



---

# **Models for Trading**

## **Exploration and Exploitation**

### **using Upper Confidence Bounds**

Peter Auer

CiT, University of Leoben



- Theme of this talk:  
The method of **Upper Confidence Bounds** for dealing with the exploration/exploitation problem
- A variant of the bandit problem with linear side information
- Special case: The Random Bandit Problem
- Online reinforcement learning

# The bandit problem with linear side information



In each trial  $t = 1, \dots, T$ :

1. receive a feature vector  $\mathbf{z}(t) \in \mathbf{R}^d$ ,  $\|\mathbf{z}(t)\|_2 \leq 1$ ,
2. select an alternative  $i(t) \in \{1, \dots, K\}$ ,
3. observe success  $x_{i(t)}(t) \in \{0, 1\}$ .

Statistical assumption:

for each alternative  $i$  the probability of success is given by an unknown linear law  $\mathbf{f}_i \in \mathbf{R}^d$ ,  $\|\mathbf{f}_i\|_2 \leq 1$ ,

$$\mathbf{P} \{x_i(t) = 1\} = \mathbf{z}(t) \cdot \mathbf{f}_i.$$

# Goal



Maximize the (expected) number of successes

$$\mathbf{E} \left[ \sum_{t=1}^T x_{i(t)}(t) \right] = \sum_{t=1}^T \mathbf{z}(t) \cdot \mathbf{f}_{i(t)}.$$

Compare with the number of successes of an optimal, omniscient strategy:

$$\text{REGRET} = \sum_{t=1}^T \max_i \{ \mathbf{z}(t) \cdot \mathbf{f}_i \} - \sum_{t=1}^T \mathbf{z}(t) \cdot \mathbf{f}_{i(t)}$$

⇒ Need to learn about the unknown  $\mathbf{f}_i$ .



## Online learning of linear functions

### Random Bandit Problem

- [Robbins, 1952], [Lai, Robbins, 1985]
- The success probability for each alternative is constant over time:

$$\mathbf{P} \{x_i(t) = 1\} = \mu_i$$

- Special case of our model where  $\mathbf{z}(t) = 1$  for all  $t$ .
- [Agrawal, 1995] uses upper confidence bounds

### Associative reinforcement learning

- [Kaelbling, 1994] uses upper confidence bounds as heuristics
- [Abe, Long, 1999] give first rigorous analysis



**Our result (using confidence bounds)**

$$\mathbb{E} [\text{REGRET}] \leq 18 \ln(T) K^{1/2} (dT)^{1/2} \quad \text{as } T \rightarrow \infty$$

**[Abe, Long, 1999] (using online learning of linear functions)**

$$\mathbb{E} [\text{REGRET}] \leq (2 + o(1)) K^{1/2} T^{3/4} \quad \text{as } T \rightarrow \infty$$



- We loose only  $\tilde{O}\left(1/\sqrt{T}\right)$  per trial against an omniscient strategy:

$$\frac{1}{T} \sum_{t=1}^T \left[ \max_i \{\mathbf{z}(t) \cdot \mathbf{f}_i\} - \sum_{t=1}^T \mathbf{z}(t) \cdot \mathbf{f}_{i(t)} \right] = \tilde{O}\left(\frac{1}{\sqrt{T}}\right)$$

- Holds even for individual inputs  $\mathbf{z}_i(t)$  for each alternative  $i$ :

$$\mathbf{P} \{x_i(t) = 1\} = \mathbf{z}_i(t) \cdot \mathbf{f}_i$$

- Holds for non-binary  $x_i(t) \in [0, 1]$  when

$$\mathbf{E} [x_i(t)] = \mathbf{z}_i(t) \cdot \mathbf{f}_i.$$

# Algorithm: Using upper confidence bounds



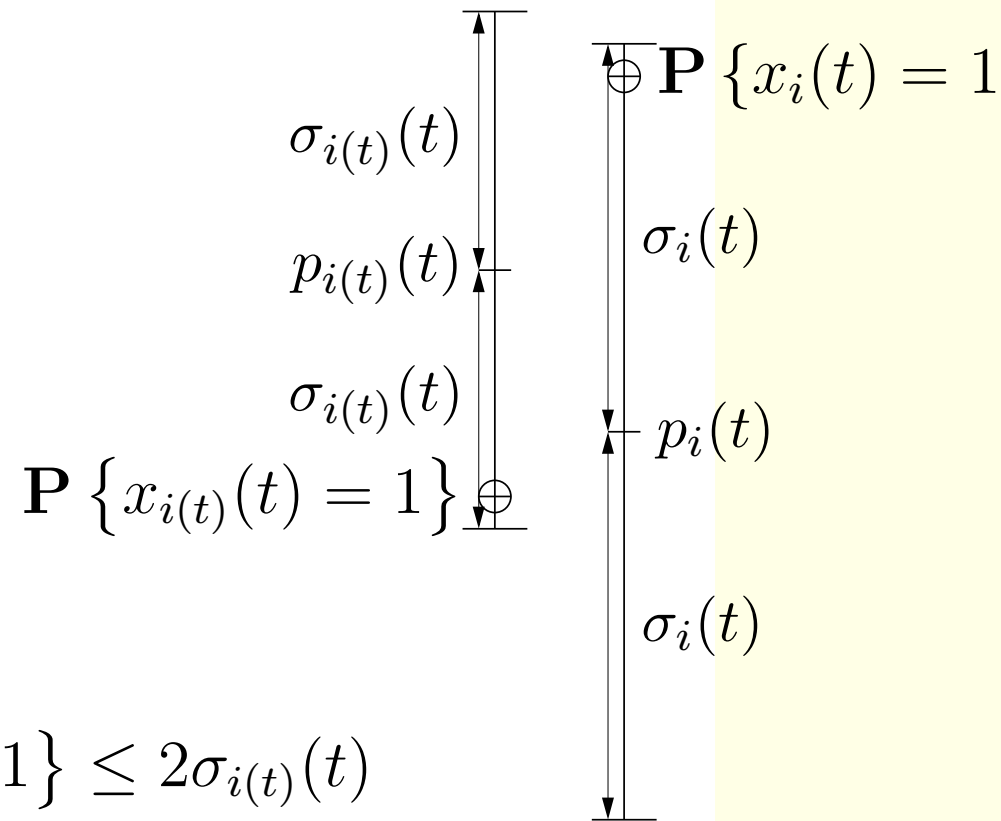
1. Calculate an estimate  $p_i(t)$  for the success probability of each alternative.
2. Calculate a confidence interval such that

$$\mathbf{P} \{x_i(t) = 1\} \in [p_i(t) - \sigma_i(t), p_i(t) + \sigma_i(t)]$$

with very high probability.

3. Choose alternative  $i(t)$  with maximal  $p_i(t) + \sigma_i(t)$ .

# Why does this work?



We have

$$\max_i \mathbf{P} \{x_i(t) = 1\} - \mathbf{P} \{x_{i(t)}(t) = 1\} \leq 2\sigma_{i(t)}(t)$$

and thus

$$\text{REGRET} \leq 2 \sum_{t=1}^T \sigma_{i(t)}(t).$$

# Bounding the widths of the confidence intervals



Consider the random bandit problem:

$$\mathbf{P} \{x_i(t) = 1\} = \mu_i$$

We need a confidence interval with

$$\mathbf{P} \{x_i(t) = 1\} = \mu_i \in [p_i(t) - \sigma_i(t), p_i(t) + \sigma_i(t)].$$

Good estimate for  $\mu_i$ :

$$p_i(t) = \frac{1}{n_i(t)} \sum_{\tau < t: i(\tau) = i} x_i(\tau), \quad n_i(t) = \#\{1 \leq \tau < t : i(\tau) = i\}$$

$$\implies \mathbf{E} [p_i(t)] \cong \frac{1}{n_i(t)} \sum_{\tau < t: i(\tau) = i} \mathbf{E} [x_i(\tau)] = \mu_i$$

# Random bandit problem (cont.)



Estimate:

$$p_i(t) = \frac{1}{n_i(t)} \sum_{\tau < t: i(\tau)=i} x_i(\tau)$$

Width: (follows from Chernoff/Hoeffding bounds)

$$\begin{aligned} \sigma_i(t) &\sim \sqrt{(\log T) \cdot \mathbf{V} [p_i(t)]} \cong \sqrt{\frac{\log T}{[n_i(t)]^2} \sum_{\tau < t: i(\tau)=i} \mathbf{V} [x_i(\tau)]} \\ &\leq \sqrt{\frac{\log T}{[n_i(t)]^2} \sum_{\tau < t: i(\tau)=i} 1} \\ &= \sqrt{\frac{\log T}{n_i(t)}} \end{aligned}$$

# Random bandit problem (cont.)

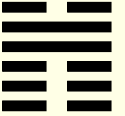


We have

$$\sigma_i(t) \sim \sqrt{\frac{\log T}{n_i(t)}}.$$

$$\begin{aligned} \text{REGRET} &\leq 2 \sum_{t=1}^T \sigma_{i(t)}(t) \sim \sqrt{\log T} \sum_{i=1}^K \sum_{\tau=1}^{n_i(T)} \frac{1}{\sqrt{\tau}} \\ &\sim \sqrt{\log T} \sum_{i=1}^K \sqrt{n_i(T)} \\ &\leq \sqrt{\log T} \cdot K \cdot \sqrt{T/K} \\ &= \sqrt{KT \log T} \end{aligned}$$

# Random bandit problem: Improved bounds



Let  $\mu^* = \max_i \mu_i$ ,  $\Delta_i = \mu^* - \mu_i$ . Then

$$\text{REGRET} \leq \sum_{i=1}^K \Delta_i \mathbf{E} [n_i(T)].$$

# Random bandit problem: Improved bounds



Let  $\mu^* = \max_i \mu_i$ ,  $\Delta_i = \mu^* - \mu_i$ . Then

$$\text{REGRET} \leq \sum_{i=1}^K \Delta_i \mathbf{E} [n_i(T)].$$

Furthermore,  $i(t) = i$  only if

$$p_i(t) + \sigma_i(t) \geq p^*(t) + \sigma^*(t)$$

# Random bandit problem: Improved bounds



Let  $\mu^* = \max_i \mu_i$ ,  $\Delta_i = \mu^* - \mu_i$ . Then

$$\text{REGRET} \leq \sum_{i=1}^K \Delta_i \mathbf{E} [n_i(T)].$$

Furthermore,  $i(t) = i$  only if

$$\mu_i + 2\sigma_i(t) \geq p_i(t) + \sigma_i(t) \geq p^*(t) + \sigma^*(t)$$

# Random bandit problem: Improved bounds



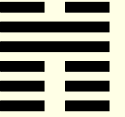
Let  $\mu^* = \max_i \mu_i$ ,  $\Delta_i = \mu^* - \mu_i$ . Then

$$\text{REGRET} \leq \sum_{i=1}^K \Delta_i \mathbf{E} [n_i(T)].$$

Furthermore,  $i(t) = i$  only if

$$\mu_i + 2\sigma_i(t) \geq p_i(t) + \sigma_i(t) \geq p^*(t) + \sigma^*(t) \geq \mu^*.$$

# Random bandit problem: Improved bounds



Let  $\mu^* = \max_i \mu_i$ ,  $\Delta_i = \mu^* - \mu_i$ . Then

$$\text{REGRET} \leq \sum_{i=1}^K \Delta_i \mathbf{E} [n_i(T)].$$

Furthermore,  $i(t) = i$  only if

$$\mu_i + 2\sigma_i(t) \geq p_i(t) + \sigma_i(t) \geq p^*(t) + \sigma^*(t) \geq \mu^*.$$

Thus  $i(t) = i$  only if  $\sigma_i(t) \geq \Delta_i/2$ .

# Random bandit problem: Improved bounds



Let  $\mu^* = \max_i \mu_i$ ,  $\Delta_i = \mu^* - \mu_i$ . Then

$$\text{REGRET} \leq \sum_{i=1}^K \Delta_i \mathbf{E} [n_i(T)].$$

Furthermore,  $i(t) = i$  only if

$$\mu_i + 2\sigma_i(t) \geq p_i(t) + \sigma_i(t) \geq p^*(t) + \sigma^*(t) \geq \mu^*.$$

Thus  $i(t) = i$  only if  $\sigma_i(t) \geq \Delta_i/2$ .

Since  $\sigma_i(t) \sim \sqrt{\frac{\log T}{n_i(t)}}$  we have  $n_i(t) \leq O\left(\frac{\log T}{\Delta_i^2}\right)$ ,

# Random bandit problem: Improved bounds



Let  $\mu^* = \max_i \mu_i$ ,  $\Delta_i = \mu^* - \mu_i$ . Then

$$\text{REGRET} \leq \sum_{i=1}^K \Delta_i \mathbf{E} [n_i(T)].$$

Furthermore,  $i(t) = i$  only if

$$\mu_i + 2\sigma_i(t) \geq p_i(t) + \sigma_i(t) \geq p^*(t) + \sigma^*(t) \geq \mu^*.$$

Thus  $i(t) = i$  only if  $\sigma_i(t) \geq \Delta_i/2$ .

Since  $\sigma_i(t) \sim \sqrt{\frac{\log T}{n_i(t)}}$  we have  $n_i(t) \leq O\left(\frac{\log T}{\Delta_i^2}\right)$ , and hence

$$\text{REGRET} \leq O\left((\log T) \sum_{i=1: i \neq i^*}^K \frac{1}{\Delta_i}\right).$$

**Linear side information:**  $\mathbf{E} [x_i(t)] = \mathbf{z}(t) \cdot \mathbf{f}_i$



Calculate  $p_i(t)$ :

Choose  $\mathbf{a}_i(t) = (a_i(1), \dots, a_i(n_i(t)))$  such that

$$\mathbf{z}(t) = \mathbf{a}_i(t) \cdot Z_i(t) \cong \sum_{\tau=1}^{n_i(t)} a_i(\tau) \cdot \mathbf{z}_i(\tau)$$

and set

$$p_i(t) := \mathbf{a}_i(t) \cdot \mathbf{x}_i(t) \cong \sum_{\tau=1}^{n_i(t)} a_i(\tau) \cdot x_i(\tau),$$

where

$$Z_i(t) = (\mathbf{z}(\tau))_{1 \leq \tau < t: i(\tau)=i}, \quad \mathbf{x}_i(t) = (x_i(\tau))_{1 \leq \tau < t: i(\tau)=i}.$$

# Calculating the variance



From

$$p_i(t) = \mathbf{a}_i(t) \cdot \mathbf{x}_i(t), \quad \mathbf{E} [x_i(t)] = \mathbf{z}(t) \cdot \mathbf{f}_i, \quad \mathbf{z}(t) = \mathbf{a}_i(t) \cdot Z_i(t)$$

we get

$$\begin{aligned} \underline{\mathbf{E} [p_i(t)]} &= \mathbf{E} [\mathbf{a}_i(t) \cdot \mathbf{x}_i(t)] \cong \mathbf{a}_i(t) \cdot \mathbf{E} [\mathbf{x}_i(t)] \\ &= \mathbf{a}_i(t) \cdot Z_i(t) \cdot \mathbf{f}_i = \mathbf{z}(t) \cdot \mathbf{f}_i = \underline{\mathbf{E} [x_i(t)]}, \\ \underline{\sigma_i(t)^2} &= (\log T) \cdot \mathbf{V} [p_i(t)] = (\log T) \cdot \mathbf{V} [\mathbf{a}_i(t) \cdot \mathbf{x}_i(t)] \\ &\cong (\log T) \cdot \sum_{\tau=1}^{n_i(t)} a_i(\tau)^2 \cdot \mathbf{V} [x_i(\tau)] \leq \underline{(\log T) \cdot \|\mathbf{a}_i(t)\|_2^2} \end{aligned}$$



Thus

$$\text{REGRET} \sim \sqrt{\log T} \sum_{t=1}^T \|\mathbf{a}_{i(t)}(t)\|_2.$$

We need  $\mathbf{a}_i(t)$  with  $\mathbf{a}_i(t) \cdot Z_i(t) = \mathbf{z}(t)$  and minimal  $\|\mathbf{a}_i(t)\|_2$ .

Simple calculus gives

$$\mathbf{a}_i(t) = \mathbf{z}(t) \cdot [Z_i(t)' \cdot Z_i(t)]^{-1} \cdot Z_i(t)'$$

It remains to bound

$$\sum_{t=1}^T \|\mathbf{a}_{i(t)}(t)\|_2 = \sum_{i=1}^K \sum_{t:i(t)=i} \sqrt{\mathbf{z}(t) \cdot [Z_i(t)' \cdot Z_i(t)]^{-1} \cdot \mathbf{z}(t)'}$$

# Calculating $\sum_{t=1}^T \|\mathbf{a}_{i(t)}(t)\|_2$ (very much simplified)



Assume that all  $\mathbf{z}_i(\tau)$  are unit vectors  $(0, \dots, 0, 1, 0, \dots, 0)$ .

Then

$$[Z_i(t)' \cdot Z_i(t)] = \begin{pmatrix} k_{i,1}(t) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & k_{i,d}(t) \end{pmatrix}$$

where  $k_{i,j}(t) = \#\{\tau < t : \mathbf{z}(\tau) = \mathbf{e}_j \text{ and } \tau < t : i(\tau) = i\}$ .

Thus, if  $i(t) = i$  and  $\mathbf{z}(t) = \mathbf{e}_j$ , then  $\|\mathbf{a}_{i(t)}(t)\|_2 = \frac{1}{\sqrt{k_{i,j}(t)}}$  and

$$k_{i,j}(t+1) = k_{i,j}(t) + 1.$$

Hence

$$\sum_{t=1}^T \|\mathbf{a}_{i(t)}(t)\|_2 \leq \sim Kd \sum_{\tau=1}^{T/Kd} \frac{1}{\sqrt{\tau}} \sim Kd \sqrt{T/Kd} = \sqrt{KdT}.$$



We consider a **Markov decision process** (MDP) on a finite set of *states*  $S$  with a finite state of *actions*  $A$  available in each state.

- ▶ There is an *initial state*  $s_0$ .
- ▶ The *transition probability*  $p_a(s, s')$  is the probability of reaching state  $s'$  when choosing action  $a$  in state  $s$ .
- ▶ The *reward* (in  $[0, 1]$ ) when choosing action  $a$  in state  $s$  has mean  $r_a(s)$ .

A **policy** is a mapping  $\pi : S \rightarrow A$ .

**Goal:** Maximize the rewards.

Need to learn optimal actions from delayed feedback.

# Motivation for online reinforcement learning



- Typical RL algorithms try to converge to a nearly optimal policy during a designated learning phase.
- Thus there is no exploration/exploitation problem, since only the performance of the final policy is considered:
  - > Q-Learning,
  - > Fiechter ('94),
  - >  $E^3$  [Kearns, Singh, 1998, 1999].
- We are interested in an online learning algorithm which is evaluated by its **regret**, i.e. the performance loss compared to an optimal policy.
- Possibly some reinforcement learning algorithms could be turned into online algorithms, but there is no analysis available.

# Discounted and undiscounted returns



A **path** is a sequence of states,  $\psi = s_0, s_1, s_2 \dots$

Let  $\mathbb{P}_\pi\{s_t = s\}$  be the probability that, when executing policy  $\pi$ , state  $s$  is reached after  $t$  steps.

RL often considers the (expected) discounted return

$$V_\pi^\gamma = \sum_{t \geq 0} \sum_{s \in S} \gamma^t \cdot \mathbb{P}_\pi\{s_t = s\} \cdot r_{\pi(s)}(s)$$

for some  $\gamma \in (0, 1)$ .

We consider the undiscounted return (up to time  $T$ )

$$V_\pi^T = \sum_{t=0}^T \sum_{s \in S} \mathbb{P}_\pi\{s_t = s\} \cdot r_{\pi(s)}(s).$$



Let

$$\mathbb{P}_A\{s_t = s, a_t = a\}$$

be the probability that, when executing learning algorithm  $A$ , state  $s$  is reached after  $t$  steps and action  $a$  is chosen in this state.

Then the (expected) return for algorithm  $A$  is

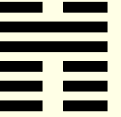
$$V_A^T = \sum_{t=0}^T \sum_{s \in S} \sum_{a \in A} \mathbb{P}_A\{s_t = s, a_t = a\} \cdot r_a(s)$$

and the regret of algorithm  $A$  up to time  $T$  is

$$\text{REGRET} = \max_{\pi} V_{\pi}^T - V_A^T.$$

# Episodic reinforcement learning

---



We consider episodes of length  $T$  such that the MDP is restarted after  $T$  steps.

We bound the performance of the learning algorithm in respect to the number of episodes  $M$ .

This simplifies analysis considerably.

We have recent results also for RL without restarts.

Then mixing times need to be considered: How fast can we reach an arbitrary state of the MDP.

# The algorithm: Upper confidence bounds again



- Our algorithm is based on upper confidence bounds for the return of a policy and achieves  $O(\log M)$  regret for episodic RL.
- In each episode  $m$  calculate upper confidence values  $\tilde{V}_\pi^T(m)$  for the returns of the policies  $\pi$ .
- In each episode execute the policy  $\pi_m$  which maximizes  $\tilde{V}_\pi^T(m)$ .
- The regret is bounded by

$$\text{REGRET} \leq \text{poly}(A, S, T, 1/\Delta) \cdot \log M$$

where  $\Delta = \min_{\pi: V_\pi^T \neq V_{\pi^*}^T} (V_{\pi^*}^T - V_\pi^T)$ .

- Previous work: [Burnetas, Katehakis, 1997]

# Why it works (1)



We have

$$\tilde{V}_{\pi_m} \geq \tilde{V}_{\pi^*} \geq V_{\pi^*}.$$

Hence, if  $\tilde{V}_{\pi_m} - V_{\pi_m} < \Delta$ , then

$$V_{\pi^*} - V_{\pi_m} \leq \tilde{V}_{\pi_m} - V_{\pi_m} < \Delta,$$

i.e.  $\pi_m$  is already an optimal policy with  $V_{\pi_m} = V_{\pi^*}$ .

Thus if  $B$  is a bound on the number of episodes  $m$  with  $\tilde{V}_{\pi_m} - V_{\pi_m} \geq \Delta$ , we get

$$\text{REGRET} \leq BT$$

(since the regret in each episode is at most  $T$ ).

## Why it works (2)



Need to count episodes with  $\tilde{V}_{\pi_m} - V_{\pi_m} \geq \Delta$ .

### Idea:

If policy  $\pi_m$  is chosen, then some new statistics about  $\pi_m$  are collected and  $\tilde{V}_{\pi_m} - V_{\pi_m}$  is reduced.

Thus there is only a limited number of episodes with  $\tilde{V}_{\pi_m} - V_{\pi_m} \geq \Delta$ .

## Why it works (3)



Fix  $\pi_m$ .

Let  $n(s)$  be the expected number of visits to state  $s$  when executing policy  $\pi_m$  for  $T$  steps. Then

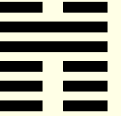
$$V = \sum_s n(s)r(s),$$

we set

$$\tilde{V} = \sum_s \tilde{n}(s)\tilde{r}(s),$$

and get

$$\tilde{V} - V \leq \sum_s [\tilde{n}(s) - n(s)] + \sum_s n(s)[\tilde{r}(s) - r(s)].$$



- Upper confidence bounds are a general technique to deal with exploration/exploitation trade-offs.
- The decrease of the confidence interval governs the regret bound.
- Model with linear side information:  
Success probability depends on  $\mathbf{z}(t) \cdot \mathbf{f}_i$ .
- Online reinforcement learning:  
Our algorithm concentrates exploration on promising states and policies while avoiding to explore mediocre states and policies.
- High-probability bounds for the adversarial bandit problem.

