Multi-Armed Bandits

Victor Boone* Bruno Gaujal Nicolas Gast

September 29, 2023

Abstract

Quick introduction to Bandits. This course is intented to last about 4 fat hours. I try to construct a comprehensible landscape of tools and notions that appears in the modern theory of bandit algorithms. They are so many things I cannot do justice for; and simply don't even mention so many others.

I intended do make a rather classical introduction to the subject. Mostly frequentist.

A good book on the subject: [Lattimore and Szepesvári, 2020].

Contents

*Refer to me for typos and questions: `victor.boone@univ-grenoble-alpes.fr`

1 A Brief Introduction to Multi-Armed Bandits

1.1 History

So there was this Thompson guy, an Canadian entomologist living in the UK, studying medical trials [Thompson, 1933]. You want to provide medicine to people while dynamically learn the efficiency of drugs, and your purpose is to cure as many patients as possible. Thompson provided a Bayesian rule to address the problem, but barely showed anything about it (checked posteriors are Beta, and that the numerically, the algorithm seems to work for the first 10-ish iterations – well, he had no computer).

Thompson did not give any follow up to his paper. Never, ever.

We have to wait up to 1952 so that Robbins [Robbins, 1952], an american mathematician, to take the problem on. As a statistician, Robbins thought of the problem as a sequential allocation task; Statistics mainly focused on the analysis of experiments in which the observed samples are fixed before-hand. What if the samples are functions of observations themselves, i.e., that the statistician chooses from whose population make samples out of its current knowledge? Robbins already defines what is now refered to as the regret, although his seminal work was still only skratching the surface of what is now know as Multi-Armed Bandit Theory.

The topic has since been exploding.

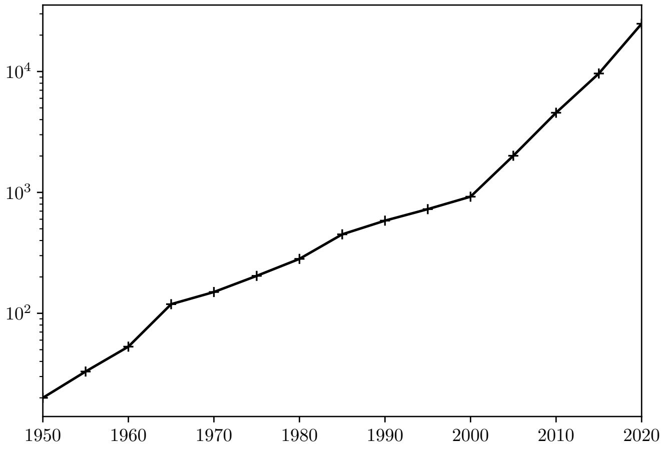

Figure 1: Number of papers refering to “multi-armed bandits” refered by Google Scholar, per 5 year slices.

1.2 Notations and Concepts

The gambler can sample from a set of unknonwn probability distributions $\{\mathrm{F}_a : a\}$ labeled by an action set $\mathcal{A}$ (finite). At time t, she picks an arms A_t and observed a reward $R_t \sim \mathrm{F}(A_t)$, generated independently of all the previous stuff. To simplify the course, we assume distributions are Bernoulli, i.e., $\mathrm{F}_a \equiv \mathrm{B}(\mu_a)$. The goal of the gambler is to maximize the expected sum of rewards, $\mathbf{E}[R_1 + \ldots + R_T]$, or, equivalently, to minimize the expected regret:

$$\mathrm{Reg}(T) := T\mu^* - \sum_{t=1}^{T} R_t. \tag{1}$$

An arm is any arm achieving maximal expected reward $\mu^* := \max_a \mu_a$, and sub-optimal otherwise. While the two tasks are obviously equivalent, the interesting point is that the regret provides a metric, by measuring how behind the algorithm is from optimal performance. The quantity $T\mu^*$ is what you would actually score if you knew everything in advance. The regret measures how far behind the gambler's performance is from the optimal one.

The number of visits of arm a at time T is $N_a(T) := \sum_{t=1}^{T-1} \mathbf{1}\,(A_t = a)$. The number of successes (resp. empirical estimate) of arm a after n pulls of it is denoted $S_{a,n}$ (resp. $\hat{\mu}_{a,n}$). The number of successes (resp. empirical estimate) of arm a at time t is $S_a(t) := S_{a,N_a(t)}$ (resp. $\hat{\mu}_a(t) := \hat{\mu}_{a,N_a(t)}$). The *optimality gap* of arm a is $\Delta_a := \mu^* - \mu_a$. It is the expected cost indured when the gambler picks arm a. The total number of arms is denoted k.

Lemma 1. *The expected regret satisfies:*

$$\mathbf{E}[\mathrm{Reg}(T)] = \sum_a \mathbf{E}[N_a(T)]\Delta_a.$$

Proof. We have $\mathbf{E}[\mathrm{Reg}(T+1)] = \mathbf{E}[\mathrm{Reg}(T)] + \mathbf{E}[\mu^* - R_T]$. The right-term is obtained as follows:

$$\begin{aligned}\mathbf{E}\left[\mu^* - R_t\right] &= \sum_a \sum_r (\mu^* - r)\mathbf{P}(A_T = a, R_T = r)\\ &= \sum_a \left(\sum_r (\mu^* - r)\mathbf{P}(R_T = r | A_T = a)\right)\mathbf{P}(A_T = a)\\ &= \sum_a (\mu^* - \mu_a)\mathbf{P}(A_T = a) = \sum_a \mathbf{E}[\mathbf{1}\,(A_T = a)]\Delta_a.\end{aligned}$$

We conclude by induction. □

Remark. Multi-arm bandits (MAB) are a special case of MDPs. The state space is here trivial, say $\{1\}$, and the action space is $\mathcal{A}$ with reward distributions $\mathrm{B}(\mu_a)$. A deterministic policy is a map $\{1\} \to \mathcal{A}$, hence a choice of arm; It is optimal (discounted, finite-horizon, average) if, and only if it picks an optimal arm.

2 Explore-Then-Commit Algorithms

How should a gambler pick actions so that $\mathbf{E}[\mathrm{Reg}(T)]$ is small? This problem is not easy and is considered as the simplest version of the *exploration-exploitation* *dilemma*. How much should I pick an arm to make sure about its average reward? How quickly should I only pick the arm that only provides the optimal observed reward?

2.1 ETC: How Would the Newcomer Solve Bandits?

The *Explore-Then-Commit* algorithm is a simple way to manage the exploration-exploitation dilemma by decoupling exploration and exploitation. The idea is to pull every arm m times, then to commit everything to the arm that is observed as empirically optimal.

Algorithm 1 Explore-Then-Commit

```
Require: $m \geq 1$ a sampling parameter.
Pick every arm $m$ times;
for $t = mk, \dots$ do
    Pick $A_t$ achieving $\max_a \hat{\mu}_a(mk)$;
end for
```

In the ETC algorithm given in Algorithm 1, how should m be tuned?

2.2 Main Tool: Concentration Inequalities

The material introduced here is about concentration inequalities for random variable supported in $[0, 1]$. We will heavily rely on this from now on. I don't prove Hoeffding's Lemma, because the proof is boring and not especially informative. The result is important and instructive. The proof is not.

Lemma 2 (Hoeffding's Lemma)**.** *Let Y a random variable with $Y \in [a, b]$ with $\mathbf{E}[Y] = 0$. Then for all $t \in \mathbf{R}$, $\mathbf{E}[e^{tY}] \leq \exp(\frac{t^2(b-a)^2}{8})$.*

Lemma 3 (Hoeffding's inequality)**.** *Let (X_k) a sequence of independent variables with $X_k \in [0, 1]$ for all $k \geq 1$. Denote $S_n := X_1 + \ldots + X_n$. Then, for all $\epsilon > 0$,*

$$\mathbf{P}\left(S_n - \mathbf{E}[S_n] \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{n}\right)$$

Proof. Classical use of the Laplace transform (also called "Chernoff method"). Let $\epsilon > 0$ and $t > 0$. We have:

$$\begin{aligned}
\mathbf{P}\left(S_n - \mathbf{E}[S_n] \geq \epsilon\right) &= \mathbf{P}\left(e^{t(S_n - \mathbf{E}[S_n])} \geq e^{t\epsilon}\right) \\
&\leq e^{-t\epsilon}\mathbf{E}\left[e^{t(S_n - \mathbf{E}[S_n])}\right] && \text{(Markov)} \\
&= e^{-t\epsilon}\mathbf{E}\left[\prod_{k=1}^{n} e^{t(X_k - \mathbf{E}[X_k])}\right] \\
&= e^{-t\epsilon}\prod_{k=1}^{n}\mathbf{E}\left[e^{t(X_k - \mathbf{E}[X_k])}\right] && \text{(independance)} \\
&\leq e^{-t\epsilon + \frac{1}{8}t^2 n}. && \text{(Hoeffding's Lemma)}
\end{aligned}$$

Minimizing in t, e.g. $t = \frac{4\epsilon}{n}$, we find $\mathbf{P}(S_n - \mathbf{E}[S_n] \geq \epsilon) \leq \exp(-\frac{2\epsilon^2}{n})$. □

We will use Hoeffding's inequality to control the deviation of empirical estimates (denoted $\hat{\mu}_a(t)$), as follows:

$$\mathbf{P}(\hat{\mu}_{a,n} - \mu_a > \epsilon) \leq \exp(-2n\epsilon^2). \tag{2}$$

2.3 Regret Guarantees of ETC and Tuning

Theorem 1. *The expected regret of ETC(m) is given by:*

$$\mathbf{E}[\mathrm{Reg}(T)] \leq \sum_a \left(m\Delta_a + 2T\Delta_a \exp\left(-\tfrac{1}{2}m\Delta_a^2\right)\right).$$

Proof. For denotational simplicity, assume that $a = 1$ is an optimal arm. This is a direct calculation:

$$\begin{aligned}
\mathbf{E}[\mathrm{Reg}(T)] &= \sum_a \left(m\Delta_a + \sum_{t=mk}^{T} \Delta_a \mathbf{E}\left[\mathbf{1}\left(A_t = a\right)\right]\right) \\
&\leq \sum_a \Delta_a \left(m + \sum_{t=mk}^{T} \mathbf{E}\left[\mathbf{1}\left(\hat{\mu}_{1,m} \leq \hat{\mu}_{a,m}\right)\right]\right) \\
&\leq \sum_a \Delta_a \left(m + \sum_{t=mk}^{T} \mathbf{E}\left[\mathbf{1}\left(\hat{\mu}_{1,m} \leq \mu_1 - \tfrac{\Delta_a}{2}\right) + \mathbf{1}\left(\hat{\mu}_{a,m} \geq \mu_a + \tfrac{\Delta_a}{2}\right)\right]\right) \\
&\leq \sum_a \Delta_a \left(m + 2T\exp(-\tfrac{1}{2}m\Delta_a^2)\right).
\end{aligned}$$

That's it! □

The algorithm is, in general, a bit annoying to tune when $k \geq 3$. But for $k = 2$ this is straight forward. Denote Δ the gap between the optimal and the suboptimal arm. We choose m in order to minimize $m + 2T\exp(-\frac{1}{8}m\Delta^2)$. The function of m is convex with unique minimum that we find by searching a zero of its derivative. We find $m^* = \frac{2}{\Delta^2}\log(T\Delta^2)$. By choosing $m = \lceil m^* \rceil$, ETC(m) achieves performance:

$$\mathbf{E}[\mathrm{Reg}(T)] \leq \frac{2}{\Delta}\log(T\Delta^2) + 3\Delta. \tag{3}$$

Remark. This proof is not tight by about a factor 2, and so is m^*.

Extensions. Thankfully the story doesn't stop here, for many reasons.

- The tuning depends on the horizon T which is supposed to be known in advance. This is not the case in many settings. There is a standard technique to overcome this technicality: The doubling trick. This adds a multiplicative factor of 2 in the regret bounds though.

- This algorithm is far from being optimal – things get even worse when one adds in the doubling trick. In fact, ETC algorithms can never be asymptotically optimal [Garivier et al., 2016], however tight the confidence intervals are chosen; and even if the exploration is non-uniform.

From the second point, we want to claim that algorithms needs to be adaptative. This is not exactly true. There is an algorithm called *Double* Explore-Then-Commit, that adds a second exploration phase after the first exploration phase that helps to "correct" the algorithm's mistakes.

3 UCB and Optimism

The use of optimism is a way to design methods achieving small regret without explicit dependance on the horizon T. Also, this leads to the design of methods that can achieve better performance than ETC. The idea is this: Say arm $a = 1$ has been pulled 100 times with average reward $\hat{\mu}_1 = 0.6$ while arm $a = 2$ has been pulled 10 times with average reward $\hat{\mu}_2 = 0.5$. Do you pull arm 1 or arm 2? To take account of the greater uncertainty about the value of arm $a = 2$, we add a bonus to the average value of arms by considering

$$I_a(t) := \hat{\mu}_a(t) + \mathrm{bonus}(t, N_a(t))$$

where the bonus is increasing in t and decreasing in $N_a(t)$. When picking the arm maximizing $I_a(t)$, this will force the algorithm to overestimate arms that are picked less. This is, roughly speaking, the idea of optimism.

3.1 The "Optimism-in-Face-of-Uncertainty" Principle

The "optimism-in-face-of-uncertainty" principle goes back at least to the seminal paper of [Lai and Robbins, 1985]. The idea is to sample an arm according to an *optimistic* estimate of its value. This optimistic value is usually chosen such that with high enough probability, the optimistic value is an upper-bound of the arm's value. Beware, the question of by how much optimistic we should be is a tricky question, that is still only partially understood today (check [Lattimore, 2018] for a recent overview, but it is only specific to bandits).

A simpler introduction to optimism is due to [Auer et al., 2002] with the famous UCB algorithm. The rule is simple. The optimistic value (or *index*) of an arm consists in its empirically observed value plus a bonus that decreases with the number of visits. Then pick the arm maximizing that index.

The exploration function is $f(t) = 1 + t\log^2(t)$.

3.2 Regret Guarantees of UCB

Theorem 2. *The expected regret of UCB is bounded as:*

$$\mathbf{E}[\mathrm{Reg}(T)] \leq \sum_{a \neq a^*} \Delta_a \inf_{\delta \in (0,1]} \left\{ \frac{\frac{1}{2}\log(T)}{(\Delta_a - \delta)^2} + \frac{2}{\delta^2} + 1 \right\}$$

Algorithm 2 UCB

```
Pick every arm once;
for $t = 1, 2, \dots$ do
    Pick $A_t$ maximizing the index $I_a(t) := \hat{\mu}_a(t) + \sqrt{\frac{\log f(t)}{2N_a(t)}}$;
end for
```

In particular, the expected regret scales with:

$$\limsup_{T\to\infty} \frac{\mathbf{E}[\mathrm{Reg}(T)]}{\log(T)} \leq \sum_{a\neq a^*} \frac{1}{2\Delta_a}.$$

Proof. Without loss of generality, we can assume that the optimal arm is $a^* = 1$, and denote $\Delta_a := \mu^* - \mu_a$. The regret satisfies $\mathbf{E}[\mathrm{Reg}(T)] = \sum_a \mathbf{E}[N_a(T)]\Delta_a$, hence we will upper bound $\mathbf{E}[N_a(T)]$. Let $a \neq 1$ a suboptimal arm. We have:

$$\begin{aligned}
N_a(T) &\leq \sum_{t=1}^{T} \mathbf{1}\left(A_t = a\right) \\
&\leq \sum_{t=1}^{T} \mathbf{1}\left(I_1(t) < \mu_1 - \delta\right) + \sum_{t=1}^{T} \mathbf{1}\left(\hat{\mu}_a(t) > \mu_a + \delta, A_t = a\right) \\
&\quad + \sum_{t=1}^{T} \mathbf{1}\left(I_1(t) \geq \mu_1 - \delta, \hat{\mu}_a(t) \leq \mu_a + \delta, A_t = a\right)
\end{aligned}$$

where $\delta \leq 1$ is an arbitrary positive number. We bound the expectation of each term separately.

For the first one, check that

$$\begin{aligned}
\mathbf{E}[-] &\leq \sum_{t=1}^{T}\sum_{n=1}^{t-1} \mathbf{E}\left[\mathbf{1}\left(\hat{\mu}_1(t) + \sqrt{\frac{\log f(t)}{2N_1(t)}} < \mu_1 - \delta, N_1(t) = n\right)\right] \\
&= \sum_{t=1}^{T}\sum_{n=1}^{t-1} \mathbf{E}\left[\mathbf{1}\left(\hat{\mu}_{1,n} + \sqrt{\frac{\log f(t)}{2n}} < \mu_1 - \delta\right)\right] \\
&\leq \sum_{t=1}^{T}\sum_{n=1}^{t-1} \exp\left(-2n \cdot \left(\sqrt{\frac{\log f(t)}{2n}} + \delta\right)^2\right) \\
&= \sum_{t=1}^{T} \frac{1}{f(t)} \sum_{n=1}^{t-1} \exp(-2n\delta^2) \leq \frac{3}{2\delta^2}.
\end{aligned}$$

The last inequality is mostly calculus. We use $\frac{e^{-c}}{1-e^{-c}} \leq \frac{1}{c}$, that holds for $c > 0$, to show that $\sum_{n\geq 1} \exp(-2n\delta^2) \leq \frac{1}{2\delta^2}$. Then a serie-integral comparison shows that $\sum_{t\geq 1} \frac{1}{f(t)} \leq 3$.

For the second one, check that:

$$\begin{aligned}
\mathbf{E}[-] &= \sum_{t=1}^{T} \mathbf{E}\left[\mathbf{1}\left(\hat{\mu}_a(t) > \mu_a + \delta, A_t = a\right)\right] \\
&= \sum_{t=1}^{T} \mathbf{E}\left[\sum_{n=0}^{t-1} \mathbf{1}\left(\hat{\mu}_{a,n} > \mu_a + \delta, A_t = a, N_a(t) = n\right)\right] \\
&\leq 1 + \sum_{n=1}^{T-1} \mathbf{E}\left[\mathbf{1}\left(\mu_{a,n} > \mu_a + \delta\right)\right] \\
&\leq 1 + \sum_{n=1}^{T-1} \exp(-2n\delta^2) \leq 1 + \frac{1}{1 - e^{-2\delta^2}} \leq 1 + \frac{1}{2\delta^2}.
\end{aligned}$$

For the last inequality, we use again $\frac{e^{-c}}{1-e^{-c}} \leq \frac{1}{c}$, that holds for $c > 0$.

For the third term, we have:

$$\begin{aligned}
\mathbf{E}[-] &\leq \sum_{t=1}^{T-1} \mathbf{E}\left[\mathbf{1}\left(\mu_a + \delta + \sqrt{\frac{\log(t)}{2N_a(t)}} \geq \mu_1 - \delta, A_t = a\right)\right] \\
&\leq \sum_{t=1}^{T-1} \mathbf{E}\left[\mathbf{1}\left(N_a(t) \leq \frac{\log(t)}{2(\Delta_a - 2\delta)^2} \geq \mu_1, A_t = a\right)\right] \\
&\leq \frac{\log(T)}{2(\Delta_a - \delta)^2}.
\end{aligned}$$

In the end, we find that for all $\delta > 0$, we have:

$$\mathbf{E}[N_a(T)] \leq 1 + \frac{2}{\delta^2} + \frac{\frac{1}{2}\log(T)}{(\Delta_a - 2\delta)^2}.$$

To optimize in $\delta > 0$ asymptotically in T, pick $\delta \equiv \delta(T) = \log^{-1/4}(T)$. □

3.3 About Tuning UCB

UCB can be tuned, because the index can be parametrized as follows:

$$\hat{\mu}_a(t) + \sqrt{\frac{\alpha \log f(t)}{N_a(t)}}$$

where $\alpha > 0$. In the above, α is the related to the power level (of t) for which the confidence interval $\hat{\mu}_a(t) \pm \sqrt{\alpha \log f(t) / N_a(t)}$ holds. In theory, choosing $\alpha < \frac{1}{2}$ may lead to $\Omega(\log(T))$ expected regret. In practical scenarios, this is quite often that α is chosen very small, sometimes close to 0. Also, $f(t)$ is usually chosen as $f(t) = 1 + t$ instead of the weird $1 + t\log^2(t)$.

Although the expected regret may be $\Omega(\log(T))$, one can show that for all $\alpha > 0$, UCB eventually ends up picking mostly optimal arms. The thing is that the probability of UCB mistaking the suboptimal arm for the optimal one may be large enough so that the *expected* regret is big. This probability nonetheless goes to 0 as $T \to \infty$.

4 Lower Bounds

So we have presented ETC. Said that it was not optimal. We suggested UCB, that has better asymptotic guarantees. Can we do better? How efficient can an algorithm be? It depends.

Consider the algorithm that only picks the arm $a = 1$. If the arm $a = 1$ is optimal by any chance, then the algorithm will have *null* regret. If it isn't, then $\mathbf{E}[\mathrm{Reg}(T)] = \Omega(T)$. This algorithm is not very interesting however, because it doesn't work on every instance. This leads to the following definition:

Definition 1. An algorithm is said to be *uniformly consistent* if for all distribution F on arms, for all $\epsilon > 0$, we have $\mathbf{E}_{\mathrm{F}}[\mathrm{Reg}(T)] = \mathrm{o}(T^{\epsilon})$; or equivalently, if whenever a is a suboptimal arm under F, we have $\mathbf{E}_{\mathrm{F}}[N_a(T)] = \mathrm{o}(T^{\epsilon})$.

We will only consider uniformly consistent algorithm from now on.

To lower bound the expected regret of an algorithm, the idea is to relate what the algorithm is doing on the bandit model F to what it is doing to another bandit model F′; Because whatever happens on F′ has some positive probability of happening on F as well. If the algorithm is consistent and arm $a = 1$ is optimal under F′, then it has a large probability to pick arm a a lot when running under F′. Since everything that happens under F′ has a positive probability of happening under F, it means that the large number of visits of $a = 1$ under F′ force the algorithm to visit $a = 1$ a lot with positive probability under F, even though a may not be optimal under F. Taking the expectation, this produces a lower bound on $\mathbf{E}_{\mathrm{F}}[N_a(T)]$.

Now, we will make this formal.

4.1 Changes of Measure

Fix a learning algorithm and assume it is deterministic for splicity, so that A_t is a deterministic function of $(A_1, R_1, \dots, A_{t-1}, R_{t-1}) := H_t$. We say that A_t is determined by the *history* H_t.

First, we need to relate what the algorithm is doing on F to what it is doing on F′, so consider two distributions F and F′ on arms. Denote f_a (resp. f'_a) the p.d.f. of arm a under F (resp. F′). Because the algorithm is deterministic, the probability of observing the current history under F is

$$\prod_{t=1}^{T-1} f_{A_t}(R_t).$$

This is called the *likelihood* of H_T. The likelihood ration between F and F′ is $\prod_{t=1}^{T-1} f_{A_t}(R_t)/f'_{A_t}(R_t)$. A quantity which is equivalent to it, and very important in statistics, is the *log-likelihood ratio* of the observations up to time T under a fixed learning algorithm, given by

$$L_T \equiv L_T(A_1, R_1, \dots, A_{T-1}, R_{T-1}) := \sum_{t=1}^{T-1} \log\left(\frac{f_{A_t}(R_t)}{f'_{A_t}(R_t)}\right). \tag{4}$$

The log-likelihood ratio is useful to change measures, as driven by Lemma 4.

Lemma 4. *Let E a $\sigma(H_T)$-measurable event. Then* $\mathbf{P}_{\mathrm{F}'}(E) = \mathbf{E}_{\mathrm{F}}[\mathbf{1}\left(E\right)\exp(-L_T)]$.

Proof. This is actually a fancy way of saying something simple. To see what happens, let us assume that there is a single arm; so that $L_T \equiv L_T(R_1, \ldots, R_{T-1}) = \sum_{t=1}^{T-1} \log(f(R_t)/f'(R_t))$ where I drop the subscript on f (there is a single action). A history is then a sequence of observed rewards of length T, and a $\sigma(H_T)$-measurable event is a set of T-histories. We have:

$$\begin{aligned}
\mathbf{P}_{\mathrm{F}'}(E) &= \sum_{h\equiv(r_1,\ldots,r_{t-1})} \mathbf{1}\left(h \in E\right) \prod_{i=1}^{t-1} f'(r_i) \\
&= \sum_{h\equiv(r_1,\ldots,r_{t-1})} \mathbf{1}\left(h \in E\right) \exp\left(-\sum_{i=1}^{t-1} \log\left(\frac{f(r_t)}{f'(r_t)}\right)\right) \prod_{i=1}^{t-1} f(r_i) \\
&\equiv \sum_{h\equiv(r_1,\ldots,r_{t-1})} \mathbf{1}\left(h \in E\right) \exp(-L_T(h)) \prod_{i=1}^{t-1} f(r_i) = \mathbf{E}_{\mathrm{F}}[\mathbf{1}\left(E\right)\exp(-L_T)].
\end{aligned}$$

The proof is the same for $k \geq 2$, but becomes denotationally much more involved. □

The next result is *the* central result from which we will derive lower bounds on achievable performance. It may look a little bit shy or technical, but it is nonetheless very important.

Lemma 5. *If E is $\sigma(H_T)$-measurable, then* $\mathbf{E}_{\mathrm{F}}[L_T] \geq \mathrm{kl}(\mathbf{P}_{\mathrm{F}}(E), \mathbf{P}_{\mathrm{F}'}(E))$.

Proof. So E is a set of T-histories. By the previous lemma, we have:

$$\begin{aligned}
\mathbf{P}_{\mathrm{F}'}(E) &= \mathbf{E}_{\mathrm{F}}[\mathbf{1}\left(E\right)\exp(-L_T)] \\
&= \sum_{h\in E} \exp(-L_T(h))\mathbf{P}_{\mathrm{F}}(H_T = h) \\
&= \mathbf{P}_{\mathrm{F}}(E) \sum_{h\in E} \exp(-L_T(h)) \tfrac{\mathbf{P}_{\mathrm{F}}(H_T=h)}{\mathbf{P}_{\mathrm{F}}(E)} \\
&\geq \mathbf{P}_{\mathrm{F}}(E) \exp\left(-\sum_{h\in E} L_T(h) \tfrac{\mathbf{P}_{\mathrm{F}}(H_T=h)}{\mathbf{P}_{\mathrm{F}}(E)}\right) \equiv \mathbf{P}_{\mathrm{F}}(E)\exp(-\mathbf{E}[L_T|E]).
\end{aligned}$$

Similarly, with the same proof for $E^{\complement}$, we get $\mathbf{P}_{\mathrm{F}'}(E^{\complement}) \geq \mathbf{P}_{\mathrm{F}}(E^{\complement})\exp(-\mathbf{E}[L_T|E^{\complement}])$. Therefore,

$$\begin{aligned}
\mathbf{E}_{\mathrm{F}}[L_T] &= \mathbf{E}_{\mathrm{F}}[L_T|E]\mathbf{P}_{\mathrm{F}}(E) + \mathbf{E}_{\mathrm{F}}[L_T|E^{\complement}]\mathbf{P}_{\mathrm{F}}(E^{\complement}) \\
&\geq \mathbf{P}_{\mathrm{F}}(E)\log\left(\frac{\mathbf{P}_{\mathrm{F}}(E)}{\mathbf{P}_{\mathrm{F}'}(E)}\right) + \mathbf{P}_{\mathrm{F}}(E^{\complement})\log\left(\frac{\mathbf{P}_{\mathrm{F}}(E^{\complement})}{\mathbf{P}_{\mathrm{F}'}(E^{\complement})}\right) \equiv \mathrm{kl}(\mathbf{P}_{\mathrm{F}}(E), \mathbf{P}_{\mathrm{F}'}(E)).
\end{aligned}$$

When distributions are not discrete, the proof is essentially the same, but the material is a bit more advanced. □

4.2 Asymptotical Lower Bound

The last element that we need to understand where this is all doing is the following expression of $\mathbf{E}_{\mathrm{F}}[L_T]$, obtained by Wald's equation (or by induction on T):

$$\mathbf{E}_{\mathrm{F}}[L_T] = \sum_a \mathbf{E}_{\mathrm{F}}[N_a(T)]\mathrm{kl}(\mu_a, \mu'_a). \tag{5}$$

Proof. Observe that $\mathbf{E}_{\mathrm{F}}[L_{T+1}] = \mathbf{E}_{\mathrm{F}}[L_T] + \mathbf{E}_{\mathrm{F}}[\log(f_{A_T}(R_T)/f'_{A_T}(R_T))]$. So we prove this by induction. We have:

$$\begin{aligned}
\mathbf{E}_{\mathrm{F}}\left[\log\left(\frac{f_{A_T}(R_T)}{f'_{A_T}(R_T)}\right)\right] &= \sum_a \sum_{r\in\{0,1\}} \log\left(\frac{f_a(r)}{f'_a(r)}\right)\mathbf{P}_{\mathrm{F}}(R_T = r|A_T = a)\mathbf{P}_{\mathrm{F}}(A_T = a) \\
&= \sum_a \left(\sum_{r\in\{0,1\}} f_a(r)\log\left(\frac{f_a(r)}{f'_a(r)}\right)\right)\mathbf{E}_{\mathrm{F}}\left[\mathbf{1}\left(A_T = a\right)\right] \\
&= \sum_a \mathrm{kl}(\mu_a, \mu'_a)\mathbf{E}_{\mathrm{F}}\left[\mathbf{1}\left(A_T = a\right)\right]
\end{aligned}$$

Overall, we get by induction: $\mathbf{E}_{\mathrm{F}}[L_{T+1}] = \sum_a \mathrm{kl}(\mu_a, \mu'_a)\sum_{t=1}^T \mathbf{E}_{\mathrm{F}}[\mathbf{1}\left(A_t = a\right)]$. □

This is very similar to the expected regret $\mathbf{E}_{\mathrm{F}}[\mathrm{Reg}(T)] = \sum_a \mathbf{E}_{\mathrm{F}}[N_a(T)]\Delta_a$. To us, the take-away from the above expression is that, combined with Lemma 5, we know that if we find an event E which is rare under F but is frequent under F′, then we will have

$$\sum_a \mathbf{E}_{\mathrm{F}}[N_a(T)]\mathrm{kl}(\mu_a, \mu'_a) \geq \mathrm{kl}(\mathbf{P}_{\mathrm{F}}(E), \mathbf{P}_{\mathrm{F}'}(E)) \gg 1.$$

Therefore, there must a $\mathbf{E}_{\mathrm{F}}[N_a(T)]\mathrm{kl}(\mu_a, \mu'_a) \gg 1$. We will be able to deduce that a properly selected $N_a(T)$ must have high enough expected value.

Theorem 3. *Fix* F *the distributions over arms, with* $\mathrm{F}_a \equiv \mathrm{B}(\mu_a)$*. Every uniformly consistent algorithm satisfies:*

$$\liminf_{T\to\infty} \frac{\mathbf{E}_{\mathrm{F}}[N_a(T)]}{\log(T)} \geq \frac{1}{\mathrm{kl}(\mu_a, \mu^*)}.$$

Proof. Let a a suboptimal arm under F and assume, without loss of generality, that the optimal arm under F is $a = 1$. Let $\delta > 0$ small enough. Let F′, similar in every way to F, excepted that $\mu'_a = \mu_1 + \delta$; hence a is optimal by exactly δ. Therefore, F′ is chosen so that:

$$\mathbf{E}_{\mathrm{F}}[L_T] = \mathbf{E}_{\mathrm{F}}[N_a(T)]\mathrm{kl}(\mu_a, \mu_1 + \delta).$$

Consider the event

$$E := \left(N_a(T) \geq \tfrac{1}{2}T\right).$$

The idea is that this event is eventually very probable under F′, but is eventually very rare under F. Fix $\epsilon > 0$.

Under F, on the one hand, we have:

$$\mathbf{E}_{\mathrm{F}}[N_a(T)] \geq \tfrac{1}{2}T\mathbf{E}\left[\mathbf{1}\left(N_a(T) \geq \tfrac{1}{2}T\right)\right].$$

On the over hand, because a is suboptimal we have $\Delta_a \mathbf{E}_{\mathrm{F}}[N_a(T)] \leq \mathbf{E}_{\mathrm{F}}[\mathrm{Reg}(T)]$, which is eventually smaller than T^ϵ by uniform consistency. Both together, we obtain

$$\mathbf{E}\left[\mathbf{1}\left(N_a(T) \geq \tfrac{1}{2}T\right)\right] \leq \tfrac{2}{\Delta_a}T^{\epsilon-1}. \tag{6}$$

Under F′, on the one hand we have

$$\begin{aligned}\mathbf{E}_{\mathrm{F}'}[N_a(T)] &\leq \tfrac{1}{2}T\left(1 - \mathbf{E}_{\mathrm{F}'}\left[\mathbf{1}\left(N_a(T) \geq \tfrac{1}{2}T\right)\right]\right) + T\mathbf{E}_{\mathrm{F}'}\left[\mathbf{1}\left(N_a(T) \geq \tfrac{1}{2}T\right)\right] \\ &= \tfrac{1}{2}T + \tfrac{1}{2}T\mathbf{E}_{\mathrm{F}'}\left[\mathbf{1}\left(N_a(T) \geq \tfrac{1}{2}T\right)\right].\end{aligned}$$

On the other hand and provided that $\delta > 0$ is small enough, we have $\mathbf{E}_{\mathrm{F}'}[\mathrm{Reg}(T)] \geq \delta(T - \mathbf{E}_{\mathrm{F}'}[N_a(T)])$ (because a is the only optimal arm, and is δ-optimal), and by uniform consistency $\mathbf{E}_{\mathrm{F}'}[\mathrm{Reg}(T)] \leq T^\epsilon$ eventually. Rearraging terms, we see that $\mathbf{E}_{\mathrm{F}'}[N_a(T)] \geq T - \delta^{-1}T^\epsilon$. Overall,

$$\begin{aligned}\mathbf{E}_{\mathrm{F}'}\left[\mathbf{1}\left(N_a(T) \geq \tfrac{1}{2}T\right)\right] &\geq \tfrac{2}{T}\mathbf{E}_{\mathrm{F}'}\left[N_a(T)\right] - 1 \\ &\geq 1 - \tfrac{2}{\delta}T^{\epsilon-1}.\end{aligned} \tag{7}$$

We now conclude the proof. We get:

$$\begin{aligned}\mathbf{E}_{\mathrm{F}}[N_a(T)]\mathrm{kl}(\mu_a, \mu_1 + \delta) &\geq \mathrm{kl}\left(\tfrac{2}{\Delta_a}T^{\epsilon-1}, 1 - \tfrac{2}{\delta}T^{\epsilon-1}\right) \\ &\underset{T\to\infty}{\sim} \log\left(T^{1-\epsilon}\right) = (1-\epsilon)\log(T).\end{aligned}$$

Therefore,

$$\liminf_{T\to\infty} \frac{\mathbf{E}_{\mathrm{F}}[N_a(T)]}{\log(T)} \geq \frac{1-\epsilon}{\mathrm{kl}(\mu_a, \mu_1 + \delta)}. \tag{8}$$

This holds for all $\delta > 0$ and $\epsilon > 0$. Conclude by summing over arms. □

A few remarks.

- This analysis is inspired from [Kaufmann et al., 2016].
- We observe that UCB is nowhere close to achieving the lower bound for Bernoulli bandits. However, we can show that UCB is optimal for bandits with Gaussian rewards, where the standard deviations are known. Our proof would need to be generalized to Gaussian distributions.
- This lower bound inspired the design of many asymptotically optimal algorithms: KL-UCB, MED, IMED are the most important. The design of MED was directly inspired by the lower bound, as written in the original paper [Honda and Takemura, 2010]. The idea is to sample arm a with probability proportional to $\exp(-N_a(t)\mathrm{kl}(\hat{\mu}_a(t), \hat{\mu}^*(t))$. These three algorithms greatly outperform UCB in practice.

- The algorithm of Thompson [Thompson, 1933], mentionned in the introduction, is asymptotically optimal [Kaufmann et al., 2012] but it was only shown recently.

5 Minimax Analysis, Quickly

One issue that people have with the lower bound of Theorem 3, is that it may be vacuous. When T is fixed, we can always choose F such that $\mathrm{kl}(\mu_a, \mu_1)$ is so small that $(\mu_1 - \mu_a) \cdot \frac{\log(T)}{\mathrm{kl}(\mu_a,\mu_1)} \gg T$. The issue is that when $\mu_1 \approx \mu_a$, $\mathrm{kl}(\mu_a, \mu_1) \approx \frac{(\mu_1-\mu_a)^2}{\mu_1(1-\mu_1)}$, so that

$$\frac{(\mu_1 - \mu_a)\log(T)}{\mathrm{kl}(\mu_a, \mu_1)} \approx \frac{\mu_1(1-\mu_1)\log(T)}{\mu_1 - \mu_a}.$$

This issue, by the way, already arises in UCB's regret upper bound. This raises one question: If a run a given algorithm, what is

$$\max_{\mathrm{F}} \mathbf{E}_{\mathrm{F}}[\mathrm{Reg}(T)] \qquad ?$$

When, of course, F lives in a fixed space of probability distributions (e.g. Bernoulli).

5.1 Getting Minimax Bounds for Free

For simplicity, assume that $\mathcal{A} = \{1, 2\}$ (only two arms) and denote $\Delta := \max \mu_a - \min \mu_a$ the suboptimality gap.

Proposition 1. *Assume that the algorithm is such that there exists a universal constant $C > 0$ such that, for all F, $\mathbf{E}_{\mathrm{F}}[\mathrm{Reg}(T)] \leq \frac{C}{\Delta}\log(T)$. Then*

$$\max_{\mathrm{F}} \mathbf{E}_{\mathrm{F}}[\mathrm{Reg}(T)] \leq \sqrt{CT\log(T)}.$$

Proof. We have $\mathbf{E}_{\mathrm{F}}[\mathrm{Reg}(T)] \leq \min\left\{\Delta T, \frac{C}{\Delta}\log(T)\right\}$. The left term ΔT is increasing with Δ, while the second $\frac{C}{\Delta}\log(T)$ is decreasing with Δ. The two are equal when $\Delta T = \frac{C}{\Delta}\log(T)$, i.e., when

$$\Delta = \sqrt{\frac{C\log(T)}{T}}.$$

So $\min\left\{\Delta T, \frac{C}{\Delta}\log(T)\right\} \leq \sqrt{CT\log(T)}$. □

5.2 A Few Words About the Lower Bound

Theorem 4. *Every algorithm must satisfy* $\max_{\mathrm{F}} \mathbf{E}_{\mathrm{F}}[\mathrm{Reg}(T)] = \Omega(\sqrt{kT})$.

Hence, the $\sqrt{T\log(T)}$ in the previous proposition is not tight. Getting rid of this extra $\sqrt{\log(T)}$ is not easy, and algorithms that are *minimax optimal* (achieving the $\sqrt{kT}$, see MOSS [Bubeck and Cesa-Bianchi, 2012]) are usually different from those why good logarithmic regret guarantees.

One active area of research is the design of algorithms that are doubly optimals: both achieving minimal optimality and asymptotic optimality.

References

[Auer et al., 2002] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-Time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.*, 47(2–3):235–256.

[Bubeck and Cesa-Bianchi, 2012] Bubeck, S. and Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems.

[Garivier et al., 2016] Garivier, A., Lattimore, T., and Kaufmann, E. (2016). On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, 29.

[Honda and Takemura, 2010] Honda, J. and Takemura, A. (2010). An Asymptotically Optimal Policy for Finite Support Models in the Multiarmed Bandit Problem.

[Kaufmann et al., 2016] Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17(1):1–42.

[Kaufmann et al., 2012] Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson Sampling: An Asymptotically Optimal Finite Time Analysis. *arXiv:1205.4217*.

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

[Lattimore, 2018] Lattimore, T. (2018). Refining the confidence level for optimistic bandit strategies. *The Journal of Machine Learning Research*, 19(1):765–796. Publisher: JMLR. org.

[Lattimore and Szepesvári, 2020] Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

[Robbins, 1952] Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535. Publisher: American Mathematical Society.

[Thompson, 1933] Thompson, W. R. (1933). On the Likelihood that One Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3-4):285–294.