Optimal algorithms for sequential resource allocation

Emilie Kaufmann,
post-doc in the DYOGENE team

Sequential allocation: some examples

Clinical trial

- K possible treatments (with unknown effect)

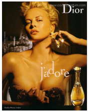


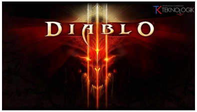



- Which treatment should be allocated to each patient based on their effect on previous patients?

Online advertisement

- K possible ads to display





- Which ad should be displayed to each user, based on the previous clicks of previous (similar) users?

Options have random outcomes

Clinical trials

The effect of treatment a is modeled by a Bernoulli distribution with probability of success p_a:

$$\begin{aligned} \mathbb{P}(X = 1) &= p_a \\ \mathbb{P}(X = 0) &= 1 - p_a \end{aligned}$$

X is a random variable indicating if the patient is cured or not.

The 'bandit' framework

One-armed bandit
= slot machine (or arm)

<u>Multi-armed bandit:</u> several arms.
Drawing arm $a \Leftrightarrow$ observing a sample from a distribution ν_a, with mean p_a

Best arm $a^* = \text{argmax}_a \; p_a$

Which arm should be drawn based on the previous observed outcomes?

Outline

A multi-armed bandit problem

What are optimal bandit algorithms?

The UCB approach

Bayesian bandit algorithms

Outline

A multi-armed bandit problem

What are optimal bandit algorithms?

The UCB approach

Bayesian bandit algorithms

Bandit model with Bernoulli rewards

K arms:

- arm $a \rightarrow$ distribution $\mathcal{B}(p_a)$ (unknown parameter)
- **Unknown** best arm

$$a^* = \text{argmax}_a \; p_a \qquad\qquad p^* = \max_a \; p_a$$

An agent:

- draws arm A_t at time t
- observes the reward: $X_t \sim \mathcal{B}(p_{A_t})$

(A_t) is his **strategy** or **bandit algorithm**:

$$A_{t+1} \;\text{ must depend on }\; A_1, X_1, \ldots, A_t, X_t.$$

A 'bandit problem'

The agent wants to ajust (A_t) to

- maximize the (expected) sum of rewards accumulated,

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$$

- or equivalently minimize his *regret*:

$$R_T = Tp^* - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$$

Outline

A multi-armed bandit problem

What are optimal bandit algorithms?

The UCB approach

Bayesian bandit algorithms

Our optimality criterion

$N_a(t)$: number of draws of arm a up to time t

$$R_T = \sum_{a=1}^{K} (p^* - p_a)\mathbb{E}[N_a(T)]$$

- [Lai and Robbins 1985]: every 'uniformly good' bandit algorithm satisfies

$$p_a < p^* \;\Rightarrow\; \mathbb{E}[N_a(T)] \geq \frac{\log T}{d(p_a, p^*)} \text{ for } T \text{ large enough}$$

with

$$d(p,q) = \mathrm{KL}(\mathcal{B}(p), \mathcal{B}(q)) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$$

(Kullback-Leibler divergence between two Bernoulli distributions).

Our optimality criterion

- Our goal: build asymptotically optimal algorithms

Definition

A bandit algorithm is **asymptotically optimal** if, for every bandit model,

$$p_a < p^* \Rightarrow \limsup_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \leq \frac{1}{d(p_a, p^*)}$$

A multi-armed bandit problem

What are optimal bandit algorithms?

The UCB approach

Bayesian bandit algorithms

Some naive strategies

- Draw each arm T/K times

$\Rightarrow$ EXPLORATION only

Some naive strategies

- Draw each arm T/K times

$\Rightarrow$ EXPLORATION only

- Always play the empirical best arm

$$A_{t+1} = \underset{a}{\text{argmax}}\ \hat{p}_a(t)$$

$\Rightarrow$ EXPLOITATION only

Some naive strategies

- Draw each arm T/K times

$\Rightarrow$ EXPLORATION only

- Always play the empirical best arm

$$A_{t+1} = \underset{a}{\operatorname{argmax}}\ \hat{p}_a(t)$$

$\Rightarrow$ EXPLOITATION only

- Draw uniformly the arms during $T/2$ time steps (EXPLORATION)
  Then choose the empirical best and draw it till the end (EXPLOITATION)

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

The UCB heuristic

- For each arm a, compute a confidence interval on the unknown parameter p_a:

$$p_a \leq UCB_a(t) \quad w.h.p$$

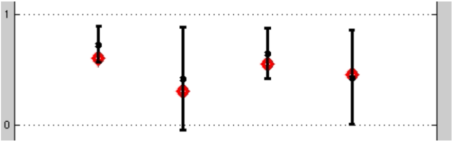


Figure: Confidence intervals on the arms means after t rounds

The UCB heuristic

- Use the *optimism-in-face-of-uncertainty principle*:

  'act as if the best possible model was the true model'

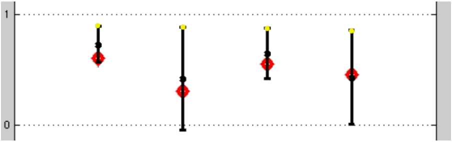
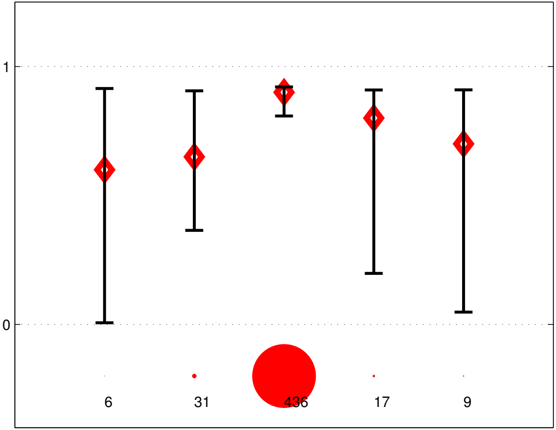


Figure: Confidence intervals on the arms means after t rounds

- The algorithm chooses at time $t+1$

$$A_{t+1} = \arg\max_a \; UCB_a(t)$$

A UCB algorithm in action



The UCB1 algorithm

- UCB1 [Auer et al. 02] uses Hoeffding bounds:

$$UCB_a(t) = \hat{p}_a(t) + \sqrt{\frac{2\log(t)}{N_a(t)}}$$

- One has:

$$\mathbb{E}[N_a(T)] \leq \underbrace{\frac{K_1}{2(p_a - p^*)^2}}_{\text{bigger than our target constant}} \times \log T + K_2,$$

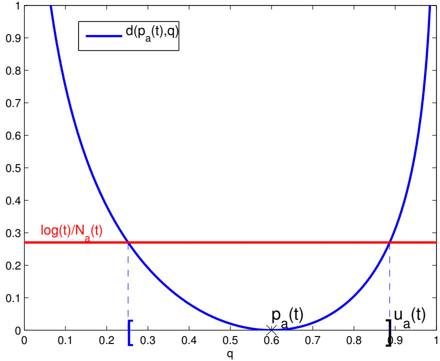
with $K_1 > 1$.

The KL-UCB algorithm

- KL-UCB [Cappé, Garivier, Maillard, Stoltz, Munos 11] uses the index:

$$u_a(t) = \underset{q > \hat{p}_a(t)}{\text{argmax}} \left\{ d\left(\hat{p}_a(t), q\right) \leq \frac{\log(t)}{N_a(t)} \right\}$$

with $d(p,q) = \text{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right)$.



The KL-UCB algorithm

- KL-UCB [Cappé, Garivier, Maillard, Stoltz, Munos 11] uses the index:

$$u_a(t) = \underset{q > \hat{p}_a(t)}{\text{argmax}} \left\{ d\left(\hat{p}_a(t), q\right) \leq \frac{\log(t)}{N_a(t)} \right\}$$

with $d(p, q) = \text{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right)$.

- One has

$$\mathbb{E}[N_a(T)] \leq \underbrace{\frac{1}{d(p_a, p^*)}}_{\text{our target constant}} \times \log T + K$$

Outline

A multi-armed bandit problem

What are optimal bandit algorithms?

The UCB approach

Bayesian bandit algorithms

Statistical Background

$X_1, \ldots X_n$ be n i.i.d observations of a Bernoulli distribution $\mathcal{B}(p)$

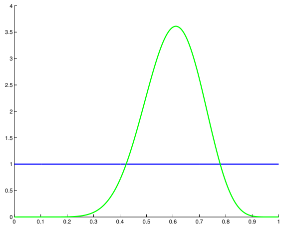
- **Frequentist point of view:** p is an unknown parameter

$$\hat{p}_n = \frac{X_1 + \cdots + X_n}{n} \quad \text{estimates} \quad p$$

- **Bayesian point of view:** p is drawn from a probability distribution (prior distribution): $p \sim \mathcal{U}([0,1])$.

  The posterior distribution incorporates the information we have on p:

$$\pi_a(t) = \mathcal{L}(p | X_1, ..., X_n)$$



Bayesian algorithms

At the end of round t,

- $\Pi_t = (\pi_1(t), \ldots, \pi_K(t))$ is the current posterior over $(p_1, ..., p_K)$
- $\pi_a(t) = \text{Beta}(S_a(t) + 1, N_a(t) - S_a(t) + 1)$ for Bernoulli bandits

A Bayesian algorithm uses Π_t to choose action A_{t+1}.

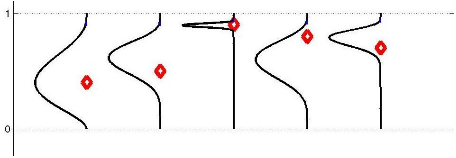
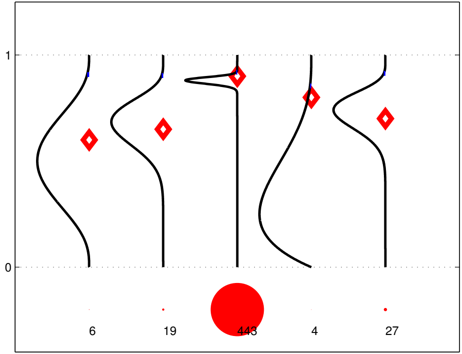


Figure: Posterior over the means of the arms after t rounds

Two optimal Bayesian algorithms

- **Bayes-UCB** chooses at time $A_{t+1} = \underset{a}{\text{argmax}}\ q_a(t)$, where

$$q_a(t) = Q\left(1 - \frac{1}{t}, \pi_a(t)\right).$$



Two optimal Bayesian algorithms

- **Thompson Sampling algorithm** uses samples from the posterior distribution:

$$\forall a \in \{1..K\}, \quad p_a(t) \sim \pi_a(t)$$
$$A_{t+1} = \text{argmax}_a \; p_a(t)$$

Our contributions

Both Bayes-UCB and Thompson Sampling are asymptotically optimal algorithms

Conclusion

You have understood:

- How bandits can save lives (initial motivation)
- How bandits can make money (current motivation)

You know how to design good bandit algorithms:

- By using the UCB approach (and good confidence intervals)
- By using Bayesian tools

Any question?