

# Use of variance estimation in the multi-armed bandit problem

Jean-Yves Audibert<sup>1</sup>   Rémi Munos<sup>2</sup>   Csaba Szepesvári<sup>3</sup>

<sup>1</sup>Certis

ParisTech - Ecole des Ponts, France

<sup>2</sup>SequeL

INRIA Futurs, France

<sup>3</sup>University of Alberta, Canada  
and MTA SZTAKI, Hungary

Neural Information Processing Systems, 2006

# Outline

- 1 The multi-armed bandit problem and UCB policies
- 2  $\beta$ -UCB policy
- 3 UCB-tuned policy

# The multi-armed bandit problem

- We have a finite number of arms (or reward providers).
- At each time step, we have to choose one of them.
- The aim is to maximize our rewards.

## Assumptions

- successive plays of an arm yield rewards which are i.i.d. realizations of the unknown distribution characterizing the arm.
- Rewards of different arms are independent.
- Rewards bounded.

## Notation (1/2)

$K$  = number of arms

$X_{k,t}$  = reward of arm  $k$  when drawn for the  $t$ -th time  
*Assumption* :  $0 \leq X_{k,t} \leq 1$

$\mu_k$  = expected reward of arm  $k$

$\sigma_k^2$  = variance of arm  $k$

$k^*$  = optimal arm :  $k^* \in \operatorname{argmax}_{k \in \{1, \dots, K\}} \mu_k$

$\mu^*$  = expected reward of the optimal arm =  $\mu_{k^*}$

## Notation (2/2)

$\Delta_k$  = expected regret of arm  $k$  :  $\Delta_k = \mu^* - \mu_k$

$\bar{X}_{k,t}$  =  $\frac{\sum_{i=1}^t X_{k,i}}{t}$

$V_{k,t}$  =  $\frac{\sum_{i=1}^t (X_{k,i} - \bar{X}_{k,t})^2}{t}$

policy = way of choosing the next arm to play

$I_t$  = arm played by the policy at time  $t$

$T_k(t)$  = number of times arm  $k$  is chosen by the policy during the first  $t$  plays

$R_n$  =  $\sum_{k=1}^K T_k(n) \Delta_k$

$\mathbb{E}R_n$  =  $\sum_{k=1}^K \mathbb{E}[T_k(n)] \Delta_k$

=  $n\mu^* - \mathbb{E}\left[\sum_{t=1}^n X_{I_t, T_{I_t}(t)}\right]$

# UCB policies

## General idea

- At time  $t$ , from past observations and some probabilistic argument, you have an upper confidence bound (UCB) on the expected rewards.
  - Play the arm having the largest UCB.
- 
- Why does it make sense?
  - Could we stay a long time drawing a wrong arm? No, since:
    - The more we draw a wrong arm  $k$  the closer the UCB gets to the expected reward  $\mu_k$ .
    - $\mu_k < \mu^* \leq \text{UCB on } \mu^*$ .

# UCB policies

## General idea

- At time  $t$ , from past observations and some probabilistic argument, you have an upper confidence bound (UCB) on the expected rewards.
  - Play the arm having the largest UCB.
- 
- Why does it make sense?
  - Could we stay a long time drawing a wrong arm? No, since:
    - The more we draw a wrong arm  $k$  the closer the UCB gets to the expected reward  $\mu_k$ ,
    - $\mu_k < \mu^* \leq \text{UCB on } \mu^*$ .

# UCB policies

## General idea

- At time  $t$ , from past observations and some probabilistic argument, you have an upper confidence bound (UCB) on the expected rewards.
  - Play the arm having the largest UCB.
- 
- Why does it make sense?
  - Could we stay a long time drawing a wrong arm? No, since:
    - The more we draw a wrong arm  $k$  the closer the UCB gets to the expected reward  $\mu_k$ ,
    - $\mu_k < \mu^* \leq \text{UCB on } \mu^*$ .

## Bernstein's type inequalities

- Let  $X, X_1, \dots, X_t$  be i.i.d. random variables taking their values in  $[0; 1]$ .
- Let  $\sigma^2 = \text{Var } X$ .
- Let  $\bar{X} = \frac{\sum_{i=1}^t X_i}{t}$  and  $S^2 = \frac{\sum_{i=1}^t (X_i - \bar{X})^2}{t}$ .

### Theorem (Bernstein inequality)

With probability at least  $1 - \beta$ , we have

$$|\mathbb{E}X - \bar{X}| \leq \sigma \sqrt{\frac{2 \log(2\beta^{-1})}{t}} + \frac{2 \log(2\beta^{-1})}{3t}.$$

### Theorem (Bernstein's type inequality)

With probability at least  $1 - \beta$ , we have

$$|\mathbb{E}X - \bar{X}| \leq S \sqrt{\frac{2 \log(3\beta^{-1})}{t}} + \frac{16 \log(3\beta^{-1})}{3t}.$$

## Sketch of the proof

- Start with Bernstein's inequality With probability at least  $1 - 2\beta/3$ , we have  $|\mathbb{E}X - \bar{X}| \leq \sigma \sqrt{\frac{2 \log(3\beta^{-1})}{t}} + \frac{2 \log(3\beta^{-1})}{3t}$ .
- It remains to control  $\sigma$  with probability  $1 - \beta/3$ . For this, we use Bernstein's inequality on the i.i.d. r.v.  $(X_i - \mathbb{E}X)^2$  and note that

$$\mathbb{E}(X_i - \mathbb{E}X)^2 = \sigma^2$$

and

$$\begin{aligned} \frac{1}{t} \sum_{i=1}^t (X_i - \mathbb{E}X)^2 &= \frac{1}{t} \sum_{i=1}^t (X_i - \bar{X})^2 + (\bar{X} - \mathbb{E}X)^2 \\ &= S^2 + (\mathbb{E}X - \bar{X})^2 \end{aligned}$$

since  $\mathbb{E}(W - a)^2 = \mathbb{E}(W - \mathbb{E}W)^2 + (\mathbb{E}W - a)^2$ .

## Sketch of the proof

- Start with Bernstein's inequality With probability at least  $1 - 2\beta/3$ , we have  $|\mathbb{E}X - \bar{X}| \leq \sigma \sqrt{\frac{2 \log(3\beta^{-1})}{t}} + \frac{2 \log(3\beta^{-1})}{3t}$ .
- It remains to control  $\sigma$  with probability  $1 - \beta/3$ . For this, we use Bernstein's inequality on the i.i.d. r.v.  $(X_i - \mathbb{E}X)^2$  and note that

$$\mathbb{E}(X_i - \mathbb{E}X)^2 = \sigma^2$$

and

$$\begin{aligned} \frac{1}{t} \sum_{i=1}^t (X_i - \mathbb{E}X)^2 &= \frac{1}{t} \sum_{i=1}^t (X_i - \bar{X})^2 + (\bar{X} - \mathbb{E}X)^2 \\ &= S^2 + (\mathbb{E}X - \bar{X})^2 \end{aligned}$$

since  $\mathbb{E}(W - a)^2 = \mathbb{E}(W - \mathbb{E}W)^2 + (\mathbb{E}W - a)^2$ .

# Definition

- $\beta > 0$
- Consider the (sub-)confidence levels

$$\beta_s \triangleq \frac{\beta}{4Ks(s+1)}.$$

- Let

$$B_{k,s} \triangleq \left( \bar{X}_{k,s} + \sqrt{\frac{2V_{k,s} \log(\beta_s^{-1})}{s}} + \frac{16 \log(\beta_s^{-1})}{3s} \right) \wedge 1$$

with the convention  $1/0 = +\infty$ .

$\beta$ -UCB policy

At time  $t$ , play an arm maximizing  $B_{k, T_k(t-1)}$ .

# Definition

- $\beta > 0$
- Consider the (sub-)confidence levels

$$\beta_s \triangleq \frac{\beta}{4Ks(s+1)}.$$

→ union bound over all arms and all possible number of draws

- Let

$$B_{k,s} \triangleq \left( \bar{X}_{k,s} + \sqrt{\frac{2V_{k,s} \log(\beta_s^{-1})}{s}} + \frac{16 \log(\beta_s^{-1})}{3s} \right) \wedge 1$$

with the convention  $1/0 = +\infty$ .

## $\beta$ -UCB policy

At time  $t$ , play an arm maximizing  $B_{k, T_k(t-1)}$ .

## Definition

- $\beta > 0$
- Consider the (sub-)confidence levels

$$\beta_s \triangleq \frac{\beta}{4Ks(s+1)}.$$

- Let

$$B_{k,s} \triangleq \left( \bar{X}_{k,s} + \sqrt{\frac{2V_{k,s} \log(\beta_s^{-1})}{s}} + \frac{16 \log(\beta_s^{-1})}{3s} \right) \wedge 1$$

with the convention  $1/0 = +\infty$ .

### $\beta$ -UCB policy

At time  $t$ , play an arm maximizing  $B_{k, T_k(t-1)}$ .

# A deviation inequality for the number of plays of non-optimal arms

## Theorem

For any non-optimal arm  $k$  (i.e.  $\Delta_k > 0$ ), consider  $u_k$  the smallest integer such that

$$\frac{u_k}{\log[4Ku_k(u_k+1)\beta^{-1}]} > \frac{8\sigma_k^2}{\Delta_k^2} + \frac{16}{\Delta_k}.$$

With probability at least  $1 - \beta$ , the  $\beta$ -UCB policy plays any non-optimal arm  $k$  at most  $u_k$  times.

# A deviation inequality for the number of plays of non-optimal arms

## Theorem

For any non-optimal arm  $k$  (i.e.  $\Delta_k > 0$ ), consider  $u_k$  the smallest integer such that

$$\frac{u_k}{\log[4K u_k (u_k + 1) \beta^{-1}]} > \frac{8\sigma_k^2}{\Delta_k^2} + \frac{16}{\Delta_k}.$$

With probability at least  $1 - \beta$ , the  $\beta$ -UCB policy plays any non-optimal arm  $k$  at most  $u_k$  times.

# A deviation inequality for the number of plays of non-optimal arms

## Theorem

For any non-optimal arm  $k$  (i.e.  $\Delta_k > 0$ ), consider  $u_k$  the smallest integer such that

$$\frac{u_k}{\log[4K u_k (u_k + 1) \beta^{-1}]} > \frac{8\sigma_k^2}{\Delta_k^2} + \frac{16}{\Delta_k}.$$

With probability at least  $1 - \beta$ , the  $\beta$ -UCB policy plays any non-optimal arm  $k$  at most  $u_k$  times.

## Lemma

Let  $w_k = \frac{8\sigma_k^2}{\Delta_k^2} + \frac{16}{\Delta_k}$ . We have  $u_k \leq 5w_k \log(w_k K \beta^{-1})$

# A deviation inequality for the number of plays of non-optimal arms

## Theorem

For any non-optimal arm  $k$  (i.e.  $\Delta_k > 0$ ), consider  $u_k$  the smallest integer such that

$$\frac{u_k}{\log[4Ku_k(u_k+1)\beta^{-1}]} > \frac{8\sigma_k^2}{\Delta_k^2} + \frac{16}{\Delta_k}.$$

With probability at least  $1 - \beta$ , the  $\beta$ -UCB policy plays any non-optimal arm  $k$  at most  $u_k$  times.

- With high probability, the number of plays of non-optimal arms is bounded by some quantity independent of the total number of plays !

# Cumulative regret bounds

## Theorem

With probability at least  $1 - \beta$ , for any time  $n$ ,

$$R_n = \sum_{k=1}^K T_k(n) \Delta_k \leq \sum_{k \in K} [u_k \wedge n] \Delta_k$$

For any  $n \geq 1$ , the  $1/n$ -UCB satisfies

$$\mathbb{E} R_n = \sum_{k=1}^K \mathbb{E}[T_k(n)] \Delta_k \leq C \log(2n) \sum_{k \neq k^*} \left(1 + \frac{\sigma_k^2}{\Delta_k}\right)$$

for some universal constant  $C > 0$ .

# Cumulative regret bounds

## Theorem

With probability at least  $1 - \beta$ , for any time  $n$ ,

$$R_n = \sum_{k=1}^K T_k(n) \Delta_k \leq \sum_{k \in K} [u_k \wedge n] \Delta_k$$

For any  $n \geq 1$ , the  $1/n$ -UCB satisfies

$$\mathbb{E}R_n = \sum_{k=1}^K \mathbb{E}[T_k(n)] \Delta_k \leq C \log(2n) \sum_{k \neq k^*} \left(1 + \frac{\sigma_k^2}{\Delta_k}\right)$$

for some universal constant  $C > 0$ .

Comparison with UCB1 [Auer, Cesa-Bianchi, Fischer (2002)]

$$\mathbb{E}R_n \leq C \log(2n) \sum_{k \neq k^*} \frac{1}{\Delta_k}.$$

# Cumulative regret bounds

## Theorem

With probability at least  $1 - \beta$ , for any time  $n$ ,

$$R_n = \sum_{k=1}^K T_k(n) \Delta_k \leq \sum_{k \in K} [u_k \wedge n] \Delta_k$$

For any  $n \geq 1$ , the  $1/n$ -UCB satisfies

$$\mathbb{E}R_n = \sum_{k=1}^K \mathbb{E}[T_k(n)] \Delta_k \leq C \log(2n) \sum_{k \neq k^*} \left(1 + \frac{\sigma_k^2}{\Delta_k}\right)$$

for some universal constant  $C > 0$ .

## Comparison with UCB1-normal [Auer, Cesa-Bianchi, Fischer (2002)]

When the rewards have normal distributions, we have

$$\mathbb{E}R_n \leq C \log(2n) \sum_{k \neq k^*} \left(\Delta_k + \frac{\sigma_k^2}{\Delta_k}\right)$$

# Discussion on the $1/n$ -UCB policy

## Drawback

it needs to know the total number of plays.

## First solution

the *doubling trick*, i.e. cut time space into intervals of length  $2^{2^l}$  and launch the algorithm independently on each of these epochs.

## Drawback of this first solution

it needs to restart the algorithm at each epoch.

## Second solution

the *UCB-tuned* policy

## Discussion on the $1/n$ -UCB policy

### Drawback

it needs to know the total number of plays.

### First solution

the *doubling trick*, i.e. cut time space into intervals of length  $2^{2^l}$  and launch the algorithm independently on each of these epochs.

### Drawback of this first solution

it needs to restart the algorithm at each epoch.

### Second solution

the *UCB-tuned* policy

## Discussion on the $1/n$ -UCB policy

### Drawback

it needs to know the total number of plays.

### First solution

the *doubling trick*, i.e. cut time space into intervals of length  $2^{2^l}$  and launch the algorithm independently on each of these epochs.

### Drawback of this first solution

it needs to restart the algorithm at each epoch.

### Second solution

the *UCB-tuned* policy

## Discussion on the $1/n$ -UCB policy

### Drawback

it needs to know the total number of plays.

### First solution

the *doubling trick*, i.e. cut time space into intervals of length  $2^{2^l}$  and launch the algorithm independently on each of these epochs.

### Drawback of this first solution

it needs to restart the algorithm at each epoch.

### Second solution

the *UCB-tuned* policy

## Definition

- Let  $p > 2$  and

$$B_{k,s,t} \triangleq \left( \bar{X}_{k,s} + \sqrt{\frac{2V_{k,s} \log(4t^p)}{s}} + \frac{16 \log(4t^p)}{3s} \right) \wedge 1$$

with the convention  $1/0 = +\infty$ .

### UCB-tuned policy

At time  $t$ , play an arm maximizing  $B_{k, T_k(t-1), t}$ .

- Underlying idea:* the larger the total number of plays is, the more confident we should be before discarding a non-promising (but still possibly optimal) arm.

## Definition

- Let  $p > 2$  and

$$B_{k,s,t} \triangleq \left( \bar{X}_{k,s} + \sqrt{\frac{2V_{k,s} \log(4t^p)}{s}} + \frac{16 \log(4t^p)}{3s} \right) \wedge 1$$

with the convention  $1/0 = +\infty$ .

### UCB-tuned policy

At time  $t$ , play an arm maximizing  $B_{k, T_k(t-1), t}$ .

- Underlying idea:* the larger the total number of plays is, the more confident we should be before discarding a non-promising (but still possibly optimal) arm.

## Definition

- Let  $p > 2$  and

$$B_{k,s,t} \triangleq \left( \bar{X}_{k,s} + \sqrt{\frac{2V_{k,s} \log(\beta_s)}{s}} + \frac{16 \log(\beta_s)}{3s} \right) \wedge 1$$

with the convention  $1/0 = +\infty$ .

$$\beta_s \triangleq \frac{\beta}{4Ks(s+1)}.$$

### UCB-tuned policy

At time  $t$ , play an arm maximizing  $B_{k, T_k(t-1), t}$ .

- Underlying idea:* the larger the total number of plays is, the more confident we should be before discarding a non-promising (but still possibly optimal) arm.

## Definition

- Let  $p > 2$  and

$$B_{k,s,t} \triangleq \left( \bar{X}_{k,s} + \sqrt{\frac{2V_{k,s} \log(4t^p)}{s}} + \frac{16 \log(4t^p)}{3s} \right) \wedge 1$$

with the convention  $1/0 = +\infty$ .

### UCB-tuned policy

At time  $t$ , play an arm maximizing  $B_{k, T_k(t-1), t}$ .

- Underlying idea:* the larger the total number of plays is, the more confident we should be before discarding a non-promising (but still possibly optimal) arm.

# Expected cumulative regret bound

## Theorem

For any time  $n$ , the UCB-tuned policy satisfies

$$\mathbb{E}R_n \leq 16 \log(4n^p) \sum_{k \neq k^*} \left( 1 + \frac{\sigma_k^2}{2\Delta_k} \right) + \frac{2p-2}{p-2} \sum_{k \neq k^*} \Delta_k.$$

Comparison with UCB1-normal [Auer, Cesa-Bianchi, Fischer (2002)]

When the rewards have normal distributions, we have

$$\mathbb{E}R_n \leq 256(\log n) \sum_{k \neq k^*} \frac{\sigma_k^2}{\Delta_k} + (6 + 8 \log n) \sum_{k \neq k^*} \Delta_k$$

# Expected cumulative regret bound

## Theorem

For any time  $n$ , the UCB-tuned policy satisfies

$$\mathbb{E}R_n \leq 16 \log(4n^p) \sum_{k \neq k^*} \left( 1 + \frac{\sigma_k^2}{2\Delta_k} \right) + \frac{2p-2}{p-2} \sum_{k \neq k^*} \Delta_k.$$

## Comparison with UCB1-normal [Auer, Cesa-Bianchi, Fischer (2002)]

When the rewards have normal distributions, we have

$$\mathbb{E}R_n \leq 256(\log n) \sum_{k \neq k^*} \frac{\sigma_k^2}{\Delta_k} + (6 + 8 \log n) \sum_{k \neq k^*} \Delta_k$$

## Sketch of the proof (1/3)

Let

$$b_{k,s,t} = \mu_k + \sqrt{\frac{8\sigma_k^2 \log(4t^\rho)}{s}} + \frac{8 \log(4t^\rho)}{s}.$$

The choice of the confidence sequence  $1/(4t^\rho)$  is such that:

$$\forall k \quad \forall 1 \leq s \leq t \quad \begin{cases} \mathbb{P}(\mu_k \leq B_{k,s,t}) \geq 1 - t^{-\rho} \\ \mathbb{P}(B_{k,s,t} \leq b_{k,s,t}) \geq 1 - t^{-\rho} \end{cases}$$

## Sketch of the proof (2/3)

Let  $k$  be such that  $\mu_k < \mu^*$ . Let  $u_{k,n} \geq 1$  be an integer-valued sequence to be chosen. We have

$$\begin{aligned}
 T_k(n) &= \sum_{t=1}^n \mathbb{1}_{I_t=k} \\
 &\leq u_{k,n} - 1 + \sum_{t=1}^n \mathbb{1}_{I_t=k; T_k(t) \geq u_{k,n}} \\
 &\leq u_{k,n} - 1 + \sum_{t=1}^n \mathbb{1}_{B_{k^*, T_{k^*}(t)}, t \leq B_{k, T_k(t), t}; T_k(t) \geq u_{k,n}} \\
 &\leq u_{k,n} - 1 + \sum_{t=1}^n \mathbb{1}_{B_{k^*, T_{k^*}(t)}, t \leq \mu^*} \\
 &\quad + \sum_{t=1}^n \mathbb{1}_{\mu^* < b_{k, T_k(t), t}; T_k(t) \geq u_{k,n}} + \sum_{t=1}^n \mathbb{1}_{b_{k, T_k(t), t} \leq B_{k, T_k(t), t}} \\
 &\leq u_{k,n} - 1 + \sum_{t=1}^n \sum_{s=1}^t \mathbb{1}_{B_{k^*, s, t} \leq \mu^*} \\
 &\quad + \sum_{t=1}^n \sum_{s=u_{k,n}}^t \mathbb{1}_{\mu^* < b_{k, s, t}} + \sum_{t=1}^n \sum_{s=1}^t \mathbb{1}_{b_{k, s, t} \leq B_{k, s, t}}
 \end{aligned}$$

## Sketch of the proof (2/3)

Let  $k$  be such that  $\mu_k < \mu^*$ . Let  $u_{k,n} \geq 1$  be an integer-valued sequence to be chosen. We have

$$\begin{aligned}
 T_k(n) &= \sum_{t=1}^n \mathbb{1}_{I_t=k} \\
 &\leq u_{k,n} - 1 + \sum_{t=1}^n \mathbb{1}_{I_t=k; T_k(t) \geq u_{k,n}} \\
 &\leq u_{k,n} - 1 + \sum_{t=1}^n \mathbb{1}_{B_{k^*, T_{k^*}(t), t} \leq B_{k, T_k(t), t}; T_k(t) \geq u_{k,n}} \\
 &\leq u_{k,n} - 1 + \sum_{t=1}^n \mathbb{1}_{B_{k^*, T_{k^*}(t), t} \leq \mu^*} \\
 &\quad + \sum_{t=1}^n \mathbb{1}_{\mu^* < b_{k, T_k(t), t}; T_k(t) \geq u_{k,n}} + \sum_{t=1}^n \mathbb{1}_{b_{k, T_k(t), t} \leq B_{k, T_k(t), t}} \\
 &\leq u_{k,n} - 1 + \sum_{t=1}^n \sum_{s=1}^t \mathbb{1}_{B_{k^*, s, t} \leq \mu^*} \\
 &\quad + \sum_{t=1}^n \sum_{s=u_{k,n}}^t \mathbb{1}_{\mu^* < b_{k, s, t}} + \sum_{t=1}^n \sum_{s=1}^t \mathbb{1}_{b_{k, s, t} \leq B_{k, s, t}}
 \end{aligned}$$

## Sketch of the proof (3/3)

Taking expectation, we obtain

$$\begin{aligned}
 \mathbb{E}T_k(n) &\leq u_{k,n} - 1 + \sum_{t=1}^n \sum_{s=1}^t t^{-p} \\
 &\quad + \sum_{t=1}^n \sum_{s=u_{k,n}}^t \mathbb{P}(\mu^* < b_{k,s,t}) + \sum_{t=1}^n \sum_{s=1}^t t^{-p} \\
 &\leq u_{k,n} - 1 + \frac{2p-2}{p-2} + \sum_{t=1}^n \sum_{s=u_{k,n}}^t \mathbb{1}_{\mu^* < b_{k,s,t}} \\
 &\leq u_{k,n} - 1 + \frac{2p-2}{p-2} + n^2 \mathbb{1}_{\mu^* < b_{k,u_{k,n},n}}
 \end{aligned}$$

since we have  $b_{k,u_{k,n},n} \geq b_{k,s,t}$  for  $1 \leq u_{k,n} \leq s \leq t \leq n$ . Taking  $u_{k,n} \triangleq 1 + \log(4n^p) \left( \frac{8\sigma_k^2}{\Delta_k^2} + \frac{16}{\Delta_k} \right)$ , i.e. close to the smallest threshold ensuring that  $\mu^* \geq b_{k,u_{k,n},n}$ , we obtain

$$\mathbb{E}T_k(n) \leq \log(4n^p) \left( \frac{8\sigma_k^2}{\Delta_k^2} + \frac{16}{\Delta_k} \right) + \frac{2p-2}{p-2},$$

hence the desired bound on  $\mathbb{E}R_n = \sum_{k \neq k^*} \mathbb{E}[T_k(n)]\Delta_k$ .

# Conclusion

- We propose a UCB algorithm satisfying: with high probability, the cumulative regret after  $n$  plays is bounded by a constant independent from  $n$  and depending weakly on the confidence level.
- We prove a logarithmic bound on the expected regret of a variant of UCB-tuned that takes advantage of variance estimates