

$$Z_t(M) = \sum_{s=1}^T (\log[M(\pi_s)](o_s)) + \eta W_s(M).$$

We have

$$\begin{aligned}
&\log\left(\sum_{M\in\mathcal{M}} \mathbb{E}_{t-1}\left[Z_t(M)\right]\right) - \log\left(\sum_{M\in\mathcal{M}} Z_{t-1}(M)\right) \\
=&\log\left(\sum_{M\in\mathcal{M}} \frac{Z_{t-1}(M)}{\sum_{M\in\mathcal{M}} Z_{t-1}(M)} \mathbb{E}_{t-1}\left[\exp\left(\log[M(\pi_t)](o_t) + W_t(M)\right)\right]\right) \\
=&\log\left(\sum_{M\in\mathcal{M}} \mu_t(M)\cdot\mathbb{E}_{t-1}\left[[M(\pi_t)](o_t)\right]\cdot\exp\left(W_t(M)\right)\right).
\end{aligned}$$

By the above equality, we have

$$\begin{aligned}
&\log\left(\sum_{M\in\mathcal{M}}\mathbb{E}_{t-1}\left[Z_t(M)\right]\right)-\log\left(\sum_{M\in\mathcal{M}}Z_{t-1}(M)\right)\\
\leq&\sum_{M\in\mathcal{M}}\mu_t(M)\cdot\mathbb{E}_{t-1}\left[[M(\pi_t)](o_t)\right]\cdot\exp\left(W_t(M)\right)-1\\
=&\sum_{M\in\mathcal{M}}\mu_t(M)\cdot\left(\mathbb{E}_{t-1}\left[[M(\pi_t)](o_t)\right]-1\right)\cdot\exp\left(W_t(M)\right)\\
&+\sum_{M\in\mathcal{M}}\mu_t(M)\cdot\exp\left(W_t(M)\right)-1\\
\leq&\mathbb{E}_{\mu_t,p_t}\left[\int_{\mathcal{O}}[M^*(\pi)](o)[M(\pi)](o)do-1\right]+2W_t(M)\\
\leq&\mathbb{E}_{\mu_t,p_t}\left[\int_{\mathcal{O}}\sqrt{[M^*(\pi)](o)[M(\pi)](o)}do-1\right]+2W_t(M)\\
=&-\frac{1}{2}\mathbb{E}_{\mu_t,p_t}\left[D^2_{\mathrm{H}}(M(\pi),M^*(\pi))\right]+2W_t(M),
\end{aligned}$$

where the first inequality is because $\log(1+z)\leq z$ for all $z\in\mathbb{R}$; the second inequality is because $e^z\leq 1+2z$ for all $z\leq[0,1]$ and $W_t(M)\in[0,1]$; and the third inequality is because $[M^*(\pi)](o)\in[0,1]$ by the the fact $M^*(\pi)$ is a probability distribution over $\mathcal{O}$.

Rearrange the above inequality, we conclude that

$$\begin{aligned}
&\mathbb{E}_{\mu_t,p_t}\left[D^2_{\mathrm{H}}(M(\pi),M^*(\pi))\right]\\
\leq&2\log\left(\sum_{M\in\mathcal{M}}Z_{t-1}(M)\right)-2\log\left(\sum_{M\in\mathcal{M}}\mathbb{E}_{t-1}\left[Z_t(M)\right]\right)+4W_t(M). \qquad (31)
\end{aligned}$$

By lemma H.5, for any $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\log\left(\sum_{M\in\mathcal{M}}Z_{t-1}(M)\right)-\log\left(\sum_{M\in\mathcal{M}}\mathbb{E}_{t-1}\left[Z_t(M)\right]\right)\right)\\
\leq&\sum_{t=1}^{T}\left(\log\left(\sum_{M\in\mathcal{M}}Z_{t-1}(M)\right)-\log\left(\sum_{M\in\mathcal{M}}Z_t(M)\right)\right)+\log\frac{1}{\delta}\\
=&\log\left(\sum_{M\in\mathcal{M}}Z_0(M)\right)-\log\left(\sum_{M\in\mathcal{M}}Z_T(M)\right)+\log\frac{1}{\delta}\\
\leq&\eta+\log|\mathcal{M}|+\log\frac{1}{\delta}. \qquad (32)
\end{aligned}$$

Taking $\delta=1/T$ in (32) and applying (31), we can show that

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_{\mu_t,p_t}\left[D^2_{\mathrm{H}}(M(\pi),M^*(\pi))\right]\\
\leq&\max\{1,2\eta+2\log(|\mathcal{M}|T)+4\sum_{t=1}^{T}\mathbb{E}_{\mu_t,p_t}\left[W_t(M)\right]\}\\
\leq&2\eta+1+2\log(|\mathcal{M}|T)+4\sum_{t=1}^{T}\mathbb{E}_{\mu_t,p_t}\left[W_t(M)\right].
\end{aligned}$$

□

G.3. Proof for Theorem C.4

Calculating the optimization error. Similar to Theorem 3.4, we can prove

$$\mathfrak{R}_T \le \frac{\log|\mathcal{M}|}{\eta} + \sum_{t=1}^{T}\left(\mathtt{MAIR}_{\rho_t,\eta}(p_t,\mu_t) + \left\langle \left.\frac{\partial \mathtt{MAIR}_{\rho_t,\eta}(p_t,\mu)}{\partial \mu}\right|_{\mu=\mu_t}, \mathbb{1}(M^*) - \mu_t \right\rangle\right), \tag{33}$$

where $\mathbb{1}(M^*)$ is the vector whose M^*- coordinate is 1 but all other coordinates are 0.

By Lemma E.3 we have

$$\begin{aligned}\frac{\partial \mathtt{MAIR}_{\rho,\eta}(p,\mu)}{\partial \mu(M)} &= f_M(\pi_M) - \mathbb{E}_{\pi\sim p}\left[f_M(\pi)\right] + \frac{1}{\eta}\mathbb{E}_{\pi\sim p, o\sim M(\pi)}\left[\log\frac{\rho(M)}{\mu(M|\pi,o)}\right]\\ &= f_M(\pi_M) - \mathbb{E}_{\pi\sim p}\left[f_M(\pi)\right] - \frac{1}{\eta}\log\frac{\mu(M)}{\rho(M)} - \frac{1}{\eta}\mathbb{E}_{\pi\sim p,\bar{o}\sim M(\pi)}\left[\log\frac{[M(\pi)](\bar{o})}{\mu_o(\bar{o}|\pi)}\right].\end{aligned}$$

By using the updating rule in (12), we have

$$\frac{\partial \mathtt{MAIR}_{\rho,\eta}(p,\mu)}{\partial \mu(M)} = -\frac{1}{\eta}\mathbb{E}_{\pi\sim p,\bar{o}\sim M(\pi)}\left[\log\frac{[M(\pi)](o)}{\mu_o(\bar{o}|\pi)}\right] = -\frac{1}{\eta}\mathbb{E}_{p\sim\pi}\left[\mathrm{KL}(M(\pi),\mu_{o|\pi})\right],$$

which implies

$$\begin{aligned}&\left\langle \left.\frac{\partial \mathtt{MAIR}_{\rho_t,\eta}(p_t,\mu)}{\partial \mu}\right|_{\mu=\mu_t}, \mathbb{1}(M^*) - \mu_t \right\rangle\\ =&\frac{1}{\eta}\mathbb{E}_{\mu_t,p_t}\left[\mathrm{KL}(M(\pi),(\mu_t)_{o|\pi})\right] - \frac{1}{\eta}\mathbb{E}_{\pi\sim p}\left[\mathrm{KL}(M^*(\pi),(\mu_t)_{o|\pi})\right].\end{aligned} \tag{34}$$

Taking (34) into (33), we have

$$\begin{aligned}&\mathfrak{R}_T\\ \le&\frac{\log|\mathcal{M}|}{\eta} + \sum_{t=1}^{T}\mathbb{E}_{\mu_t,p_t}\left[f_M(\pi_M) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}\left(M^*(\pi),(\mu_t)_{o|\pi}\right) - \frac{1}{\eta}\mathrm{KL}(\mu_t,\rho_t)\right].\end{aligned}$$

So we have

$$\mathfrak{R}_T \le \frac{\log|\mathcal{M}|}{\eta} + \sum_{t=1}^{T}\mathbb{E}_{\mu_t,p_t}\left[f_M(\pi_M) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}(\mu_t,\rho_t)\right]. \tag{35}$$

Refined analysis of Algorithm 6. At the same time, we have

$$\begin{aligned}&\mathbb{E}_{\mu_t,p_t}\left[f_M(\pi_M) - f_M(\pi) - \frac{1}{\eta}\mathrm{KL}(\mu_t,\rho_t)\right]\\ =&\mathbb{E}_{\mu_t,p_t}\left[f_M(\pi_M) - f_M(\pi) - \frac{1}{\eta}D^2_{\mathrm{H}}\left(M^*(\pi),M(\pi)\right)\right]\\ &-\frac{1}{\eta}\mathrm{KL}(\mu_t,\rho_t) + \frac{1}{\eta}\mathbb{E}_{\pi\sim p_t,M\sim\mu_t,}\left[D^2_{\mathrm{H}}(M(\pi),M^*(\pi))\right].\end{aligned} \tag{36}$$

Applying Theorem G.1 with

$$W_s(M) = \eta\left(f_M(\pi_M) - \mathbb{E}_{p_s}\left[f_M(\pi)\right]\right),$$

we have that

$$\frac{1}{\eta}\sum_{t=1}^{T}\mathbb{E}_{\mu_t,p_t}\left[D^2_{\mathrm{H}}(M(\pi),M^*(\pi))\right] \le 2 + \frac{2\log(|\mathcal{M}|T)+1}{\eta} + 4\sum_{t=1}^{T}\mathbb{E}_{\mu_t,p_t}\left[f_M(\pi_M) - f_M(\pi)\right]. \tag{37}$$

Combining (35), (36) and (37), we have

$$
\begin{aligned}
&\mathfrak{R}_T\\
\leq&\frac{\log|\mathcal{M}|}{\eta}+\sum_{t=1}^{T}\mathbb{E}_{\mu_t,p_t}\left[f_M(\pi_M)-f_M(\pi)-\frac{1}{\eta}\mathrm{KL}(\mu_t,\rho_t)\right]\\
\leq&\sum_{t=1}^{T}\mathbb{E}_{\mu_t,p_t}\left[5\left(f_M(\pi_M)-f_M(\pi)\right)-\frac{1}{\eta}D_{\mathrm{H}}^2(M(\pi),M^*(\pi))-\frac{1}{\eta}\mathrm{KL}(\mu_t,\rho_t)\right]\\
&+2+\frac{2\log(|\mathcal{M}|T)+1}{\eta}.
\end{aligned}
$$

Therefore, we prove Theorem C.4. □

G.4. Proof of Theorem C.5

Consider the Bayesian posterior sampling strategy induced by $\mu\in\Delta(M)$, which samples $M\sim\mu$ and plays π_M. Denote the induced decision probability as

$$
\mu_{\pi_M}(\pi)=\sum_{M\in\mathcal{M},\pi_M=\pi}\mu(M).
$$

For arbitrary $\mu\in\Delta(M)$, denote the $|\Pi|-$dimensional vectors X,Y by

$$
\begin{aligned}
X(\pi)&=\mathbb{E}_{\mu}\left[f_M(\pi_M)-f_M(\pi)\right],\\
Y(\pi)&=\mathbb{E}_{\mu}\left[D_{\mathrm{H}}^2(\mathbb{P}(o|M,\pi),\mathbb{P}_{\mu}(o|\pi))\right].
\end{aligned}
$$

Then

$$
\begin{aligned}
&\mathbb{E}_{\mu,\pi\sim\rho_{\pi_M}}\left[5\left(f_M(\pi_M)-f_M(\pi)\right)-\frac{1}{\eta}D_{\mathrm{H}}^2(M^*(\pi),M(\pi))\right]-\frac{1}{\eta}\mathrm{KL}(\mu,\rho)\\
\leq&\left\langle\rho_{\pi_M},5X\right\rangle-\frac{1}{\eta}\left\langle\rho_{\pi_M},Y\right\rangle-\frac{1}{\eta}\mathrm{KL}(\mu_{\pi_M},\rho_{\pi_M})\\
\leq&\left\langle\mu_{\pi_M},5X\right\rangle+10\eta-\frac{1}{\eta}\left\langle\rho_{\pi_M},Y\right\rangle-\frac{1}{2\eta}\mathrm{KL}(\mu_{\pi_M},\rho_{\pi_M})\\
\leq&\left\langle\mu_{\pi_M},5X\right\rangle+10\eta-\frac{1}{\eta}\left\langle\rho_{\pi_M},Y\right\rangle-\frac{1}{2\eta}D_{\mathrm{H}}^2(\mu_{\pi_M},\rho_{\pi_M})\\
\leq&\left\langle\mu_{\pi_M},5X\right\rangle-\frac{(1-\eta)}{(1+\eta)\eta}\left\langle\mu_{\pi_M},Y\right\rangle+10\eta\\
\leq&5\mathbb{E}_{\mu,\pi\sim\mu_{\pi_M}}\left[f_M(\pi_M)-f_M(\pi)-\frac{1}{2\eta}D_{\mathrm{H}}^2(M^*(\pi),M(\pi))\right]+2\eta,
\end{aligned}
$$

where the first inequality is because KL divergence between induced decision distributions of two model distributions will be no larger than KL divergence between the two model distributions; the second inequality is by Lemma H.8 and the fact $f_M(\pi)\in[0,1]$ for all M and π; the third inequality is by Lemma H.6; the fourth inequality is a consequence of Lemma H.7 and the AM-GM inequality; and the last inequality uses the condition $\eta\leq\frac{1}{10}$.

Combining the above inequality with Theorem C.4, we prove Theorem C.5.

□

H. Technical Backgrounds

H.1. Conditional entropy

In the discussion after Definition 2.3, we utilize the following important result stating that conditional entropy is concave. The reference provides a succinct proof to this result.

Lemma H.1 (Conditional entropy of a probability measure is concave, (Song, 2019))**.** *Let $\mathbb{P}$ be a probability measure on locally compact space $\mathcal{X}$, and let $\mathfrak{E}, \mathfrak{F}$ be countable partitions of the space. Define the entropy with respect to the partition $\mathfrak{E}$ as*

$$H(\mathbb{P}, \mathfrak{E}) = -\sum_{E \in \mathfrak{E}} \mathbb{P}(E) \log \mathbb{P}(E),$$

and the conditional entropy as

$$H(\mathbb{P}, \mathfrak{E}|\mathfrak{F}) = \sum_{F \in \mathfrak{F}} \mathbb{P}(F) H(\mathbb{P}(\cdot|F), \mathfrak{E}).$$

Then the conditional entropy $H(\mathbb{P}, \mathfrak{E}|\mathfrak{F})$ is a concave function with respect to $\mathbb{P}$.

H.2. Minimax theorem

We introduce the classical minimax theorem for convex-concave game.

Lemma H.2 (Sion's minimax theorem for values, (Sion, 1958))**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be convex and compact sets, and $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ a function which for all $y \in \mathcal{Y}$ is convex and continuous in x and for all $x \in \mathcal{X}$ is concave and continuous in y. Then*

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \psi(x, y) = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \psi(x, y).$$

H.3. Convex analysis

Let $\mathcal{W}$ be a $d-$dimensional convex decision set. Let the potential function $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \infty$ be a proper convex function that is Legendre [Rockafellar, 2015, §26]. We assume $\mathcal{W} \subset \text{dom}(\Psi) := \{u \in \mathbb{R}^d : \Psi(u) < \infty\}$ and bounded diameter of potential, e.g.,

$$\texttt{diam}(\mathcal{W}) := \sup_{u, v \in \mathcal{W}} \Psi(u) - \Psi(v) < \infty.$$

Define the Fenchel-Legendre dual of Ψ as

$$\Psi^*(a) = \sup_{u \in \mathbb{R}^d} \langle a, u \rangle - \Psi(u), \quad \forall a \in \mathbb{R}^d.$$

Define the Bregman divergences with respect to Ψ and Ψ^* as

$$D_\Psi(u, v) = \Psi(u) - \Psi(v) - \langle \nabla \Psi(v), u - v \rangle,$$
$$D_{\Psi*}(a, b) = \Psi^*(a) - \Psi^*(b) - \langle \nabla \Psi^*(b), a - b \rangle.$$

Lemma H.3 (Properties of Legendre function, (Lattimore & Szepesvári, 2020))**.** *If Ψ is a Legendre function, then*

(a) $\nabla \Psi$ is a bijection between $\text{int}(\text{dom}(\Psi))$ *and* $\text{int}(\text{dom}(\Psi^*))$ *with the inverse $(\nabla \Psi)^{-1} = \nabla \Psi^*$. That is, for $u \in$* $\text{int}(\text{dom}(\Psi))$*, if $a = \nabla \Psi(u)$, then $a \in$* $\text{int}(\text{dom}(\Psi^*))$ *and $\nabla \Psi^*(a) = u$;*

(b) $D_\Psi(u, v) = D_{\Psi_*}(\nabla \Psi(v), \nabla \Psi(u))$ for all $u, v \in$ $\text{int}(\text{dom}(\Psi))$*; and*

(c) the Fenchel conjugate Ψ^* is Legendre.

Note that the property (a) in Lemma H.3 is a foundational results in convex optimization—in order to optimize a convex function, one only needs to optimize its Fenchel dual function (in the sense of making gradient small). For example, mirror descent, dual averaging and follow the regularized leader are procedures based on this principle. This property is a special case of Lemma E.3, the "envelop theorem."

We also introduce a property of Bregman divergence.

Lemma H.4 (Generalized Pythagorean theorem)**.** *For a convex function $\Psi : \mathcal{W} \to \mathbb{R} \cup \infty$ and $u, v, w \in \mathcal{W}$, we have*

$$D_\Psi(u, v) - D_\Psi(v, w) - D_\Psi(w, u) = \langle u - w, \nabla \Psi(w) - \nabla \Psi(v) \rangle.$$

H.4. Concentration inequality

We introduce a one-sided martingale concentration inequality from [Foster et al., 2021, Lemma A.4] for a sequence of random variables.

Lemma H.5 (Martingale concentration inequality)**.** *For any sequence of real-valued random variables $\{X_t\}_{t=1}^T$ adapted to a filtration $\{\mathfrak{F}_t\}_{t=1}^T$, it holds that for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $T' \leq T$,*

$$\sum_{t=1}^{T'} X_t \leq \sum_{t=1}^{T'} \log\left(\mathbb{E}_{t-1}[\exp(X_t)]\right) + \log\frac{1}{\delta}.$$

H.5. Information theory

We have the following result stating that the Hellinger distance between two probability measures are smaller than the KL divergence between those two probability measures.

Lemma H.6 (Hellinger distance smaller than KL divergence)**.** *For probability measures $\mathbb{P}$ and $\mathbb{Q}$, the following inequalities hold:*

$$D_{\mathrm{H}}^2(\mathbb{P},\mathbb{Q}) \leq \mathrm{KL}(\mathbb{P},\mathbb{Q}).$$

We introduce a localized version of Pinsker-type inequality using Hellinger distance (which will be stronger than using the KL divergence).

Lemma H.7 (Multiplicative Pinsker-type inequality for Hellinger distance, (Foster et al., 2021))**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on compact space $\mathcal{X}$. For all $h : \mathcal{X} \to \mathbb{R}$ with $0 \leq h(X) \leq R$ almost surely under $\mathbb{P}$ and $\mathbb{Q}$, we have*

$$|\mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}[h(X)]| \leq \sqrt{2R(\mathbb{E}_{\mathbb{P}}[h(X)] + \mathbb{E}_{\mathbb{Q}}[h(X)]) \cdot D_{\mathrm{H}}^2(\mathbb{P},\mathbb{Q})}.$$

We introduce a standard one-sided bound using KL divergence. Compared with Lemma H.7, the upper bound in Lemma H.8 only depends on the probability measure q, while the bound is one-sided and it does not take the square-root from as in Lemma H.7.

Lemma H.8 (Drifted error bound using KL divergence)**.** *For any $p, q \in \Delta(\Pi)$, $\eta > 0$, and any vector $y \in \mathbb{R}^{\Pi}$ where $y(\pi) \leq 1/\eta$ for all $\pi \in \Pi$, we have*

$$\langle y, p-q \rangle - \frac{1}{\eta}\mathrm{KL}(p,q) \leq \eta \sum_{\pi\in\Pi} q(\pi) y(\pi)^2.$$

Proof of Lemma H.8: consider the KL divergence $\psi_{q,\eta}(p) = \frac{1}{\eta}\mathrm{KL}(p||q)$, it is known that the convex conjugate duality of ψ_q is the log partition function

$$\begin{aligned}\psi_{q,\eta}^*(y) &:= \sup_{p\in\Delta(\Pi)} \left\{ \langle y, p \rangle - \frac{1}{\eta}\mathrm{KL}(p||q) \right\} \\ &= \frac{1}{\eta}\log\left(\sum_{\pi\in\Pi} q(\pi)\exp(\eta y(\pi))\right). \qquad (38)\end{aligned}$$

We have

$$\begin{aligned}
&\langle y, p\rangle - \frac{1}{\eta}\mathrm{KL}(p||q)\\
&\leq \frac{1}{\eta}\log\left(\sum_{\pi\in\Pi} q\exp(\eta y(\pi))\right)\\
&\leq \frac{1}{\eta}\log\left(\sum_{\pi} q(\pi)(1+\eta y(\pi)+\eta^2 y(\pi)^2)\right)\\
&= \frac{1}{\eta}\log\left(1+\eta\langle y, q\rangle + \eta^2\sum_{\pi\in\Pi} q(\pi)y(\pi)^2\right)\\
&\leq \langle y, q\rangle + \eta\sum_{\pi\in\Pi} q(\pi)y(\pi)^2,
\end{aligned} \tag{39}$$

where the first equation is because of (38); the second inequality is because $e^z \leq 1+z+z^2$ for all $z \leq 1$ and the last inequality is due to $\log(1+z) \leq z$ for all $z \in \mathbb{R}$. Therefore we have

$$\langle y, p-q\rangle - \frac{1}{\eta}\mathrm{KL}(p||q) \leq \eta\sum_{\pi\in\Pi} q(\pi)y(\pi)^2$$

for all $y \in \mathbb{R}^{|\Pi|}$ where $y(\pi) \leq 1/\eta$ for all π. □