Dynamic Ressource Allocation

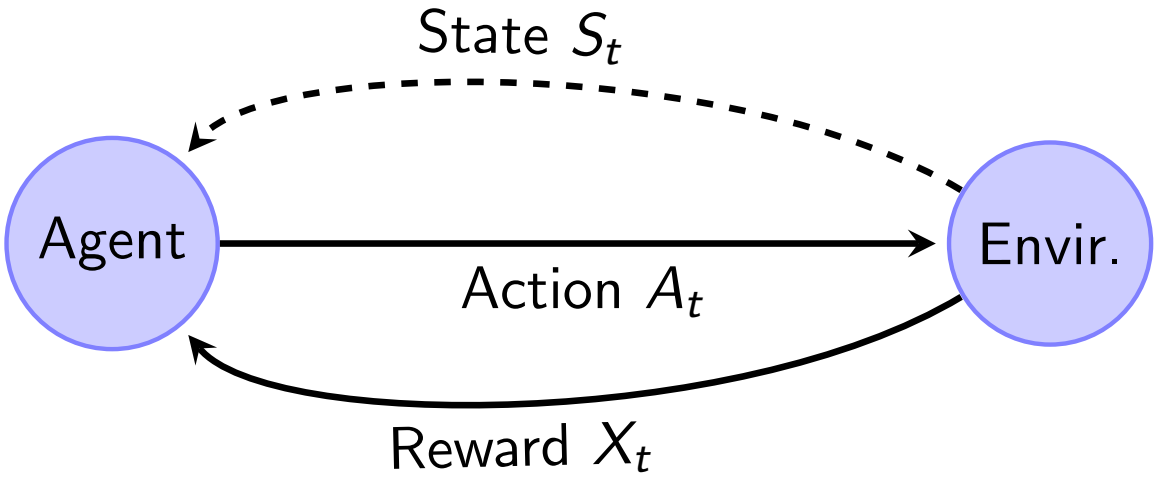
Olivier Cappé, Aurélien Garivier, Sébastien Gerchinovitz, Emilie Kaufmann

Institut de Mathématiques de Toulouse

February 20th, 2014

Roadmap

Reinforcement Learning



Exploration
|
Exploitation

$\neq$ statistical learning (maximizing the reward is the main goal)
$\neq$ game theory (the environment is stochastic)

Dynamic Ressource Allocation: basic model

Environment K options with parameters $\theta = (\theta_1, \dots, \theta_K)$ such that for any possible choice of option $A_t \in \{1, \dots, K\}$ at time t, one receives the reward

$$X_t = X_{A_t,t}$$

where, for any $1 \leq a \leq K$ and $s \geq 1$, $X_{a,s} \sim P_{\theta_a}$, and the $(X_{a,s})_{a,s}$ are independent.

Reward disbtributions can form a parametric family (ex: canonical exponential family) or not (ex: general bounded rewards)

Example Bernoulli rewards: $\theta \in [0,1]^K$, $X_{a,s} \sim \mathcal{B}(\theta_a)$

Strategy The agent's actions follow a dynamical strategy $\pi = (\pi_1, \pi_2, \dots)$ such that

$$A_t = \pi_t(X_1, \dots, X_{t-1})$$

Performance Evaluation, Regret

Cumulated Reward $S_n = \sum_{t=1}^{n} X_t$

Our goal Choose π so as to maximize

$$E[S_n] = \sum_{t=1}^{n} \sum_{a=1}^{K} \mathbb{E}\Big[\mathbb{E}[X_t \mathbb{1}\{A_t = a\} | X_1, \ldots, X_{t-1}]\Big]$$

$$= \sum_{a=1}^{K} \mu(\theta_a) \mathbb{E}[N_n^{\pi}(a)]$$

where $N_t^{\pi}(a) = \sum_{s \leq t} \mathbb{1}\{A_s = a\}$ is the number of draws of options a up to time n, and $\mu(\theta_a) = \mathbb{E}[X_{a,t}]$

Regret Minimization equivalent to minimizing

$$R_n = n\mu^* - E[S_n] = \sum_{a:\mu(\theta_a)<\mu^*} \big(\mu^* - \mu(\theta_a)\big) \mathbb{E}[N_n^{\pi}(a)]$$

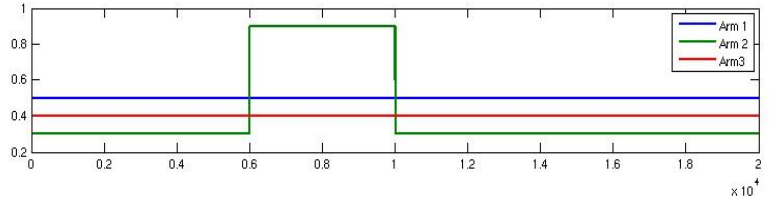
where $\mu^* \in \max\{\mu(\theta_a) : 1 \leq a \leq K\}$

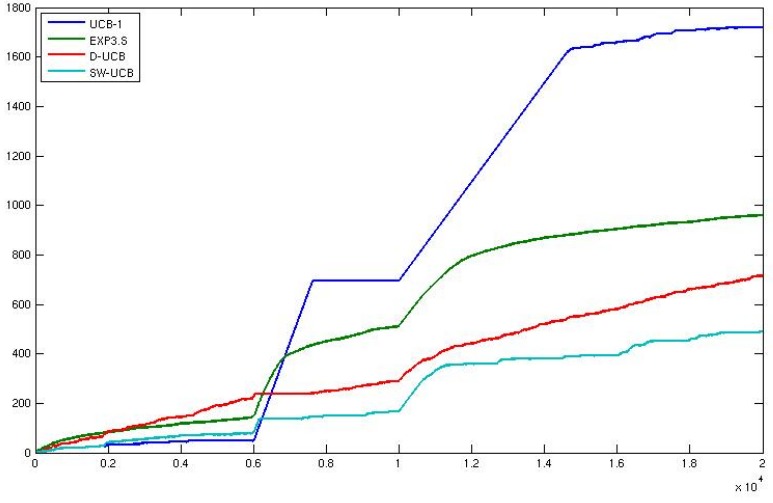
Non-stationary bandits

Model: abrupt changes in reward distributions

Goal: tracking the best arm

Example: channel sensing, Scanning Tunnelling Microscope





(Generalized) Linear Bandits

Contextual information

$$\mathbb{E}[X_t|A_t] = \mu(m'_{A_t}\theta_*)$$

where $\theta_* \in \mathbb{R}^d$ is an unknown parameter and $\mu : \mathbb{R} \to \mathbb{R}$ a link function

Example for binary rewards

$$\mu(x) = \frac{\exp(x)}{1 + \exp(x)}$$

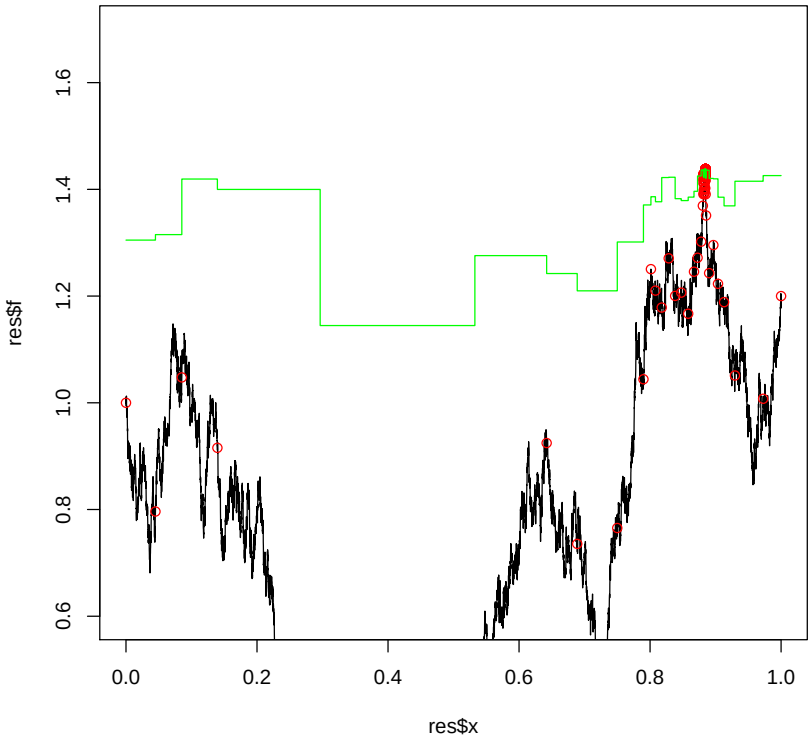
Application targeted advertisement

Goal regret bound depending on dimension d and not on the number of actions

Stochastic Optimization

Goal find the maximum (or another statistic) of a function $f : C \subset \mathbb{R}^d \to \mathbb{R}$ observed possibly with noise

Example maximal electro-magnetic field exposure (SAR)



Model f is sampled from a Gaussian Process (or low norm in RKHS)

Active learning strategies inspired by bandit algorithms

Markov Decision Processes

The system has a *state* S_t with a Markov dynamics:

$$S_{t+1} \sim P(\cdot; S_t, A_t) \text{ et } R_t = r(S_t, A_t) + \varepsilon_t$$

- Many applications
- Several possible uses of bandits inside the MDP
- Bandit-inspired algorithms

Exploration with probabilistic experts

Search space $B \subset \Omega$ (discrete set)

Probabilistic experts $P_a \in \mathfrak{M}_1(\Omega)$ for $a \in \{1, \dots, K\}$

Requests at time t, calling expert A_t yields a random $X_t \sim P_{A_t}$

Goal find as many elements of B with as few requests as possible:

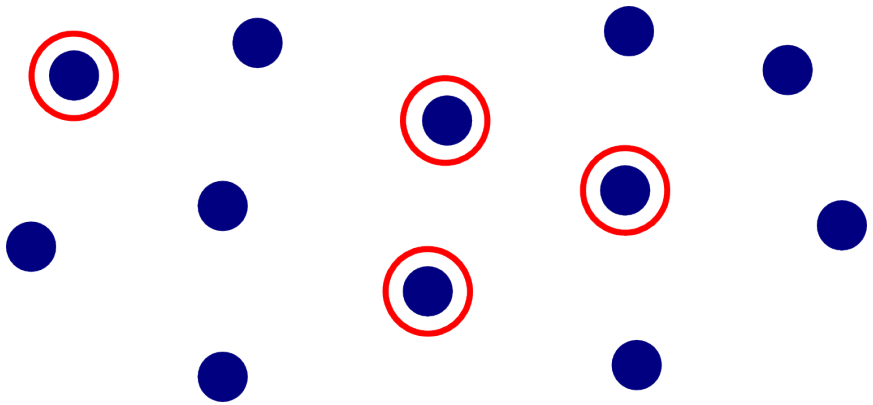
$$F_n = \mathrm{Card}\left(B \cap \{X_1, \dots, X_n\}\right)$$

$\neq$ bandits finding an item twice is useless!

Initial motivation power system security

Multi-action bandits

- **Sequential task**: choose a set of actions $\mathcal{S}_t \subset \{1, \dots, K\}$ with some constraints, and get the sum of rewards $\sum_{a \in \mathcal{S}_t} X_{a,t}$.



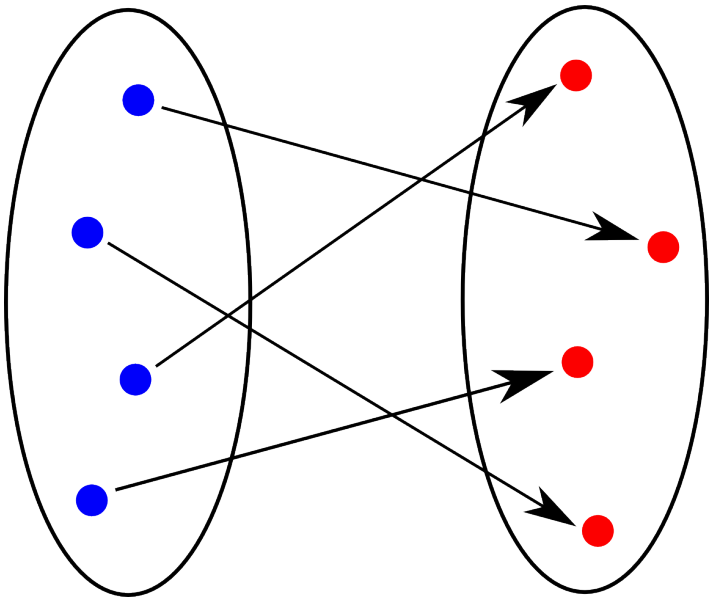
- **Goal**: minimize the regret

$$\max_{\mathcal{S} \subset \{1,\dots,K\}} \sum_{t=1}^{n} \sum_{a \in \mathcal{S}} X_{a,t} - \sum_{t=1}^{n} \sum_{a \in \mathcal{S}_t} X_{a,t} \ .$$

- **Keypoint**: the number of action sets $\mathcal{S} \subset \{1, \dots, K\}$ is large $\rightsquigarrow$ *combinatorial* bandits (statistical and algorithmic issues).

Another combinatorial bandit problem: perfect matchings

- **Sequential task**: choose a one-to-one mapping $\sigma_t : \mathcal{A} \to \mathcal{B}$, and get the sum of rewards $\sum_{a \in \mathcal{A}} X_{(a,\sigma_t(a)),t}$.



- **Goal**: minimize the regret

$$\max_{\sigma:\mathcal{A}\to\mathcal{B}} \sum_{t=1}^{n} \sum_{a\in\mathcal{A}} X_{(a,\sigma(a)),t} - \sum_{t=1}^{n} \sum_{a\in\mathcal{A}} X_{(a,\sigma_t(a)),t} \ .$$

Roadmap

The Lower Bound of Lai & Robbins

Theorem [Lai&Robbins, '85]

If π is a consistent strategy then, for any $\theta \in [0,1]^K$,

$$\liminf_{n\to\infty} \frac{R_n(\theta)}{\log(n)} \geq \sum_{a:\theta_a<\theta^*} \frac{\theta^* - \theta_a}{\mathrm{kl}(\theta_a, \theta^*)}$$

where

$$\mathrm{kl}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

denotes the Kullback-Leibler divergence between the Bernoulli distributions $\mathcal{B}(p)$ and $\mathcal{B}(q)$, for $0 \leq p, q \leq 1$

Optimism in the Face of Uncertainty

Optimism is a heuristic principle popularized by [Lai&Robins '85; Agrawal '95] which consists in letting the agent play as if the environment was the most favorable among all environments that are still sufficiently likely given the observations accumulated so far.

Surprisingly, this simple heuristic principle can be instantiated into algorithms that are robust, efficient and easy to implement in many scenarios pertaining to reinforcement learning.

Upper Confidence Bound Strategies

UCB [Lai&Robins '85; Auer&al '02]

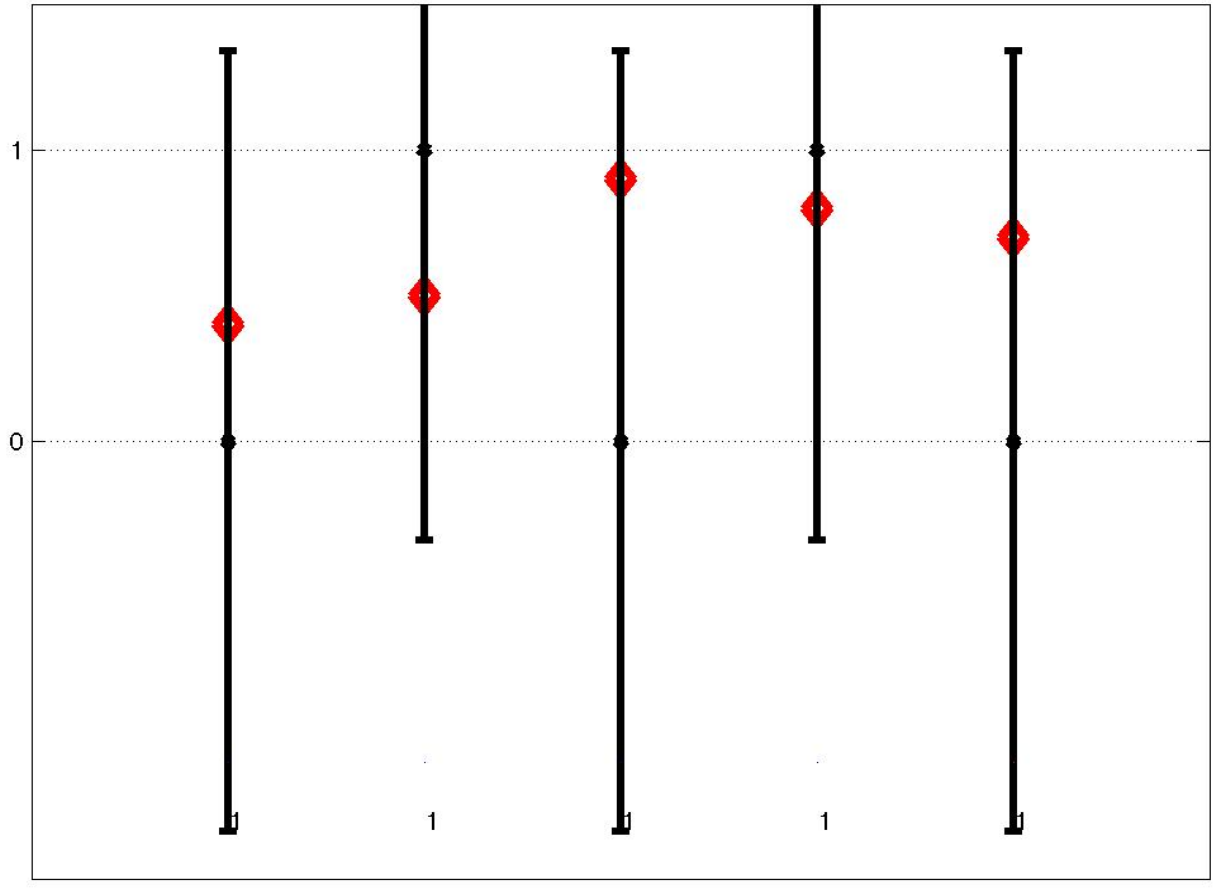
- Construct an upper confidence bound for the expected reward of each option:

$$\underbrace{\frac{S_t(a)}{N_t(a)}}_{\text{estimated reward}} + \underbrace{\sqrt{\frac{\log(t)}{2N_t(a)}}}_{\text{exploration bonus}}$$

- Choose the option with the highest UCB

- It is an *index strategy* [Gittins '79]
- Its behavior is easily interpretable and intuitively appealing

UCB in Action



Performance of UCB

The regret of UCB can be shown to verify

$$E[R_n] = O(\log(n))$$

(finite-time regret bound) and

$$\limsup_{n\to\infty} \frac{\mathbb{E}[R_n]}{\log(n)} \leq \sum_{a:\theta_a<\theta^*} \frac{1}{2(\theta^* - \theta_a)}$$

where the rhs. is greater than suggested by the bound by Lai & Robbins

Best known variant: KL-UCB reaches the (asymptotic) lower bound

Softmax methods

- **UCB algos**: pick an option $A_t \in \arg\max_a \mathrm{crit}_t(a)$
- **Softmax algos**: to be more robust, assign probabilities $p_{a,t}$ to all options $a \in \{1, \dots, K\}$ ($p_{a,t}$ increases with $\mathrm{crit}_t(a)$), and then pick $A_t \sim p_t$ at random.

Exp3 (Auer et. al 2002)

At each round $t \geq 1$,

- Using past data, compute the weight vector p_t as

$$p_{a,t} = \frac{\exp\left(\eta_t \sum_{s=1}^{t-1} \tilde{X}_{a,s}\right)}{\sum_{b=1}^{K} \exp\left(\eta_t \sum_{s=1}^{t-1} \tilde{X}_{b,s}\right)}, \quad 1 \leq a \leq K ;$$

  where $\tilde{X}_{a,s} = \frac{X_{a,s}}{p_{a,s}} \mathbb{1}_{\{A_s = a\}}$ is an unbiased estimator of $X_{a,s}$.
- Pick $A_t \sim p_t$ at random.

Theoretical guarantees without stochastic assumptions

- **Robust guarantees**: the algorithm Exp3 suffers small regret even for reward sequences $(X_{a,t})$ that don't satisfy the earlier stochastic assumptions. Namely, whatever the rewards:

$$\max_{1\leq a\leq K} E\left[\sum_{t=1}^{n} X_{a,t}\right] - E\left[\sum_{t=1}^{n} X_{A_t,t}\right] \leq C\sqrt{n\,K\ln K}$$

- The distribution-free $\sqrt{n}$ rate is optimal in a worst-case sense.
- High-probability bounds can be derived with variants of Exp3.

- **Similarity between softmax algorithms**: they all are regularization methods, close to projected gradient descent.
  $\rightsquigarrow$ Can be implemented efficiently for combinatorial bandits.

Bayesian algorithms

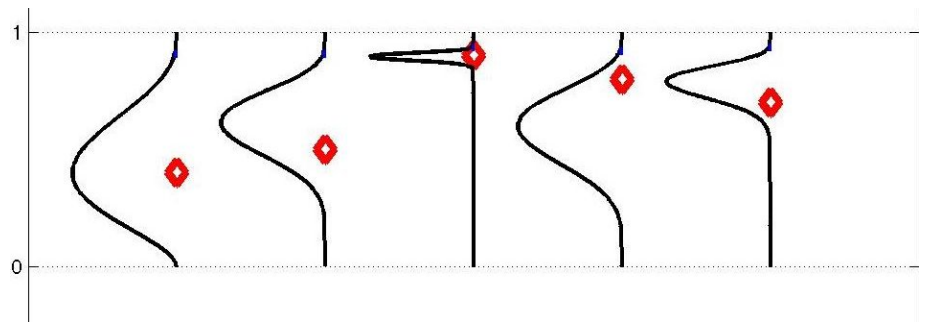
Assume the parameter of each arm is drawn from a prior distribution:

$$\forall a \in \{1, \ldots, K\}, \quad \theta_a \sim \pi_a^0$$

At the end of round t,

- $\Pi_t = (\pi_1^t, \ldots, \pi_K^t)$ is the current posterior over $(\theta_1, ..., \theta_K)$
- $\Lambda_t = (\lambda_1^t, \ldots, \lambda_K^t)$ is the current posterior over the means $(\mu(\theta_1), ..., \mu(\theta_K))$

A Bayesian algorithm uses Π_{t-1} to choose action A_t.



The Bernoulli case $\theta = \mu$ and $\Pi_t = \Lambda_t$

$$\pi_a^0 = \mathcal{U}([0,1]) \quad \text{and} \quad \pi_a^t = \text{Beta}(S_a(t) + 1, N_a(t) - S_a(t) + 1)$$

Bayes-UCB: a Bayesian optimistic algorithm

- $\Lambda_t = (\lambda_1^t, \ldots, \lambda_K^t)$ is the current posterior over the means $(\mu_1, ..., \mu_K)$

Bayes-UCB [Kaufmann et al'12]

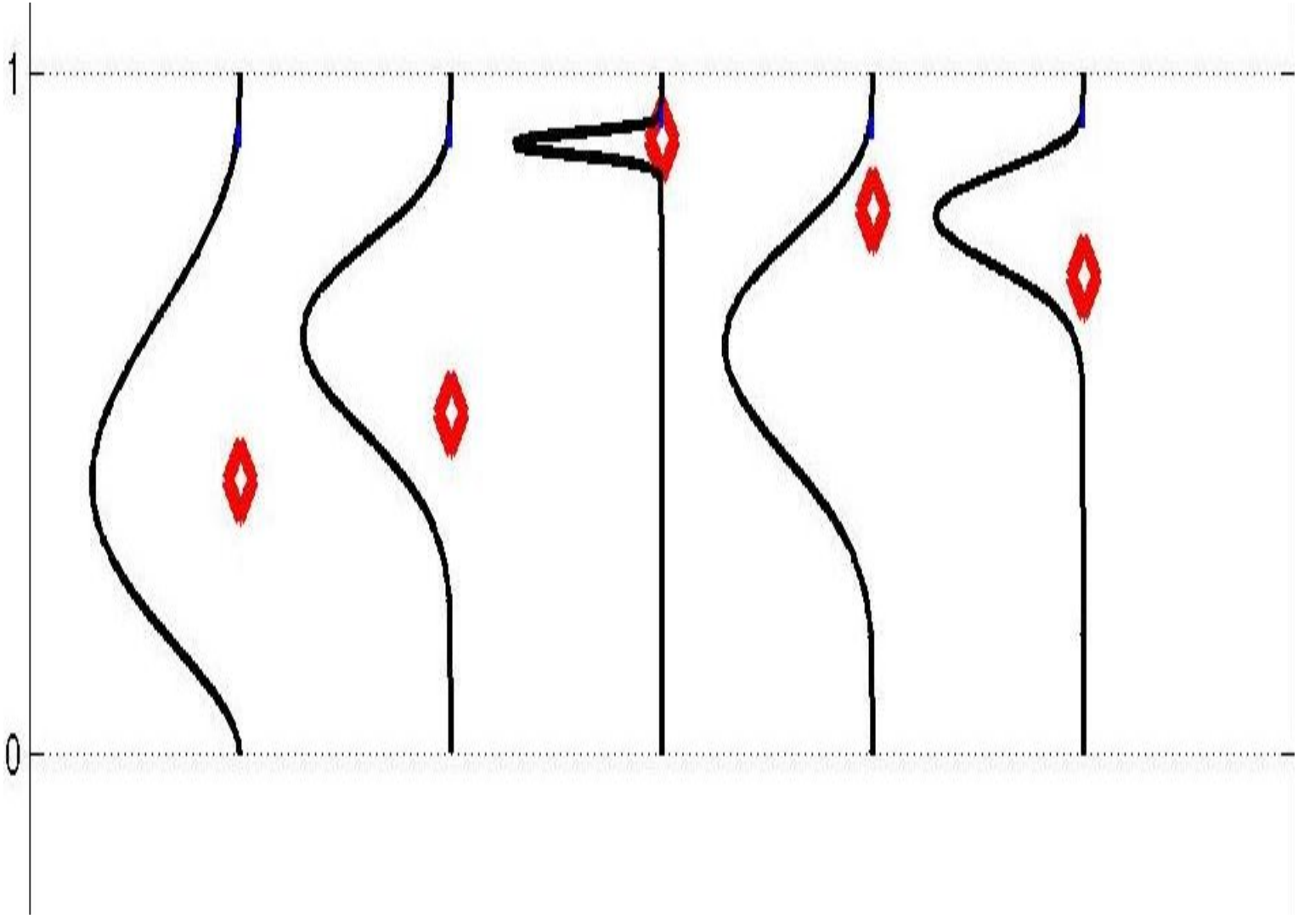
- Construct a Bayesian upper confidence bound for the expected reward of each option:

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

  where $Q(\alpha, \pi)$ is the quantile of order α of distribution π
- Choose arm $A_t = \underset{a=1...K}{\text{argmax}}\ q_a(t)$

- Bayes-UCB matches the lower bound of Lai and Robbins for Bernoulli distributions

Bayes-UCB in Action



Thompson Sampling: a Bayesian randomized algorithm

- $\Pi_t = (\pi_1^t, \ldots, \pi_K^t)$ is the current posterior over the parameters $(\theta_1, ..., \theta_K)$

Thompson Sampling [Thompson'33]

- Sample a bandit problem from the current posterior:

$$\forall a \in \{1..K\}, \quad \theta_a(t) \sim \pi_a^{t-1}$$

- Act optimaly in this sampled model:

$$A_t = \text{argmax}_a \ \mu(\theta_a(t))$$

- TS matches the lower bound of Lai and Robbins for Bernoulli distributions
- TS is very easy to implement
- TS is performant in more complex models