Adversarial Extensions to Information Directed Sampling: A Primer

1 Introduction

The Bayesian perspective offers a simple and elegant framework for tracking uncertainty in online optimization problems. In structured bandit settings, it naturally allows learners to take advantage of correlations between arms using a technique called *information directed sampling* (Russo and Van Roy 2014). The algorithm is not limited to contexts with known priors, however: recent work has adapted the technique for adversarial bandits, allowing for robust learning with minimal assumptions. This survey will introduce information directed sampling and its original information theoretic analysis before showing how minor modifications allow the results to generalize naturally to an adversarial, frequentist setting.

1.1 Bayesian Regret

I will adopt the highly general definition of Bayesian Regret used in Bubeck et al. 2015. Let $\mathcal{F}$ be a prior distribution over sequences of reward functions $f_1, \dots f_T$, where each reward function f_t maps actions A_t from a set $\mathcal{A}$ to rewards Y_t in $\mathbb{R}$. I will denote $Y_{t,a} = f_t(a)$. Let A^* be the best action in hindsight: $A^* = \max_{a \in \mathcal{A}} \sum_{i=1}^T f_i(a)$. This is a random variable, as it depends on the functions sampled from $\mathcal{F}$. Let $A_1, A_2, \dots A_T$ be the actions chosen from a policy π. The *Bayesian Regret* for policy π is defined as

$$BR(\pi) = E\left[\sum_{i=1}^{T} f_i(A^*) - f_i(A_i)\right]$$

In the traditional stochastic multi-armed bandit problem, the $f_t(a) = X_{a,t}$, where for every a, $X_{a,1}, \dots X_{a,T}$ are independent, identically distributed random variables. In general, however, the functions chosen at each round can be correlated or time dependent, making this model much more broadly applicable.

1.2 Beyond the Best Arm

Information directed sampling is motivated by settings in which the reward for different arms are correlated. In such settings, in can be beneficial to play arms you know are suboptimal in order to learn more about what the optimal arm might be. A simple example of this setting comes from a variation of the linear bandit problem. Let the true parameter $\theta^* \in \mathbb{R}^d$ be a one-hot vector chosen uniformly at random. The action set consists of vectors in $\{0,1\}^d$ normalized to be a unit vector in the L_1 norm: $\mathcal{A} = \{\frac{x}{\|x\|_1} | x \in \{0,1\}^d\}$. The reward for action $a \in \mathcal{A}$ is $a^T\theta^*$.

We can interpret this setting as the task of recommending an assortment of products to a new customer. The action set corresponds to sets of products to offer the customer, and the reward is the expected utility the customer gains from the (assumedly randomly chosen) product they buy. The true parameter θ^* corresponds to the product with the highest utility for this customer.

Algorithms that seek to play the best arm at each time like Thompson Sampling and UCB will only play 1-sparse actions (as θ^* is known to come from this set). This can require up to d time-steps to find the best action, as only one arm can be ruled out at each step. A more effective search strategy would be a binary search: start with a $d/2$-sparse action and halve the number of nonzero elements at each step. This rules out half of the remaining actions at each step, finding the result in at most $\log d$ time-steps. We will see later that algorithms based on information directed sampling recover exactly this strategy.

2 The Information Ratio

Policies based solely on estimates of the best arm fail cases like the previous section because they do not account for the *information gain* $g_t(\pi)$ of playing according to policy π at time t. Let $\mathcal{F}_t = \sigma(A_1, Y_1, \dots A_{t-1}, Y_{t-1})$ and A^* be the optimal action. I will use the notation $P_t(\cdot)$ to mean $P(\cdot|\mathcal{F}_t)$. Let $\alpha_t(a) = P_t(A^* = a)$ be the posterior distribution over A^* at time t. The following definitions of information gain are equivalent.

1. It is the expected reduction in entropy $E_t[H(\alpha_t) - H(\alpha_{t+1})]$. Note that this is an $\mathcal{F}_t$ measurable random variable.

2. It is the posterior mutual information between A^* and Y_t:

   $$I_t(A^*, Y_t) = D[P_t(A^*, Y_t) \| P_t(A^*)P_t(Y_t)]$$

   Because it is defined in terms of conditional probability measures, note that this is still an $\mathcal{F}_t$-

measurable random variable. We can derive this from (1) using the conditional entropy formulation of mutual information:

$$\begin{aligned} I_t(A^*, Y_t) &= H(P_t(A^*)) - E_t[H(P_{t+1}(A^*)] \\ &= H(\alpha_t) - E_t[H(\alpha_{t+1})] \end{aligned}$$

3. It is the expectation (over all possible values of A^*) of the KL divergence between $P_t(Y_t|A^*)$ and $P_t(Y_t)$. We can see this from the definition using posterior mutual information:

   $$g_t(\pi) = D[P_t(A^*)P(Y_t|A^*)\|P_t(A^*)P_t(Y_t)]$$

   The $P_t(A^*)$ terms cancel out, leaving $E[D[P_t(Y_t|A^*)\|P_t(Y_t)]]$

Instead of sampling from a policy that minimizes only expected one step regret $\Delta_t = E_t[f_t(A^*) - E_{a\sim\pi}[f_t(a)]]$ as in Thompson sampling, we can divide by a penalty term rewarding information gain. The result, introduced in (Russo and Van Roy 2014), is the *information ratio*, defined at time t as

$$\Gamma_t(\pi) = \frac{\Delta_t^2}{g_t(\pi)}$$

We can also think of the information gain as a function of a specific action a rather than a full randomized policy π. In this case, I will use the overloaded notation $g_t(a)$. The information gain for a policy π is the expected information gain for actions sampled sampled from this policy. As above, we can write three equivalent definitions for this "per-action" information gain:

1. $g(a) = E_t[H(\alpha_t) - H(\alpha_{t+1})|A_t = a]$
2. $g(a) = I_t(A^*, Y_t|A_t = a) = D[P_t(A^*, Y_t|A_t = a)\|P_t(A^*)P_t(Y_t|A_t = a)]$
3. $g(a) = E_t[D[P_t(Y_t|A_t = a, A^*)\|P_t(Y_t|A_t = a)]]$

Similarly, I will write $\Delta_t(a)$ to mean $\Delta_t = E_t[f_t(A^*) - f_t(a)]$.

2.1 The Importance of Randomization

What if, instead of an importance ratio, we used an importance *difference* $\Delta_t - \eta g_t(a)$ for some $\eta > 0$? Because it is a linear function of the probabilities of each action, this difference would be optimized by the deterministic policy which plays $\min_a E[f_t(A^*) - f_t(a)] - \eta g_t(a)$. Unfortunately, policies where A_t is chosen deterministically given $\mathcal{F}_t$ can have linear regret. Consider a two arm setting in which $f_t(a_1) \sim$ Bernoulli(1/2) and $f_t(a_2) \sim$ Bernoulli(p), where the unknown $p \sim \text{Unif}(\{\frac{1}{4}, \frac{3}{4}\})$. If a deterministic policy ever plays a_1 at time t_0, it learns nothing about p, and the posteriors for p and A^* will be unchanged. This means that the policy will continue to play a_1 at every future $t > t_0$, giving linear regret. On the other hand, if we never play a_1, we will always play a_2, once again giving linear regret.

This shows that the randomization we get by optimizing an information *ratio* rather than a difference is fundamentally important to the performance of the algorithm. Using a difference is still possible, but we need to add a regularization term to avoid collapsing to deterministic policies. Specifically, (Xu and Zeevi 2023) define the *Algorithmic Information Ratio* as

$$\begin{aligned} \text{AIR}_q(\pi, \nu) = E_{f\ \sim\nu}[f(A^*) - E_{a\sim\pi} f(a)] \\ - \eta E_{f\ \sim\nu}[D[P(A^*|Y_t)\|P(A^*)] + D[P(A^*)\|q]] \end{aligned}$$

where ν is the prior distribution and q_t is a reference distribution for A^*. This provides a lower bound on the importance ratio. Specifically, for any $x \geq 0, y > 0$,

$$\frac{x^2}{y} = \sup_{\eta>0}(2\eta x - \eta^2 y)$$

We can see this because the supremum will be obtained when $\nabla(2\eta x - \eta^2 y) = 0$, or $\eta = x/y$. At this point, $2\eta x - \eta^2 y = \frac{x^2}{y}$. A consequence of this observation is that

$$\text{AIR} \leq \frac{1}{\eta}\text{IR}$$

We will come back to the AIR later on, as it allows for an adversarial setting, rather than a Bayesian one.

2.2 Bounding Regret with the Information Ratio

The information ratio allows for a convenient upper bound on regret (Russo 2016).

Theorem 1. *If the information ratio is bounded by* Γ *almost surely for each* $t \in \{1 \ldots T\}$*, then for any policy* π*,*

$$E[\mathit{Regret}(T, \pi)] \leq \sqrt{\Gamma H(A^*) T}$$

Proof. First, we can see that the cumulative information gain is upper bounded by the prior entropy.

$$\begin{aligned} E[\sum_{t=1}^{T} g_t] &= E[\sum_{t=1}^{T} E_t[H(\alpha_t) - H(\alpha_{t+1})] \\ &= E[\sum_{t=1}^{T} H(\alpha_t) - H(\alpha_{t+1})] \text{ by the tower property} \\ &= H(\alpha_1) - H(\alpha_{T+1}) \\ &\leq H(\alpha_1) \text{ by the non-negativity of entropy} \end{aligned}$$

This lets us write

$$\begin{aligned} E[\text{Regret}(T,\pi)] &= E\sum_{t=1}^{T} \Delta_t \\ &= E\sum_{t=1}^{T} \sqrt{\Gamma_t}\sqrt{g_t} \\ &\le \sqrt{E\sum_{t=1}^{T} \Gamma_t}\sqrt{E\sum_{t=1}^{T} g_t} \text{ by Holder} \\ &\le \sqrt{\Gamma H(\alpha_1) T} \end{aligned}$$

□

3 The Information Ratio of Thompson Sampling

Using Theorem 1 above, we can bound the regret of any online learning algorithm by finding its information ratio. The first analysis of this form was done for Thompson sampling in Russo 2016. We start with a decomposition of Δ_i for Thompson sampling.

Lemma 2.

$$\Delta_t = \sum_{a \in \mathcal{A}} \alpha_t(a)(E[f_t(a)|A^* = a] - E[f_t(a)])$$

Proof. In Thompson sampling, $P_t(A_t = a)$ is the same as $P_t(A^* = a)$ for all a. This means that

$$\begin{aligned} \Delta_t &= \sum_{a} \alpha_t(a) E[f_t(a^*)|A^* = a] \\ &\qquad - \sum_{a} P(A_t = a) E[f_t(a)]) \\ &= \sum_{a} \alpha_t(a)(E[f_t(a^*)|A^* = a] - E[f_t(a)]) \end{aligned}$$

□

A similar decomposition holds for the information gain.

Lemma 3.

$$g_t(\pi) = \sum_{a,a^*} \alpha_t(a)\alpha_t(a^*) D[P_t(f_t(a)|A^* = a) \| P_t(f_t(a))]$$

Proof.

$$\begin{aligned} g_t(\pi) &= \sum_{a} P_t(A_t = a) g_t(a) \\ &= \sum_{a} P_t(A_t = a) \sum_{a^*} P_t(A^* = a^*) \\ &\quad D[P_t(f_t(a)|A^*) \| P_t(f_t(a))] \end{aligned}$$

□

Working from these decompositions, we can bound the information ratio for Thompson sampling.

Theorem 4. *For Thompson sampling with stochastic k-armed bandits,* $\Gamma_t \le |\mathcal{A}|/2$ *for all* t.

Proof. Using the Cauchy Schwartz inequality on the result of Lemma 2 gives

$$\begin{aligned} \Delta_t^2 &= \Bigg(\sum_{a} \alpha_t(a)(E[f_t(a^*)|A^* = a] \\ &\qquad - \sum_{a} P(A_t = a) E[f_t(a)])\Bigg)^2 \\ &\le |\mathcal{A}| \sum_{a^*} \alpha_t(a^*)^2 (E[f_t(a^*)|A^* = a^*] - E[f_t(a)])^2 \\ &\le |\mathcal{A}| \sum_{a,a^*} \alpha_t(a^*)\alpha_t(a)(E[f_t(a^*)|A^* = a^*] \\ &\qquad - E[f_t(a)])^2 \end{aligned}$$

By Pinsker's inequality,

$$\begin{aligned} &(E[f_t(a^*)|A^* = a^*] \\ &\quad - E[f_t(a)])^2 \le \frac{1}{2} D(P(f_t(a^*)|A^* = a^*) \| P(f_t(a))) \end{aligned}$$

This means

$$\Delta_t^2 \le \frac{|\mathcal{A}|}{2} \sum_{a,a^*} \alpha_t(a)^2 D(P(f_t(a^*)|A^* = a^*) \| P(f_t(a)))$$

We can substitute the result of Lemma 3 into our previous expression to find that $\Delta_t^2 \le \frac{|\mathcal{A}|}{2} g_t$, and therefore $\Gamma_t \le |\mathcal{A}|/2$. □

A similar proof applies when applying Thompson sampling to linear bandit problems.

Theorem 5. *Say* $\mathcal{A} \subset \mathbb{R}^d$ *for which* $|A| = k$*. For all* t*, assume* $E[f_t(a)] = a^T\theta$ *for a shared latent variable* θ*. Then for all* t*,* $\Gamma_t \le d/2$ *almost surely.*

Proof. Define $M \in \mathbb{R}^{k \times s}$ as

$$M_{i,j} = \sqrt{\alpha_t(i)\alpha_t(j)}(E[f_t(a_i)|A^* = a_j] - E[f_t(a_i)])$$

We can also write this as an inner product. Let $\mu = E[\theta]$ and $\mu^j = E[\theta|A^* = a_j]$. Then by linearity of expectation,

$$M_{i,j} = \sqrt{\alpha_t(i)\alpha_t(j)}((\mu^j - \mu)^T a_i)$$

By Lemma 2, we can see that $\Delta_t = \text{Trace } M_{i,j}$. Similarly, by Lemma 3 and Pinsker's inequality,

$$\begin{aligned} g_t(\pi) &\geq 2 \sum_{i,j} \alpha_i \alpha_j (E[f_t(a_i)|A^* = a_j] - E[f_t(a_i)])^2 \\ &= 2\|M\|F \end{aligned}$$

where $\|M\|F$ indicates the Frobenious norm $\sqrt{\text{Trace}(M^T M)}$.

For any matrix M

$$\text{Trace} M \leq \sqrt{\text{Rank}(M)} \|M\|_F$$

This means that $\Gamma_t \leq \frac{\text{Rank}(M)}{2}$. It remains to show that $\text{Rank} M \leq d$. We can decompose M as

$$\begin{bmatrix} \sqrt{\alpha_1}(\mu^1 - \mu)^T \\ \vdots \\ \sqrt{\alpha_k}(\mu^k - \mu)^T \end{bmatrix} \begin{bmatrix} \sqrt{\alpha(1)} a_1 & \ldots & \sqrt{\alpha(k)} a_k \end{bmatrix}$$

This shows that M is the product of a $K \times d$ matrix and a $d \times K$ matrix, so the rank is at most d. □

4 Information Directed Sampling

We can find an algorithm that can achieve a smaller information ratio than Thompson sampling (and therefore a smaller regret) by explicitly finding a policy that minimizes the information ratio at every step. This is called *information directed sampling* or IDS. The algorithm was introduced for the stationary setting in (Russo and Van Roy 2014), and was generalized to arbitrary prior distributions over reward functions in (Bubeck et al. 2015). For each time t we do the following.

1. First, compute the posterior reward distributions given $\mathcal{F}_t$. This lets us compute $E_t[f_t(a)]$ for each a.

2. Using the posterior reward distributions, calculate α_t, the posterior distribution of A^* given $\mathcal{F}_t$. This lets us compute $E[f_t(A^*)]$.

3. Compute $g_t(a)$ for each action a.

4. Together, these quantities let us compute $\Gamma_t(\pi)$ for any π. Find the minimizing π and sample it to get A_t.

The minimization problem in each round of Information Directed Sampling can be simplified by noting

Theorem 6. *The minimizing policy π for IDS will have at most two nonzero components.*

Proof. First, note that the following optimization problems are minimized by the same π.

1. Minimize $\Gamma_t(\pi)$ subject to $\pi^T 1 = 1$.

2. Minimize $\rho(\pi) = (\pi^T \Delta_i)^2 - (\pi^T g)\Gamma^*$, where Γ^* is the optimal objective for (1).

We can see this because if (1) achieves objective Γ^*, then (2) has objective 0. But (2) is non-negative because $(\pi^T \Delta_i)^2 - (\pi^T g)\Gamma^* \propto \Gamma_t - \Gamma^*$. Similarly, if (2) achieves objective 0, then $\Gamma_t - \Gamma^* = 0$, so (1) achieves objective Γ^*. This shows that it suffices to analyze minimizers of (2).

Consider a policy π^* minimizing (2). We can show that every component of the gradient where $\pi^* > 0$ must have the same value d^*. Say there were two nonzero components i and j where $\frac{\partial \rho(\pi^*)}{\partial \pi_i} > \frac{\partial \rho(\pi^*)}{\partial \pi_i}$. Decrease the probability of component i by ϵ while increasing the probability of component j to compensate, staying on the simplex. The objective value will change by $\epsilon(\frac{\partial \rho(\pi^*)}{\partial \pi_j} - \frac{\partial \rho(\pi^*)}{\partial \pi_i}) + O(\epsilon^2)$. We can choose ϵ to be small enough that this will always be negative, showing that π^* cannot be local minimum.

From this argument, we can see that for any component i where $\pi_i^* > 0$, $\frac{\partial}{\partial \pi_i} \rho(\pi^*) = d^* = 2(\Delta^T \pi^*)\Delta_i - g_i^T \Gamma^*$, so $g_i = \frac{2\Delta^T \pi^* \Delta_i - d^*}{\Gamma^*}$. Order the actions that have nonzero probability in π^* in decreasing order of g, giving simplex indices $i_1, i_2, \ldots i_m$. Then $\sum_i \pi_i^* g_i = \beta g_{i_1} + (1-\beta) g_{i_m}$ for some $\beta \in [0,1]$. By substitution, $\sum_i \pi_i^* \Delta_i = \beta \Delta_{i_1} + (1-\beta)\Delta_{i_m}$ as well. This shows that a policy that plays action i_1 with probability β and action i_m with probability $1 - \beta$ has the same expected values of Δ and g as π^*, which means it will attain the same minimum Γ^* in problem 1. □

4.1 Regret

When the rewards for each action are uncorrelated, this approach matches the regret of Thompson sampling. Say the policy minimizing the information ratio at a particular time t is π_{IS}. Let the policy that Thompson sampling would have chosen given the same information $\mathcal{F}_t$, be π_{TH}. Then $\Gamma_t(\pi_{\text{IS}}) \leq \Gamma_t(\pi_{TH})$ for all possible histories $A_1, Y_1, \ldots A_{t-1}, Y_{t-1}$, regardless of what policy these previous actions were sampled from. As $\Gamma_t(\pi_{TH}) \leq |\mathcal{A}|/2$, our regret bound using the information ratio tells us that information directed sampling has regret at most $\sqrt{|\frac{\mathcal{A}|}{2} H(A^*)T}$.

For more structured problems, however, IDS offers an improvement over Thompson Sampling. Returning to the pathological example of linear bandits with known sparsity in the introduction, we can show that the information ratio when using IDS has a constant upper bound. This stands in stark contrast to Thompson sampling which, as we showed in the previous section, has an information ratio that scales with the dimension of the space.

To be precise, we will show that the learning algorithm prescribed by information directed sampling in this case goes as follows:

- Let Θ_t be the set of possible indices for the nonzero element of θ^* consistent with the rewards up to time t.
- Choose half of the elements in Θ_t. Let A_t be a vector with nonzero values at these indices.

Theorem 7. *For IDS on 1-sparse bandit problems,*

$$\Gamma_t \leq \frac{1}{\log 2}$$

Proof. First, we can show that the binary search described above both maximizes the expected information gain at each step and minimizes the expected one-step regret.

Say $|\Theta_t|= m$. Given only the observations in $\mathcal{F}_t$, each element of Θ_t is equally likely. This means that for any a for which all nonzero elements are in Θ_t, $E[a^T\theta^*] = \frac{1}{m}$. Binary search actions therefore minimize Δ_t. Expected one step regret is highest at $t = 1$, where $\Delta_t = 1 - \frac{1}{d}$.

To show that a binary search maximizes information gain, consider an arbitrary action a at time t. Let f be the fraction of indices in Θ_t for which elements in a are nonzero. If, after playing a, we observe a nonzero reward, we will have mf remaining possibilities in Θ_t. If we observe a reward of zero, this leaves $m(1 - f)$ possibilities. This means that the expected entropy in α_{t+1} is

$$f \log mf + (1 - f) \log m(1 - f)$$

The entropy of α_t is just $\log m$. So the expected decrease in entropy is

$$-f \log f - (1 - f) \log(1 - f)$$

This corresponds to the entropy of a Bernoulli distribution with parameter f, for which the known maximizer is $f = \frac{1}{2}$, yielding the binary search strategy. The information gain with IDS in this setting is therefore $\log 2$, making

$$\Gamma_t = \frac{(1 - \frac{1}{d})^2}{\log 2} \leq \frac{1}{\log 2}$$

□

4.2 Example: IDS for Beta-Bernoulli Bandits

To build further intuition about how information directed sampling is performed in practice, consider a stationary stochastic bandit problem where the prior over each arm comes from a Beta distribution. Let each arm a take samples from a Bernoulli distribution with mean μ_a, where $P_t(\mu_a) \sim \text{Beta}(A_{a,t}, B_{a,t})$. Let the highest arm mean be μ^*. To find $g_t(a)$ for each a, we can use the decomposition specified in definition (3). Let $M_{a|a'} = E[\mu_a|\mu'_a = \mu^*]$.

$$\begin{aligned} g_t(a) &= \sum_{a'} \alpha_t(a') D[P_t(Y_{t,a}|A^*) \| P_t(Y_{t,a})] \\ &= \sum_{a'} \alpha_t(a') \bigg(M_{a|a'} \log M_{a|a'} \frac{A_{a,t} + B_{a,t}}{A_{a,t}} \\ &\qquad + (1 - M_{a|a'}) \log(1 - M_{a|a'}) \frac{A_{a,t} + B_{a,t}}{B_{a,t}} \bigg) \end{aligned}$$

To find $\alpha_t(a)$ for each action a:

$$\begin{aligned} \alpha_t(a) &= P_t\left(\bigcap_{a' \neq a} \{\mu_{a'} \leq \mu_a\} \right) \\ &= \int P_t(\mu_a = x) \prod_{a' \neq a} P_t(\mu_{a'} \leq x)\, dx \end{aligned}$$

In general, this integral can be difficult to compute; (Russo and Van Roy 2014) advise using Monte Carlo methods to approximate it. A similar procedure gives

$$M_{a'|a} = \frac{1}{\alpha(a)} \int_0^1 x P(\mu_{a'} = x, \mu_a = \mu^*)\, dx$$

$$= \frac{1}{\alpha(a)} \int_0^1 x P(\mu_a = x) \int_0^x P(\mu_{a'} = y) \prod_{b \neq a, a'} P(\mu_b \leq x)\, dy\, dx$$

Finally, we need $E_t[\mu^*]$. This can be computed with

$$E_t[\mu^*] = \sum_a \alpha_t(a) M_{a|a}$$

Knowing this, we can find $\Delta_t(a) = E_t[\mu^*] - E_t[\mu_a]$.

Let $\overrightarrow{\Delta}$ be a vector where component a is $\Delta_t(a)$ and $\overrightarrow{g}$ be a vector where component a is $g_t(a)$. It remains to find π to minimize

$$\frac{(pi^T \overrightarrow{\Delta})^2}{\pi^T \overrightarrow{g}}$$

We can loop over all pairs of indices i and j and consider the policies $\pi = pe_i + (1 + p)e_j$ for $p \in [0, 1]$. For each pair of indices, choosing an optimal p is a convex optimization problem that is easily approximated with a binary search.

5 Adversarial Learning

The information ratio is a fundamentally Bayesian concept, as it is defined in terms of a prior over the optimal action A^*. This imposes an assumption about the environment; if this assumption fails to match reality, the theoretical guarantees developed so far may no longer hold. To make a bandit algorithm as robust as possible, instead of assuming that the functions $f_1, \dots f_T$ are sampled from a prior, we can analyze how the would perform if they were chosen adversarially. In other words, we want to choose policies $\pi_1, \dots \pi_t$ to minimize

$$\sup_{f_{1:T} \in \mathcal{M}} R(\pi_{1:T}, f_{1:T})$$

where R indicates the regret and $\mathcal{M}$ is a compact class of priors. Bubeck et al. 2015 show that adversarial regret is equivalent to worst case Bayesian regret over all possible priors $\mathcal{F}$ with support in $\mathcal{M}$.

Theorem 8.

$$\min_{\pi_{1:T}} \sup_{f_{1:T} \in \mathcal{M}} R(\pi_{1:T}, f_{1:T}) = \sup_{\mathcal{F}} \min_{\pi_{1:T}} R(\pi_{1:T}, f_{1:T})$$

I will omit the proof, which is a generalization of Sion's minimax theorem. This gives us an idea how how to create algorithms with low adversarial regret: first, consider the worst possible prior $\mathcal{F}$ we could have. Then choose your policy by optimizing the information ratio. Unfortunately, this strategy by itself runs into some problems. An adversary can choose a sequence of priors that lead to arbitrarily small information gain, pushing the information ratio to infinity. This is formalized in (Foster et al. 2023).

6 Using the AIR

The problem above can be avoided if the information gain is used in a *difference* rather than a ratio. As we saw previously, this is possible with the *algorithmic information ratio*. A simplified form of the AIR is

$$\begin{aligned} AIR_q(\pi, \nu) = E_{f \sim \nu}[f(A^*) - E_{a \sim \pi} f(a)] \\ - \eta E_{f \sim \nu}[D[P(A^*|Y_t) \| q]] \end{aligned}$$

where q is a reference distribution. Unlike the standard information ratio, this quantity stays finite even for arbitrarily similar arms.

When we perform information directed sampling using the algorithmic information ratio rather than the standard information ratio and adapt a different worst case prior at each step, the result is called *adaptive minimax sampling*. Specifically, we initialize q_1 to be uniform over all actions, and repeat the following steps at each time t.

1. Find a saddle point ν, π of $AIR_q(\pi, \nu)$.
2. Sample an action from π and observe Y_t.
3. Update q_{t+1} to be the posterior $P(\nu|Y_t)$.

This is essentially the same algorithm as information directed sampling. As we use the algorithmic information ratio instead of the information ratio, however, we get to choose the prior adaptively at each step, producing an algorithm that is prior-free. The question of how exactly to find a saddle point in step 1, however, was never addressed in (Xu and Zeevi 2023); the authors only give an abstract formulation, leaving an implementation for future work. At a high level, $\mathcal{M}$ must be a family of distributions parameterized by some θ for which the minimax problem is tractable.

6.1 Upper Bounds on Regret with AIR

While the algorithmic information ratio is always smaller than the information ratio, it continues to provide an upper bound on regret. Specifically,

Theorem 9. *$R \leq \eta \log|\mathcal{A}| + \sum_{t=1}^T AIR_q$ for any reference distribution q.*

Proof. For any sequence of q_t,

$$\begin{aligned} \sum_{t=1}^T \log \frac{q_{t+1}(A^*)}{q_t(A^*)} &= \log \frac{q_T(A^*)}{q_1(A^*)} \\ &\leq \log|\mathcal{A}| \end{aligned}$$

If we let $q_{t+1} = P(A^*|A_t)$ and take an expectation, we find that

$$\sum_{t=1}^T D[P_t(A^*|A_t), q_t] \leq \log|\mathcal{A}|$$

From the definition of AIR, we know that $\Delta_t = \text{AIR}_q(\pi, P_t(f_t)) + \eta D[P_t(A^*|A_t), q_t])$. As regret is $R_T = \sum_{t=1}^T E\Delta_t$, we get $R_T = \sum_{t=1}^T \text{AIR}_{t,q} + \eta \log|\mathcal{A}|$. □

6.2 Regret of Adaptive Minimax Sampling

As we know that AIR $\leq \eta$IR, we can use the previous theorem to bound the regret of adaptive minimax sampling.

Theorem 10. *The regret of adaptive minimax sampling is*

$$\leq 2\sqrt{\log|\mathcal{A}| T \Gamma_{IDS}}$$

where Γ_{IDS} is an upper bound of the information ratio that information directed sampling sees at any step.

Proof.

$$R \leq \eta \log|\mathcal{A}| + \sum_{t=1}^{T} \mathrm{AIR}_{t,q}$$
$$\leq \eta \log|\mathcal{A}| + \frac{T\Gamma_{\mathrm{IDS}}}{\eta}$$

For ease of notation, let $X = \log|\mathcal{A}|$ and $Y = T\Gamma_{\mathrm{IDS}}$. We can choose η to minimize the result by taking the derivative.

$$X - Y/\eta^2 = 0$$
$$\eta = \sqrt{Y/X}$$

Using this η reduces the upper bound to

$$= \left(\frac{Y}{X}\right)^{1/2} X + \left(\frac{Y}{X}\right)^{-1/2} Y$$
$$= 2\sqrt{Y}\sqrt{X}$$

□

6.3 Adversarial Thompson Sampling

The idea of using Bayesian online optimization algorithms in an adversarial setting by choosing a worst case prior at each step can be applied to Thompson sampling as well in an algorithm called *adaptive posterior sampling* also introduced in (Xu and Zeevi 2023). Initialize π_1 to be uniform over all actions. At each time t

1. Sample an action $A_t \sim \pi_t$ revealing $f_t(A_t)$
2. Find the prior ν minimizing $\mathrm{AIR}_{\pi_t}\pi_t, \nu$.
3. Let π_{t+1} be the posterior distribution of A^* given $f_t(A_t)$, assuming ν as a prior.

This variation allows us to effectively perform Thompson sampling without a prior over the bandit environment, improving its robustness. Especially in nonstationary settings, there may be no obvious choice of prior over the functions f_i, making adaptive posterior sampling easier to use than Thompson sampling in practice.

6.4 Regret of Adaptive Posterior Sampling

Choosing a worst case prior at each time allows Adaptive Posterior Sampling to achieve lower regret than Thompson Sampling.

Theorem 11. *The regret of Adaptive Posterior Sampling is*

$$\leq 2\sqrt{\log|\mathcal{A}|(\Gamma_{TH}/2 + 2)T}$$

where Γ_{TH} is an upper bound on the information ratio obtained in Thompson sampling.

Proof. By (Xu and Zeevi 2023), the AIR for adaptive posterior sampling is at most

$$\frac{1}{\eta}\Gamma_{\mathrm{TH}} + \frac{2}{\eta}$$

Once again, we can use the generic regret bound in terms of the AIR to get.

$$R \leq \eta \log|\mathcal{A}| + \frac{T\Gamma_{\mathrm{TH}}}{\eta} + \frac{2T}{\eta}$$

Deriving with respect to η as before, we find that the expression is minimized at

$$\eta = \sqrt{\frac{T\Gamma_{\mathrm{TH}} + 2T}{\log|\mathcal{A}|}}$$

□

6.5 Lower Bounds on Regret with DEC

A slightly different form of the AIR is the *Decision-Estimation Coefficient* (DEC) (Foster et al. 2023). Instead of balancing the one step regret term with the information ratio, it balances it with a term

$$-\eta D_H^2(P(f_t(A_t)), Q_t(f_t(A_t)))$$

This captures estimation error in the chosen arm. Here, P is the true distribution from which the f_t are sampled, and Q_t is a guess the learner makes about the true distribution at time t. While the AIR provided an upper bound on regret, the DEC provides a lower bound. While the proof is beyond the scope of this survey, we can show that

Theorem 12. *For any online learning algorithm, there exits a choice of prior ν from which the f_t are sampled so that the expected regret is at least the $\frac{DEC}{6} \cdot T$ for all choices of η.*

7 Conclusion

By explicitly balancing one step regret with information gain, the information ratio and its variants (such as the AIR and DEC) allow a single online learning algorithm – information directed sampling – to achieve low regret across stochastic, structured and non-stationary settings. If, instead of keeping the prior fixed during learning, we assume a worst case prior at every step, the same algorithm masters the adversarial setting as well. That said, this approach has its shortcomings. Computing the information ratio at each step will in general require numerical integration, which must be repeated a quadratic number of times in the number of arms during optimization. Without knowledge of a prior capturing a problem's structure, we must solve a costly minimax problem at each step. And after all this, the resulting regret bounds are often no better than that of Thompson sampling on standard problems. Overall, perhaps information directed sampling and adaptive minimax sampling should be seen primarily as algorithm design tools, providing a template which must be refined for each concrete setting to provide practical learning algorithms.

References

[1] Sébastien Bubeck et al. “Bandit Convex Optimization: Root T Regret in One Dimension”. In: *Proceedings of The 28th Conference on Learning Theory*. Conference on Learning Theory. ISSN: 1938-7228. PMLR, June 26, 2015, pp. 266–278.

[2] Dylan J. Foster et al. *The Statistical Complexity of Interactive Decision Making*. arXiv:2112.13487. type: article. arXiv, July 11, 2023.

[3] Daniel Russo. “An Information-Theoretic Analysis of Thompson Sampling”. In: *Journal of Machine Learning Research* (2016).

[4] Daniel Russo and Benjamin Van Roy. “Learning to Optimize via Information-Directed Sampling”. In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014.

[5] Yunbei Xu and Assaf Zeevi. “Bayesian Design Principles for Frequentist Sequential Learning”. In: (June 15, 2023).