Best Arm Identification in Multi-Armed Bandits

Sébastien Bubeck[1]

joint work with Jean-Yves Audibert[2,3] & Rémi Munos[1]

[1] INRIA Lille, SequeL team
[2] Univ. Paris Est, Imagine
[3] CNRS/ENS/INRIA, Willow project

Pure exploration bandit game

Parameters available to the forecaster: the number of rounds n and the number of arms K.

Parameters unknown to the forecaster: the reward distributions (over $[0,1]$) $\nu_1, \ldots, \nu_K$ of the arms. We assume that there is a unique arm i^* with maximal mean.

For each round $t = 1, 2, \ldots, n$;

1. The forecaster chooses an arm $I_t \in \{1, \ldots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

At the end of the n rounds the forecaster outputs a recommendation $J_n \in \{1, \ldots, K\}$.

Goal: Find the best arm, i.e, the arm with maximal mean. We denote

$$e_n = \mathbb{P}(J_n \neq i^*).$$

Pure exploration bandit game

Parameters available to the forecaster: the number of rounds n and the number of arms K.

Parameters unknown to the forecaster: the reward distributions (over $[0,1]$) $\nu_1, \ldots, \nu_K$ of the arms. We assume that there is a unique arm i^* with maximal mean.

For each round $t = 1, 2, \ldots, n$;

1. The forecaster chooses an arm $I_t \in \{1, \ldots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

At the end of the n rounds the forecaster outputs a recommendation $J_n \in \{1, \ldots, K\}$.

Goal: Find the best arm, i.e, the arm with maximal mean. We denote

$$e_n = \mathbb{P}(J_n \neq i^*).$$

Pure exploration bandit game

Parameters available to the forecaster: the number of rounds n and the number of arms K.

Parameters unknown to the forecaster: the reward distributions (over $[0,1]$) $\nu_1, \ldots, \nu_K$ of the arms. We assume that there is a unique arm i^* with maximal mean.

For each round $t = 1, 2, \ldots, n$;

1. The forecaster chooses an arm $I_t \in \{1, \ldots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

At the end of the n rounds the forecaster outputs a recommendation $J_n \in \{1, \ldots, K\}$.

Goal: Find the best arm, i.e, the arm with maximal mean. We denote

$$e_n = \mathbb{P}(J_n \neq i^*).$$

Pure exploration bandit game

Parameters available to the forecaster: the number of rounds n and the number of arms K.

Parameters unknown to the forecaster: the reward distributions (over $[0,1]$) $\nu_1, \dots, \nu_K$ of the arms. We assume that there is a unique arm i^* with maximal mean.

For each round $t = 1, 2, \dots, n$;

1. The forecaster chooses an arm $I_t \in \{1, \dots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

At the end of the n rounds the forecaster outputs a recommendation $J_n \in \{1, \dots, K\}$.

Goal: Find the best arm, i.e, the arm with maximal mean. We denote

$$e_n = \mathbb{P}(J_n \neq i^*).$$

Pure exploration bandit game

Parameters available to the forecaster: the number of rounds n and the number of arms K.

Parameters unknown to the forecaster: the reward distributions (over $[0, 1]$) $\nu_1, \dots, \nu_K$ of the arms. We assume that there is a unique arm i^* with maximal mean.

For each round $t = 1, 2, \dots, n$;

1. The forecaster chooses an arm $I_t \in \{1, \dots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

At the end of the n rounds the forecaster outputs a recommendation $J_n \in \{1, \dots, K\}$.

Goal: Find the best arm, i.e, the arm with maximal mean. We denote

$$e_n = \mathbb{P}(J_n \neq i^*).$$

Pure exploration bandit game

Parameters available to the forecaster: the number of rounds n and the number of arms K.

Parameters unknown to the forecaster: the reward distributions (over $[0, 1]$) $\nu_1, \dots, \nu_K$ of the arms. We assume that there is a unique arm i^* with maximal mean.

For each round $t = 1, 2, \dots, n$;

1. The forecaster chooses an arm $I_t \in \{1, \dots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

At the end of the n rounds the forecaster outputs a recommendation $J_n \in \{1, \dots, K\}$.

Goal: Find the best arm, i.e, the arm with maximal mean. We denote

$$e_n = \mathbb{P}(J_n \neq i^*).$$

Pure exploration bandit game

Parameters available to the forecaster: the number of rounds n and the number of arms K.

Parameters unknown to the forecaster: the reward distributions (over $[0,1]$) $\nu_1, \dots, \nu_K$ of the arms. We assume that there is a unique arm i^* with maximal mean.

For each round $t = 1, 2, \dots, n$;

1. The forecaster chooses an arm $I_t \in \{1, \dots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

At the end of the n rounds the forecaster outputs a recommendation $J_n \in \{1, \dots, K\}$.

Goal: Find the best arm, i.e, the arm with maximal mean. We denote

$$e_n = \mathbb{P}(J_n \neq i^*).$$

Motivating examples

- Clinical trials for cosmetic products. During the **test phase**, several several formulæ for a cream are **sequentially tested**, and after a finite time **one is chosen** for commercialization.
- Channel allocation for mobile phone communications. Cellphones can **explore the set of channels** to find the best one to operate. Each **evaluation** of a channel is **noisy** and there is a **limited number** of evaluations before the communication starts on **the chosen channel**.

Motivating examples

- Clinical trials for cosmetic products. During the **test phase**, several several formulæ for a cream are **sequentially tested**, and after a finite time **one is chosen** for commercialization.
- Channel allocation for mobile phone communications. Cellphones can **explore the set of channels** to find the best one to operate. Each **evaluation** of a channel is **noisy** and there is a **limited number** of evaluations before the communication starts on **the chosen channel**.

Summary of the talk

- Let μ_i be the mean of ν_i, and $\Delta_i = \mu_{i^*} - \mu_i$ the suboptimality of arm i.
- Main theoretical result: it requires of order of $H = \sum_{i \neq i^*} 1/\Delta_i^2$ rounds to find the best arm. Note that this result is well known for $K = 2$.
- We present two new forecasters, **Successive Rejects (SR)** and **Adaptive UCB-E (Upper Confidence Bound Exploration)**.
- SR is parameter free, and has optimal guarantees (up to a logarithmic factor).
- Adaptive UCB-E has no theoretical guarantees but it experimentally outperforms SR.

Summary of the talk

- Let μ_i be the mean of ν_i, and $\Delta_i = \mu_{i^*} - \mu_i$ the suboptimality of arm i.
- Main theoretical result: it requires of order of $H = \sum_{i \neq i^*} 1/\Delta_i^2$ rounds to find the best arm. Note that this result is well known for $K = 2$.
- We present two new forecasters, **Successive Rejects (SR)** and **Adaptive UCB-E (Upper Confidence Bound Exploration)**.
- SR is parameter free, and has optimal guarantees (up to a logarithmic factor).
- Adaptive UCB-E has no theoretical guarantees but it experimentally outperforms SR.

Summary of the talk

- Let μ_i be the mean of ν_i, and $\Delta_i = \mu_{i^*} - \mu_i$ the suboptimality of arm i.
- Main theoretical result: it requires of order of $H = \sum_{i \neq i^*} 1/\Delta_i^2$ rounds to find the best arm. Note that this result is well known for $K = 2$.
- We present two new forecasters, **Successive Rejects (SR)** and **Adaptive UCB-E (Upper Confidence Bound Exploration)**.
- SR is parameter free, and has optimal guarantees (up to a logarithmic factor).
- Adaptive UCB-E has no theoretical guarantees but it experimentally outperforms SR.

Summary of the talk

- Let μ_i be the mean of ν_i, and $\Delta_i = \mu_{i^*} - \mu_i$ the suboptimality of arm i.
- Main theoretical result: it requires of order of $H = \sum_{i \neq i^*} 1/\Delta_i^2$ rounds to find the best arm. Note that this result is well known for $K = 2$.
- We present two new forecasters, **Successive Rejects (SR)** and **Adaptive UCB-E (Upper Confidence Bound Exploration)**.
- SR is parameter free, and has optimal guarantees (up to a logarithmic factor).
- Adaptive UCB-E has no theoretical guarantees but it experimentally outperforms SR.

Summary of the talk

- Let μ_i be the mean of ν_i, and $\Delta_i = \mu_{i^*} - \mu_i$ the suboptimality of arm i.
- Main theoretical result: it requires of order of $H = \sum_{i \neq i^*} 1/\Delta_i^2$ rounds to find the best arm. Note that this result is well known for $K = 2$.
- We present two new forecasters, **Successive Rejects (SR)** and **Adaptive UCB-E (Upper Confidence Bound Exploration)**.
- SR is parameter free, and has optimal guarantees (up to a logarithmic factor).
- Adaptive UCB-E has no theoretical guarantees but it experimentally outperforms SR.

Lower Bound

Theorem

Let $\nu_1, \ldots, \nu_K$ be Bernoulli distributions with parameters in $[1/3, 2/3]$. For any forecaster, there exists a numerical constant $c > 0$ such that, up to a permutation of the arms,

$$e_n \geq \exp\left(-c \frac{n \log(K)}{H}\right).$$

Informally, any algorithm requires at least (of order of) $H/\log(K)$ rounds to find the best arm.

Uniform strategy

For each $i \in \{1, \dots, K\}$, select arm i during $\lfloor n/K \rfloor$ rounds.

Theorem

There exists a numerical constant $c > 0$ such that the uniform strategy satisfies:

$$e_n \leq \exp\left(-c \frac{n \min_i \Delta_i^2}{K}\right).$$

Informally, the uniform strategy finds the best arm with (of order of) $K/\min_i \Delta_i^2$ rounds. For large K, this can be significantly larger than $H = \sum_{i \neq i^*} 1/\Delta_i^2$.

Uniform strategy

For each $i \in \{1, \ldots, K\}$, select arm i during $\lfloor n/K \rfloor$ rounds.

Theorem

There exists a numerical constant $c > 0$ such that the uniform strategy satisfies:

$$e_n \leq \exp\left(-c\frac{n \min_i \Delta_i^2}{K}\right).$$

Informally, the uniform strategy finds the best arm with (of order of) $K/\min_i \Delta_i^2$ rounds. For large K, this can be significantly larger than $H = \sum_{i \neq i^*} 1/\Delta_i^2$.

Uniform strategy

For each $i \in \{1, \dots, K\}$, select arm i during $\lfloor n/K \rfloor$ rounds.

Theorem

There exists a numerical constant $c > 0$ such that the uniform strategy satisfies:

$$e_n \leq \exp\left(-c\frac{n \min_i \Delta_i^2}{K}\right).$$

Informally, the uniform strategy finds the best arm with (of order of) $K / \min_i \Delta_i^2$ rounds. For large K, this can be significantly larger than $H = \sum_{i \neq i^*} 1/\Delta_i^2$.

Successive Rejects (SR)

Let $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $A_1 = \{1, \ldots, K\}$, $n_0 = 0$ and $n_k = \lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \rceil$ for $k \in \{1, \ldots, K-1\}$.

For each phase $k = 1, 2, \ldots, K-1$:

(1) For each $i \in A_k$, select arm i during $n_k - n_{k-1}$ rounds.

(2) Let $A_{k+1} = A_k \setminus \arg\min_{i \in A_k} \widehat{X}_{i,n_k}$, where $\widehat{X}_{i,s}$ represents the empirical mean of arm i after s pulls.

Let J_n be the unique element of A_K.

Theorem

There exists a numerical constant $c > 0$ such that SR satisfies:

$$e_n \leq \exp\left(-c \frac{n}{\overline{\log}(K) H}\right).$$

Successive Rejects (SR)

Let $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $A_1 = \{1, \ldots, K\}$, $n_0 = 0$ and $n_k = \lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \rceil$ for $k \in \{1, \ldots, K-1\}$.

For each phase $k = 1, 2, \ldots, K-1$:

(1) For each $i \in A_k$, select arm i during $n_k - n_{k-1}$ rounds.

(2) Let $A_{k+1} = A_k \setminus \arg\min_{i \in A_k} \widehat{X}_{i,n_k}$, where $\widehat{X}_{i,s}$ represents the empirical mean of arm i after s pulls.

Let J_n be the unique element of A_K.

Theorem

There exists a numerical constant $c > 0$ such that SR satisfies:

$$e_n \leq \exp\left(-c \frac{n}{\overline{\log}(K) H}\right).$$

Successive Rejects (SR)

Let $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $A_1 = \{1, \ldots, K\}$, $n_0 = 0$ and $n_k = \lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \rceil$ for $k \in \{1, \ldots, K-1\}$.

For each phase $k = 1, 2, \ldots, K-1$:

(1) For each $i \in A_k$, select arm i during $n_k - n_{k-1}$ rounds.

(2) Let $A_{k+1} = A_k \setminus \arg\min_{i \in A_k} \widehat{X}_{i,n_k}$, where $\widehat{X}_{i,s}$ represents the empirical mean of arm i after s pulls.

Let J_n be the unique element of A_K.

Theorem

There exists a numerical constant $c > 0$ such that SR satisfies:

$$e_n \leq \exp\left(-c \frac{n}{\overline{\log}(K) H}\right).$$

Successive Rejects (SR)

Let $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $A_1 = \{1, \dots, K\}$, $n_0 = 0$ and $n_k = \lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \rceil$ for $k \in \{1, \dots, K-1\}$.

For each phase $k = 1, 2, \dots, K-1$:

(1) For each $i \in A_k$, select arm i during $n_k - n_{k-1}$ rounds.

(2) Let $A_{k+1} = A_k \setminus \arg\min_{i \in A_k} \widehat{X}_{i,n_k}$, where $\widehat{X}_{i,s}$ represents the empirical mean of arm i after s pulls.

Let J_n be the unique element of A_K.

Theorem

There exists a numerical constant $c > 0$ such that SR satisfies:

$$e_n \leq \exp\left(-c \frac{n}{\overline{\log}(K) H}\right).$$

Successive Rejects (SR)

Let $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $A_1 = \{1, \ldots, K\}$, $n_0 = 0$ and $n_k = \lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \rceil$ for $k \in \{1, \ldots, K-1\}$.

For each phase $k = 1, 2, \ldots, K-1$:

(1) For each $i \in A_k$, select arm i during $n_k - n_{k-1}$ rounds.

(2) Let $A_{k+1} = A_k \setminus \arg\min_{i \in A_k} \widehat{X}_{i,n_k}$, where $\widehat{X}_{i,s}$ represents the empirical mean of arm i after s pulls.

Let J_n be the unique element of A_K.

Theorem

There exists a numerical constant $c > 0$ such that SR satisfies:

$$e_n \leq \exp\left(-c \frac{n}{\overline{\log}(K) H}\right).$$

Successive Rejects (SR)

Let $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $A_1 = \{1, \dots, K\}$, $n_0 = 0$ and $n_k = \lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \rceil$ for $k \in \{1, \dots, K-1\}$.

For each phase $k = 1, 2, \dots, K-1$:

(1) For each $i \in A_k$, select arm i during $n_k - n_{k-1}$ rounds.

(2) Let $A_{k+1} = A_k \setminus \arg\min_{i \in A_k} \widehat{X}_{i,n_k}$, where $\widehat{X}_{i,s}$ represents the empirical mean of arm i after s pulls.

Let J_n be the unique element of A_K.

Theorem

There exists a numerical constant $c > 0$ such that SR satisfies:

$$e_n \leq \exp\left(-c \frac{n}{\overline{\log}(K) H}\right).$$

Successive Rejects (SR)

Let $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $A_1 = \{1, \ldots, K\}$, $n_0 = 0$ and $n_k = \lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \rceil$ for $k \in \{1, \ldots, K-1\}$.

For each phase $k = 1, 2, \ldots, K-1$:

(1) For each $i \in A_k$, select arm i during $n_k - n_{k-1}$ rounds.

(2) Let $A_{k+1} = A_k \setminus \arg\min_{i \in A_k} \widehat{X}_{i,n_k}$, where $\widehat{X}_{i,s}$ represents the empirical mean of arm i after s pulls.

Let J_n be the unique element of A_K.

Theorem

There exists a numerical constant $c > 0$ such that SR satisfies:

$$e_n \leq \exp\left(-c \frac{n}{\overline{\log}(K) H}\right).$$

UCB-E

Parameter: exploration rate $c > 0$.

For $t \geq 1, i \in \{1, \dots, K\}$ let $B_{i,t} = \widehat{X}_{i,T_i(t)} + \sqrt{\frac{c\ n/H}{T_i(t)}}$, where $T_i(t)$ represents the number of times we selected arm i up to time t.

For each round $t = 1, 2, \dots, n$:
Draw $I_t \in \mathrm{argmax}_{i \in \{1,\dots,K\}} B_{i,t-1}$.

Let $J_n \in \mathrm{argmax}_{i \in \{1,\dots,K\}} \widehat{X}_{i,T_i(n)}$.

Theorem

For c small enough, there exists a numerical constant $c' > 0$ such that UCB-E satisfies $e_n \leq \exp\left(-c' n/H\right)$.

UCB-E finds the best arm with (of order of) H rounds, but it requires the knowledge of H.

UCB-E

Parameter: exploration rate $c > 0$.

For $t \geq 1, i \in \{1, \ldots, K\}$ let $B_{i,t} = \widehat{X}_{i,T_i(t)} + \sqrt{\frac{c\ n/H}{T_i(t)}}$, where $T_i(t)$ represents the number of times we selected arm i up to time t.

For each round $t = 1, 2, \ldots, n$:
Draw $I_t \in \mathrm{argmax}_{i \in \{1,\ldots,K\}} B_{i,t-1}$.

Let $J_n \in \mathrm{argmax}_{i \in \{1,\ldots,K\}} \widehat{X}_{i,T_i(n)}$.

Theorem

For c small enough, there exists a numerical constant $c' > 0$ such that UCB-E satisfies $e_n \leq \exp(-c' n/H)$.

UCB-E finds the best arm with (of order of) H rounds, but it requires the knowledge of H.

UCB-E

Parameter: exploration rate $c > 0$.

For $t \geq 1, i \in \{1, \dots, K\}$ let $B_{i,t} = \widehat{X}_{i,T_i(t)} + \sqrt{\frac{c\ n/H}{T_i(t)}}$, where $T_i(t)$ represents the number of times we selected arm i up to time t.

For each round $t = 1, 2, \dots, n$:
Draw $I_t \in \mathrm{argmax}_{i \in \{1,\dots,K\}} B_{i,t-1}$.

Let $J_n \in \mathrm{argmax}_{i \in \{1,\dots,K\}} \widehat{X}_{i,T_i(n)}$.

Theorem

For c small enough, there exists a numerical constant $c' > 0$ such that UCB-E satisfies $e_n \leq \exp\left(-c' n/H\right)$.

UCB-E finds the best arm with (of order of) H rounds, but it requires the knowledge of H.

UCB-E

Parameter: exploration rate $c > 0$.

For $t \geq 1, i \in \{1, \dots, K\}$ let $B_{i,t} = \widehat{X}_{i,T_i(t)} + \sqrt{\frac{c\ n/H}{T_i(t)}}$, where $T_i(t)$ represents the number of times we selected arm i up to time t.

For each round $t = 1, 2, \dots, n$:
Draw $I_t \in \mathrm{argmax}_{i \in \{1,\dots,K\}} B_{i,t-1}$.

Let $J_n \in \mathrm{argmax}_{i \in \{1,\dots,K\}} \widehat{X}_{i,T_i(n)}$.

Theorem

For c small enough, there exists a numerical constant $c' > 0$ such that UCB-E satisfies $e_n \leq \exp\left(-c' n/H\right)$.

UCB-E finds the best arm with (of order of) H rounds, but it requires the knowledge of H.

UCB-E

Parameter: exploration rate $c > 0$.

For $t \geq 1, i \in \{1, \dots, K\}$ let $B_{i,t} = \widehat{X}_{i,T_i(t)} + \sqrt{\frac{c\, n/H}{T_i(t)}}$, where $T_i(t)$ represents the number of times we selected arm i up to time t.

For each round $t = 1, 2, \dots, n$:
Draw $I_t \in \mathrm{argmax}_{i \in \{1,\dots,K\}} B_{i,t-1}$.

Let $J_n \in \mathrm{argmax}_{i \in \{1,\dots,K\}} \widehat{X}_{i,T_i(n)}$.

Theorem

For c small enough, there exists a numerical constant $c' > 0$ such that UCB-E satisfies $e_n \leq \exp\left(-c' n/H\right)$.

UCB-E finds the best arm with (of order of) H rounds, but it requires the knowledge of H.

UCB-E

Parameter: exploration rate $c > 0$.

For $t \geq 1, i \in \{1, \dots, K\}$ let $B_{i,t} = \widehat{X}_{i,T_i(t)} + \sqrt{\frac{c\ n/H}{T_i(t)}}$, where $T_i(t)$ represents the number of times we selected arm i up to time t.

For each round $t = 1, 2, \dots, n$:
Draw $I_t \in \mathrm{argmax}_{i \in \{1,\dots,K\}} B_{i,t-1}$.

Let $J_n \in \mathrm{argmax}_{i \in \{1,\dots,K\}} \widehat{X}_{i,T_i(n)}$.

Theorem

For c small enough, there exists a numerical constant $c' > 0$ such that UCB-E satisfies $e_n \leq \exp(-c' n/H)$.

UCB-E finds the best arm with (of order of) H rounds, but it requires the knowledge of H.

UCB-E

Parameter: exploration rate $c > 0$.

For $t \geq 1, i \in \{1, \dots, K\}$ let $B_{i,t} = \widehat{X}_{i,T_i(t)} + \sqrt{\frac{c\ n/H}{T_i(t)}}$, where $T_i(t)$ represents the number of times we selected arm i up to time t.

For each round $t = 1, 2, \dots, n$:
Draw $I_t \in \mathrm{argmax}_{i \in \{1,\dots,K\}} B_{i,t-1}$.

Let $J_n \in \mathrm{argmax}_{i \in \{1,\dots,K\}} \widehat{X}_{i,T_i(n)}$.

Theorem

For c small enough, there exists a numerical constant $c' > 0$ such that UCB-E satisfies $e_n \leq \exp\left(-c' n/H\right)$.

UCB-E finds the best arm with (of order of) H rounds, but it requires the knowledge of H.

Adaptive UCB-E

Parameter: exploration rate $c > 0$.

Definitions: For $k \in \{1, \dots, K-1\}$, let $n_k = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \right\rceil$, $t_0 = 0$, $t_1 = K n_1$, and for $k > 1$, $t_k = n_1 + \dots n_{k-1} + (K-k+1) n_k$. For $i \in \{1, \dots, K\}$ and $a > 0$, let $B_{i,t}(a) = \widehat{X}_{i,T_i(t)} + \sqrt{\frac{a}{T_i(t)}}$ for $t \geq 1$.

Algorithm: For each phase $k = 0, 1, \dots, K-1$:
Let $\widehat{H}_k = K$ if $k = 0$, and otherwise $\widehat{H}_k = \max_{K-k+1 \leq i \leq K} i \widehat{\Delta}^{-2}_{<i>,k}$, where $\widehat{\Delta}_{i,k} = \left(\max_{1 \leq j \leq K} \widehat{X}_{j,T_j(t_k)}\right) - \widehat{X}_{i,T_i(t_k)}$ and $< i >$ is an ordering such that $\widehat{\Delta}_{<1>,k} \leq \dots \leq \widehat{\Delta}_{<K>,k}$.

For $t = t_k + 1, \dots, t_{k+1}$: Draw
$I_t \in \mathrm{argmax}_{i \in \{1,\dots,K\}} B_{i,t-1}(c\, n / \widehat{H}_k)$.

Adaptive UCB-E

Parameter: exploration rate $c > 0$.

Definitions: For $k \in \{1, \dots, K-1\}$, let $n_k = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \right\rceil$, $t_0 = 0$, $t_1 = Kn_1$, and for $k > 1$, $t_k = n_1 + \dots n_{k-1} + (K-k+1)n_k$. For $i \in \{1, \dots, K\}$ and $a > 0$, let $B_{i,t}(a) = \widehat{X}_{i,T_i(t)} + \sqrt{\frac{a}{T_i(t)}}$ for $t \geq 1$.

Algorithm: For each phase $k = 0, 1, \dots, K-1$:
Let $\widehat{H}_k = K$ if $k = 0$, and otherwise $\widehat{H}_k = \max_{K-k+1 \leq i \leq K} i \widehat{\Delta}_{<i>,k}^{-2}$, where $\widehat{\Delta}_{i,k} = \left(\max_{1 \leq j \leq K} \widehat{X}_{j,T_j(t_k)} \right) - \widehat{X}_{i,T_i(t_k)}$ and $<i>$ is an ordering such that $\widehat{\Delta}_{<1>,k} \leq \dots \leq \widehat{\Delta}_{<K>,k}$.

For $t = t_k + 1, \dots, t_{k+1}$: Draw
$I_t \in \mathrm{argmax}_{i \in \{1,\dots,K\}} B_{i,t-1}(c\, n/\widehat{H}_k)$.

Adaptive UCB-E

Parameter: exploration rate $c > 0$.

Definitions: For $k \in \{1, \dots, K-1\}$, let $n_k = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \right\rceil$, $t_0 = 0$, $t_1 = Kn_1$, and for $k > 1$, $t_k = n_1 + \dots n_{k-1} + (K-k+1)n_k$. For $i \in \{1, \dots, K\}$ and $a > 0$, let $B_{i,t}(a) = \widehat{X}_{i,T_i(t)} + \sqrt{\frac{a}{T_i(t)}}$ for $t \geq 1$.

Algorithm: For each phase $k = 0, 1, \dots, K-1$:
Let $\widehat{H}_k = K$ if $k = 0$, and otherwise $\widehat{H}_k = \max_{K-k+1 \leq i \leq K} i \widehat{\Delta}_{<i>,k}^{-2}$, where $\widehat{\Delta}_{i,k} = \left(\max_{1 \leq j \leq K} \widehat{X}_{j,T_j(t_k)} \right) - \widehat{X}_{i,T_i(t_k)}$ and $<i>$ is an ordering such that $\widehat{\Delta}_{<1>,k} \leq \dots \leq \widehat{\Delta}_{<K>,k}$.

For $t = t_k + 1, \dots, t_{k+1}$: Draw
$I_t \in \operatorname{argmax}_{i \in \{1, \dots, K\}} B_{i,t-1}(c\, n/\widehat{H}_k)$.

Adaptive UCB-E

Parameter: exploration rate $c > 0$.

Definitions: For $k \in \{1, \dots, K-1\}$, let $n_k = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \right\rceil$,
$t_0 = 0$, $t_1 = Kn_1$, and for $k > 1$,
$t_k = n_1 + \dots n_{k-1} + (K-k+1)n_k$. For $i \in \{1, \dots, K\}$ and $a > 0$, let $B_{i,t}(a) = \widehat{X}_{i,T_i(t)} + \sqrt{\frac{a}{T_i(t)}}$ for $t \geq 1$.

Algorithm: For each phase $k = 0, 1, \dots, K-1$:
Let $\widehat{H}_k = K$ if $k = 0$, and otherwise $\widehat{H}_k = \max_{K-k+1 \leq i \leq K} i \widehat{\Delta}^{-2}_{<i>,k}$,
where $\widehat{\Delta}_{i,k} = \left(\max_{1 \leq j \leq K} \widehat{X}_{j,T_j(t_k)} \right) - \widehat{X}_{i,T_i(t_k)}$ and $< i >$ is an ordering such that $\widehat{\Delta}_{<1>,k} \leq \dots \leq \widehat{\Delta}_{<K>,k}$.

For $t = t_k + 1, \dots, t_{k+1}$: Draw
$I_t \in \mathrm{argmax}_{i \in \{1,\dots,K\}} B_{i,t-1}(c\, n/\widehat{H}_k)$.

Adaptive UCB-E

Parameter: exploration rate $c > 0$.

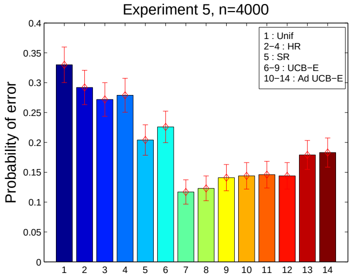
Definitions: For $k \in \{1, \dots, K-1\}$, let $n_k = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \right\rceil$, $t_0 = 0$, $t_1 = Kn_1$, and for $k > 1$, $t_k = n_1 + \dots n_{k-1} + (K-k+1)n_k$. For $i \in \{1, \dots, K\}$ and $a > 0$, let $B_{i,t}(a) = \widehat{X}_{i,T_i(t)} + \sqrt{\frac{a}{T_i(t)}}$ for $t \geq 1$.

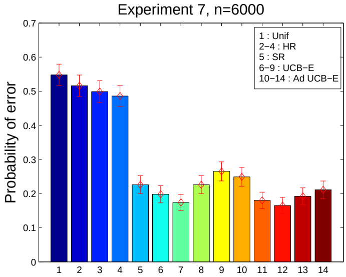
Algorithm: For each phase $k = 0, 1, \dots, K-1$:
Let $\widehat{H}_k = K$ if $k = 0$, and otherwise $\widehat{H}_k = \max_{K-k+1 \leq i \leq K} i \widehat{\Delta}^{-2}_{<i>,k}$, where $\widehat{\Delta}_{i,k} = \left(\max_{1 \leq j \leq K} \widehat{X}_{j,T_j(t_k)} \right) - \widehat{X}_{i,T_i(t_k)}$ and $< i >$ is an ordering such that $\widehat{\Delta}_{<1>,k} \leq \dots \leq \widehat{\Delta}_{<K>,k}$.

For $t = t_k + 1, \dots, t_{k+1}$: Draw
$I_t \in \mathrm{argmax}_{i \in \{1, \dots, K\}} B_{i,t-1}(c\, n / \widehat{H}_k)$.

Experiments

- Experiment 5: Arithmetic progression, $K = 15$, $\mu_i = 0.5 - 0.025i$, $i \in \{2, \ldots, 15\}$.
- Experiment 7: Three groups of bad arms, $K = 30$, $\mu_{2:6} = 0.45$, $\mu_{7:20} = 0.43$, $\mu_{21:30} = 0.38$.





Conclusion

- It requires at least $H/\log(K)$ rounds to find the best arm.
- SR is a parameter free algorithm, it requires less than $H\log(K)$ rounds to find the best arm.
- UCB-E requires only H rounds but also the knowledge of H to tune its parameter.
- Adaptive UCB-E does not have theoretical guarantees but it experimentally outperforms SR.