Bandits: Best Arms Identification

Classical Settings and methods

Mengyan Zhang

Australian National University; Data61, CSIRO

Currently @ MSRA Social Computing

Outline

- Definitions and Basic Settings
- Motivations and Applications
- Classical algorithms
  - Multi-armed bandits
    - UCB-based
    - Successive Rejects Type
    - Gap based – a unified algorithm
  - Black-box function optimization for many/continuous arms
  - Contextual bandits: Linear rewards

Multi-armed Bandits:
Sequential decision making

In each round t ϵ {1, .., N},

1. an agent selects an arm $A_t = i \; \epsilon \; \{1, .., K\}$ according policy π

2. then receive a reward $X_{i,T_i(t)}$ sampled from unknown distribution F_i

3. update estimations over distribution F_i based on historical observations

Multi-armed Bandits

Sequential decision making

In each round t ϵ {1, .., N},

For a given user

1. an agent selects an arm $A_t = i \; \epsilon \; \{1, .., K\}$ according policy π

Recommender system   Item (e.g. news)   Recommendation strategy

2. then receive a reward $X_{i,T_i(t)}$ sampled from unknown distribution F_i

CTR/click; non-click

3. update estimations over distribution F_i based on historical observations

Multi-Armed Bandits

Simple regret $r_t = \mu^* - \mu_{A_t}$ where $\mu^* = \max_{k \in \{1,\ldots,K\}} \mu_k$

Best Arm Identification

to recommend best arm(s)
at the end of exploration stage

Fixed Budget:

the number of round for exploration phase is fixed and known

Fixed Confidence:

the confidence level of quality of returned arms is fixed

Regret minimization: maximize the cumulative reward
(i.e. minimize cumulative regret)

Cumulative regret $R_T = \sum_{t=1}^{T} r_t$

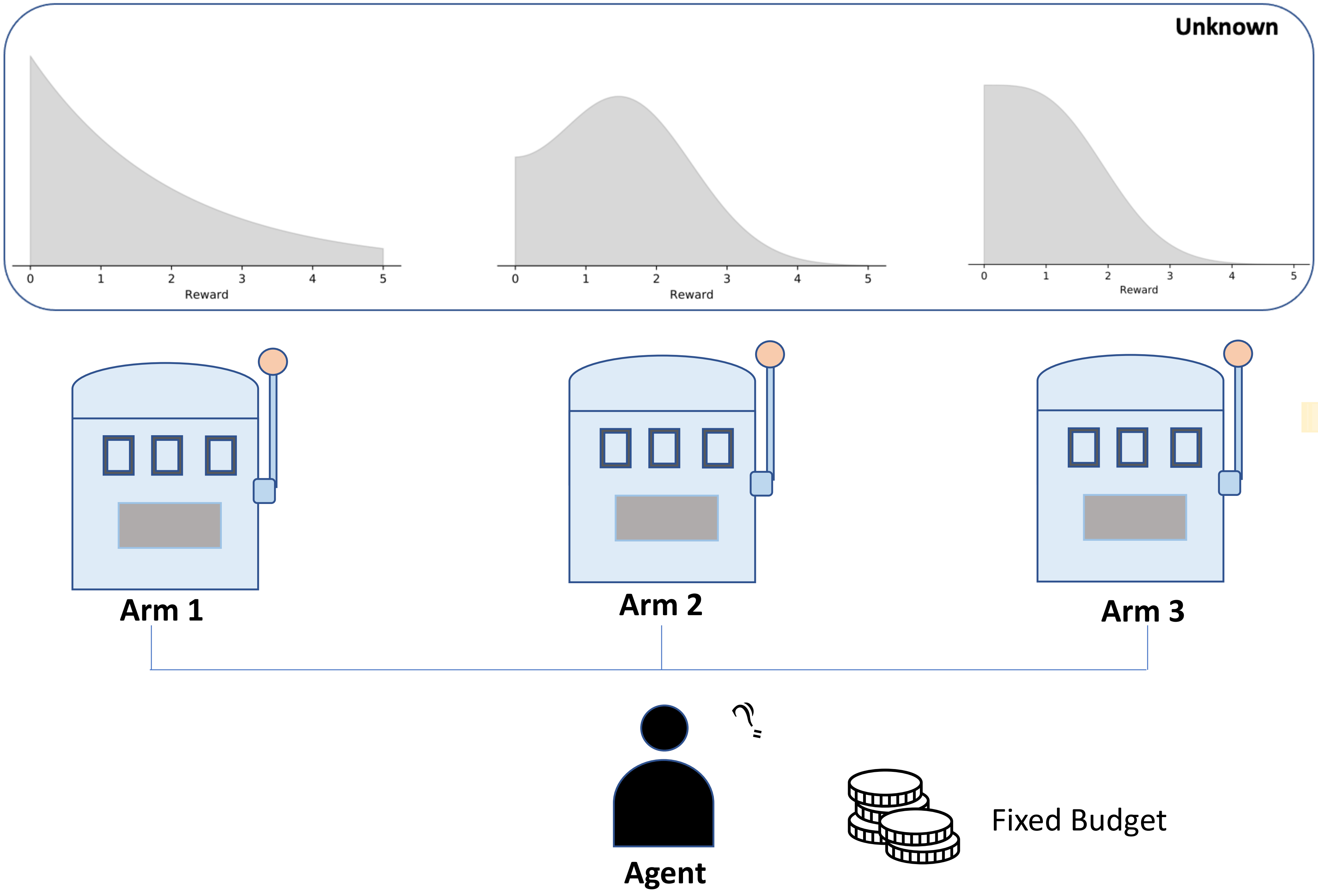
Best choice with
the current information

Other possibilities
have not been tried

Regret Minimization:
Exploitation vs. Exploration ?

Best Arm Identification:
How to allocate samples adaptively ?

Exploration Stage



Fixed Budget

or

Recommendation Stage

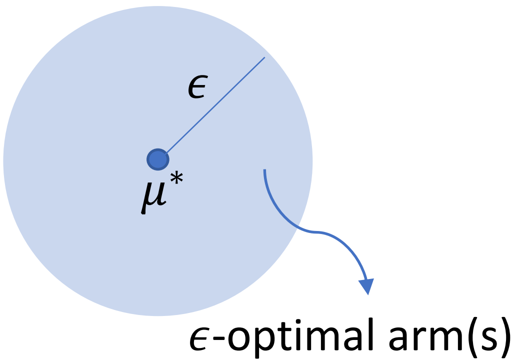
Best Arms?

Fixed Confidence

Recommend optimal arms with high probability

Best Arm Identification

to recommend best arm(s) at the end of exploration stage



Simple regret $\quad r_t = \mu^* - \mu_{A_t} \quad$ where $\quad \mu^* = \max_{k \in \{1,\ldots,K\}} \mu_k$

Fixed Budget: the number of round for exploration phase is fixed and known

-> To maximize the probability of returning the ϵ-optimal arm(s)

i.e. T is given, minimize Probability of error $\quad \delta_T = \mathrm{P}\left[r_T \geq \epsilon\right]$

δ_T and $\mathrm{E}[r_T]$ behave similarly

When $\epsilon = 0, \delta_T = \mathrm{P}\left[A_T \neq i^*\right]$

$$\mathrm{E}[r_T] = \sum_{i \neq i^*} \mathrm{P}[A_T = i]\Delta_i$$

Suboptimality gap $\quad \Delta_i = \mu^* - \mu_i$

$$\Delta_{min}\ \delta_T \leq \mathrm{E}[r_T] \leq \Delta_{\max}\ \delta_T$$

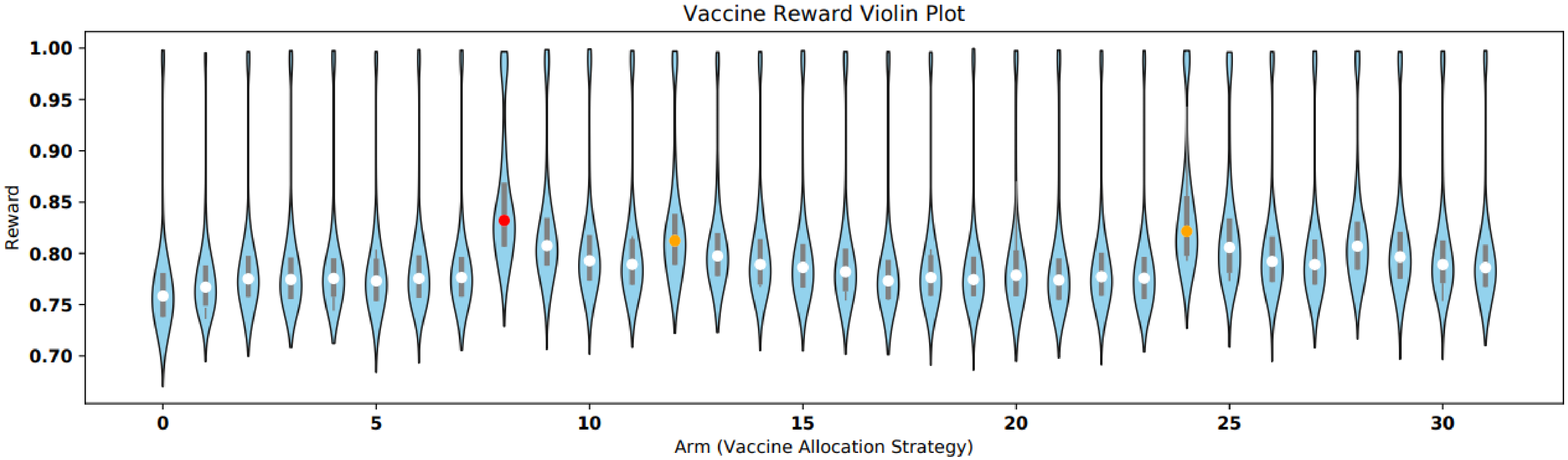
Fixed Confidence: the confidence level of quality of returned arms is fixed

-> To minimize the number of rounds needed

i.e. δ is given, minimize budget T s.t. $\quad \mathrm{P}\left[r_T \geq \epsilon\right] \leq \delta$

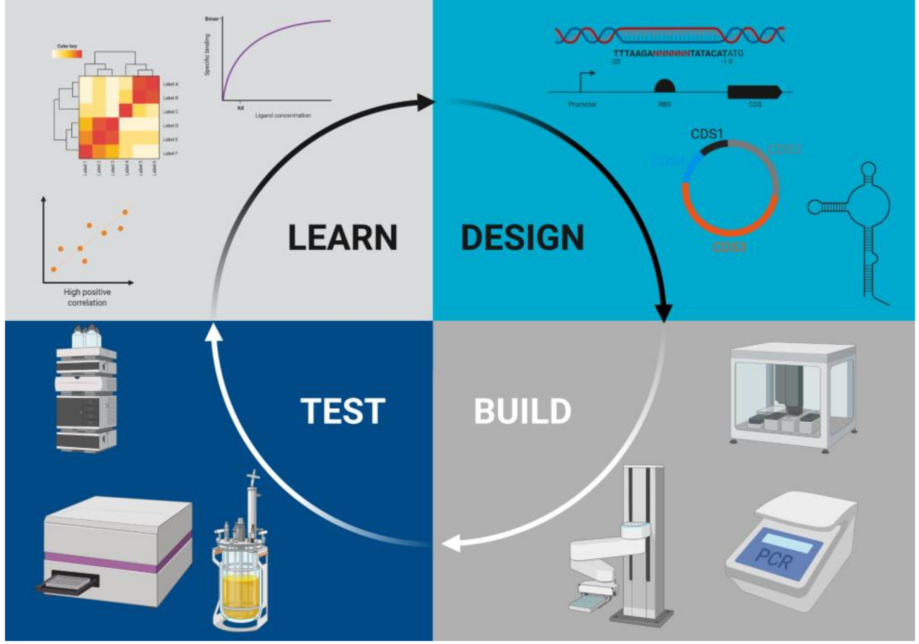
Applications: Vaccine testing

- Identify optimal strategies (highest mean/median reward) for allocation vaccines
- Arm: vaccine allocation strategy
- Reward: proportion of individuals that did not experience symptomatic infection



Zhang, M. ; Ong, C.S.. Quantile Bandits for Best Arms Identification. ICML2021.

Applications: Biological design

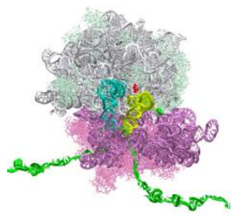


With fixed budget, design Ribosome Binding Site (RBS) sequences

↓

Optimize the protein expression level

Identify the DNA sequences with highest possible protein expression level



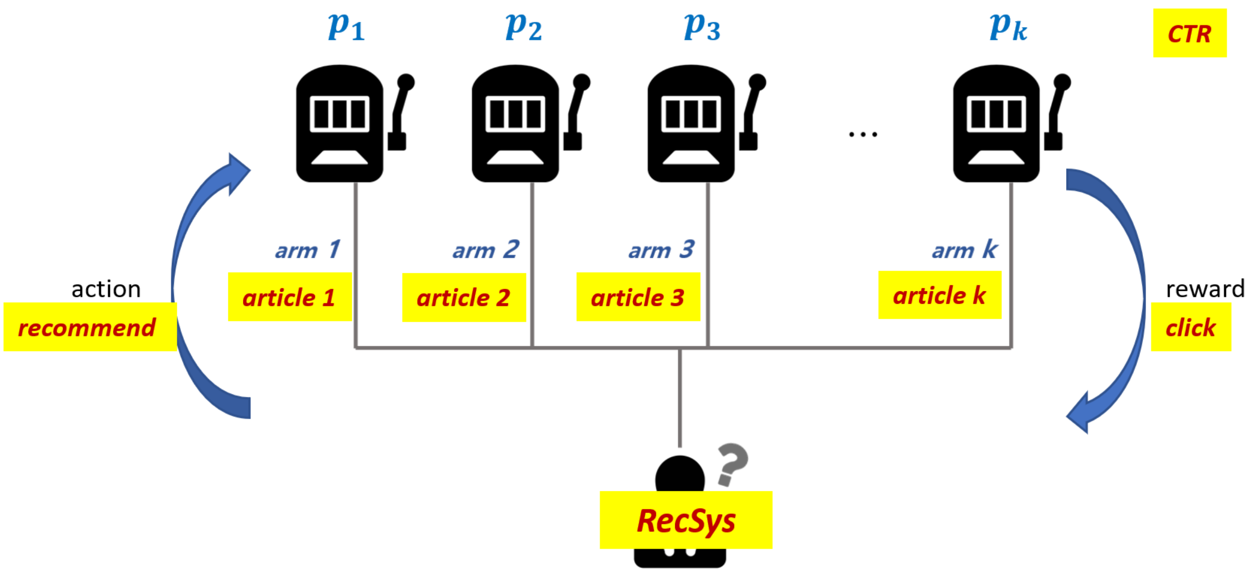
Arm: RBS sequence	Reward: Normalized* Protein Expression Level
TTTAAGAGTTATATATACAT	1.58
TTTAAGAATATGCTATACAT	1.42
TTTAAGACTCGGATATACAT	0.14
TTTAAGAGTTTTTTATACAT	2.88

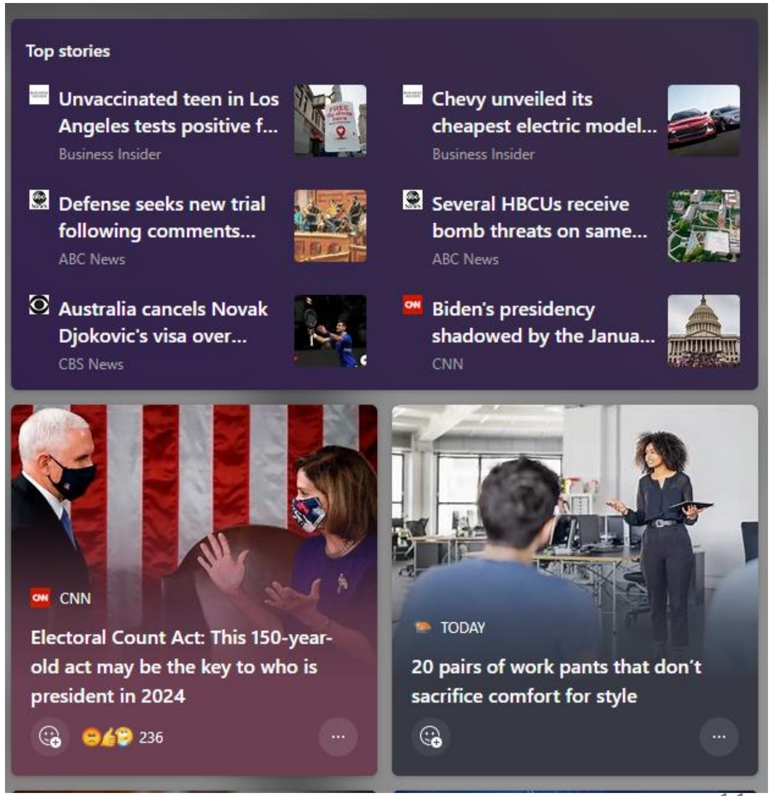
* zero mean and unit variance normalization $z = \frac{x-\mu}{\sigma}$

Zhang, M., and Ong, C. S. Opportunities and Challenges in Designing Genomic Sequences. ICML Workshop on Computational Biology 2021.

Applications: Recommendation System

- identify the most popular items (with potential high CTR) above some level of confidence using fewest possible samples
- Arm: item (e.g. news)
- Reward: click/ preference





Nan Li, Peng Hu, Wenming Xiao, Liangjie Zhang Exploration & Exploitation in Recommender System 2021 Slides

Outline

- Definitions and Basic Settings
- Motivations and Applications
- Classical algorithms
  - Multi-armed bandits
    - UCB-based
    - Successive Rejects Type
    - Gap based – a unified algorithm
  - Black-box function optimization for many/continuous arms
  - Contextual bandits: Linear rewards

Best Arm Identification in Multi-Armed Bandits

- K arms associated with K unknown reward distributions respectively (assume rewards are in [0,1] and there is a unique optimal arm),
- T rounds/budget,
- The agent select one arm A_t according to policy π -> observe reward drawn from v_{A_t} idependently from the past (actions and observations)

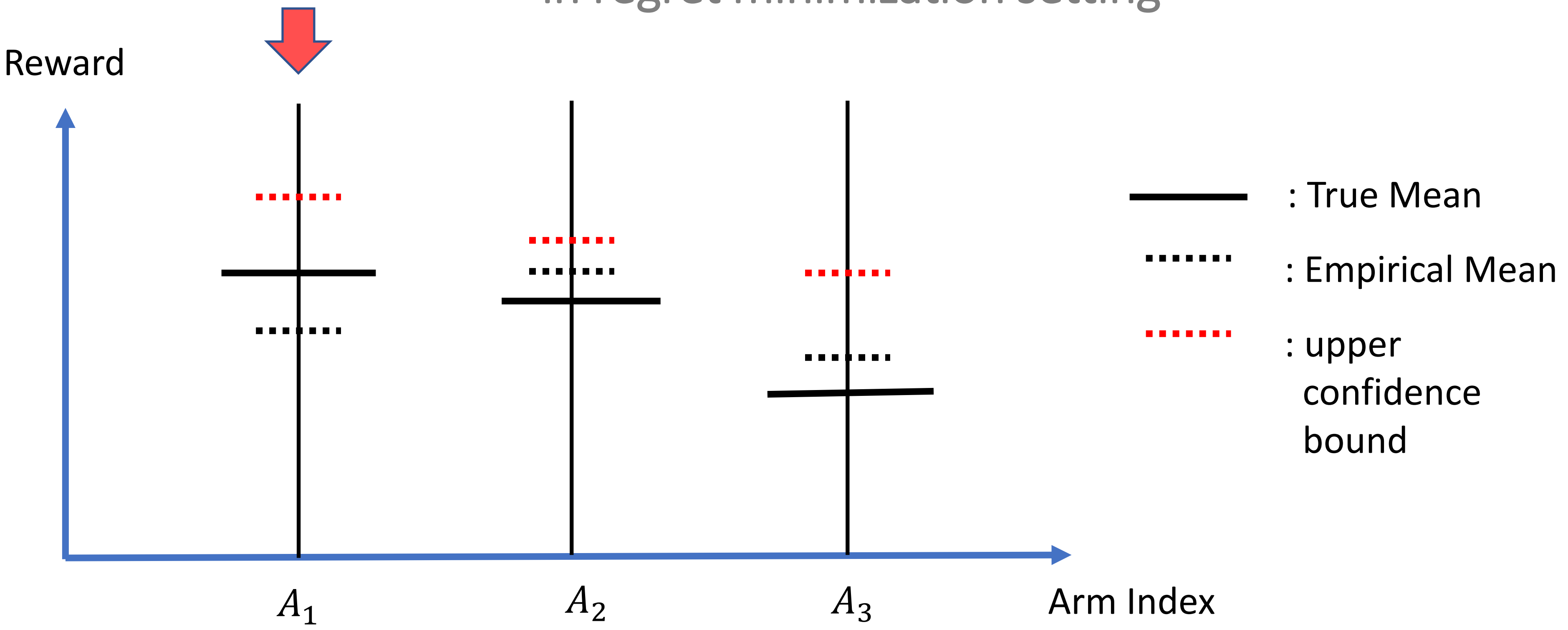
Simple regret $r_T = \mu^* - \mu_{A_T}$ where $\mu^* = \max_{k \in \{1,\ldots,K\}} \mu_k$

Probability or error $\delta_T = \mathrm{P}\left[A_T \neq i^*\right]$

Audibert, Jean-Yves, and Sébastien Bubeck. Best Arm Identification in Multi-Armed Bandits. COLT2010.

Upper Confidence Bound (UCB)

in regret minimization setting



$$\underset{i\,\in\kappa}{\operatorname{argmax}}\ \text{Central Tendency} + \beta\ \text{Confidence Width}$$

Exploitation

Exploration

e.g. Empirical Mean

UCB-E (Upper Confidence Bound Exploration) algorithm

$$\mathrm{argmax}_{i\in\{1,\ldots,K\}} \quad B_{i,s} = \widehat{X}_{i,s} + \sqrt{\frac{a}{s}}$$

Number of draws

Regret minimisation -> a ~ O(log T)
BAI, need more exploration -> a ~ O(T)

Suboptimality gap $\Delta_i = \mu^* - \mu_i$

Hardness of the task

$$H_1 = \sum_{i=1}^{K} \frac{1}{\Delta_i^2} \quad \text{and} \quad H_2 = \max_{i\in\{1,\ldots,K\}} i\Delta_{(i)}^{-2}.$$

Probability or error $\delta_T \sim \exp(-c\,T/H_1)$

UCB-E: $a = \frac{25}{36}\,(T-K)/H_1$

H_1 is unknown and hard to be online estiamted!

Audibert, Jean-Yves, and Sébastien Bubeck. Best Arm Identification in Multi-Armed Bandits. COLT2010.

Successive Rejects

$$\overline{log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$$

$$n_k = \left\lceil \frac{1}{\overline{log}(K)} \frac{T-K}{K+1-k} \right\rceil, \mathrm{n}_0 = 0$$

- Devide budget T into K-1 phases

$n_k \ - n_{k-1}$

- In each phase k, pull equally often each arm which has one been rejected yet
- At the end of each phase, reject the arm with lowest empirical mean
- Recommend the last surviving arm

Suboptimality gap $\Delta_i = \mu^* - \mu_i$

Hardness of the task

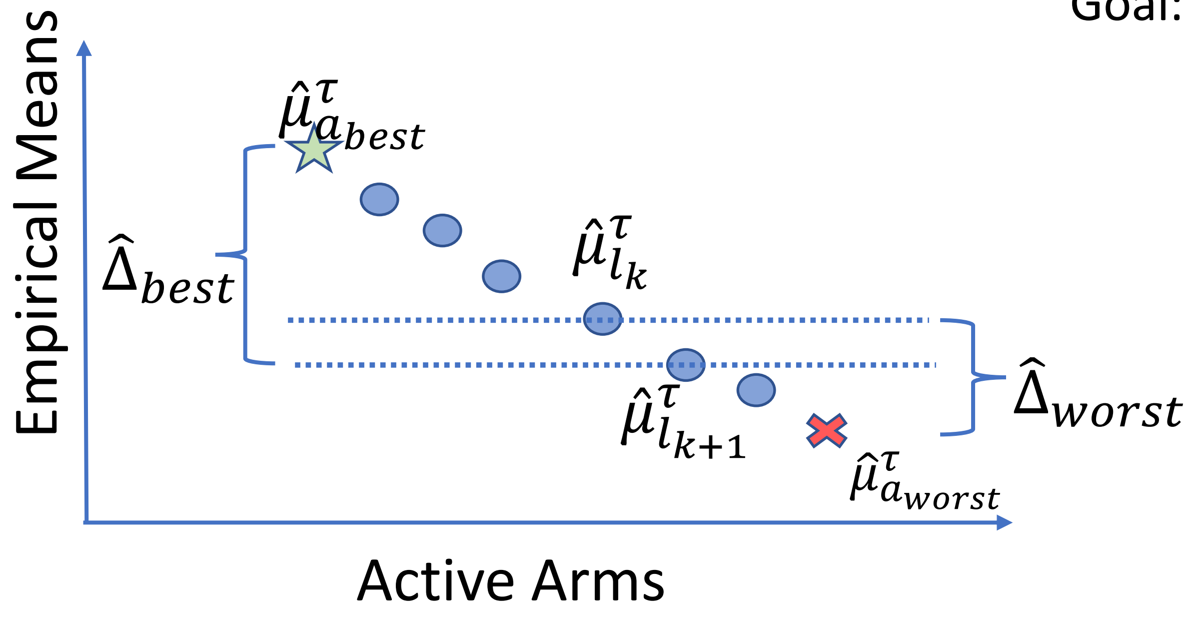
$$H_1 = \sum_{i=1}^{K} \frac{1}{\Delta_i^2} \quad \text{and} \quad H_2 = \max_{i \in \{1,\ldots,K\}} i\Delta_{(i)}^{-2}.$$

$$H_2 \leq H_1 \leq \log(2K) H_2.$$

Probability or error $\delta_T \sim \exp(-c\, T/H_2)$

Audibert, Jean-Yves, and Sébastien Bubeck. Best Arm Identification in Multi-Armed Bandits. COLT2010.

Multi-arm Identification – Successive Accepts and Rejects



Goal: recommend m >= 1 arms with fixed budget T

l_k: remaining number of arms to be found in current phase k

$\hat{\Delta}_{best} > \hat{\Delta}_{worst}$: Accept

$\hat{\Delta}_{best} \leq \hat{\Delta}_{worst}$: Reject

$$\Delta_i^{\langle m\rangle} = \begin{cases} \mu_i - \mu_{m+1} & \text{if } i \leq m \\ \mu_m - \mu_i & \text{if } i > m \end{cases},$$

$$H_1^{\langle m\rangle} = \sum_{i=1}^{K} \frac{1}{\left(\Delta_i^{\langle m\rangle}\right)^2},$$

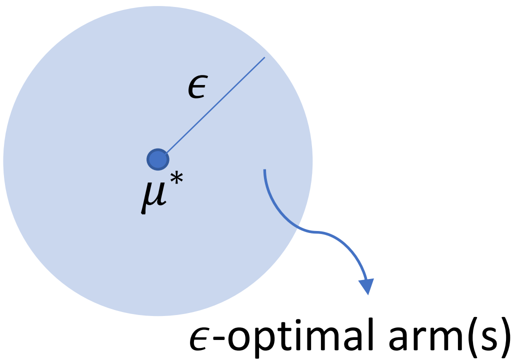
$$H_2^{\langle m\rangle} = \max_{i \in \{1,\ldots,K\}} i \left(\Delta_{(i)}^{\langle m\rangle}\right)^{-2},$$

Probability or error $\delta_T \sim \exp(-c\, T / H_2^{<m>})$

Bubeck, Sébastien, Tengyao Wang, and Nitin Viswanathan. Multiple Identifications in Multi-Armed Bandits. ICML2013.

Best Arm Identification

to recommend best arm(s) at the end of exploration stage



$$\boxed{\text{Simple regret} \quad r_t = \mu^* - \mu_{A_t} \quad \text{where} \quad \mu^* = \max_{k \in \{1,\ldots,K\}} \mu_k}$$

Fixed Budget: the number of round for exploration phase is fixed and known

-> To maximize the probability of returning the ϵ-optimal arm(s)

i.e. T is given, minimize Probability of error $\delta_T = \mathrm{P}\,[r_T \geq \epsilon]$

δ_T and $\mathrm{E}[r_T]$ behave similarly

When $\epsilon = 0, \delta_T = \mathrm{P}\,[A_T \neq i^*]$

$$\mathrm{E}[r_T] = \sum_{i \neq i^{\wedge} *} \mathrm{P}[A_T = i]\Delta_i$$

Suboptimality gap $\Delta_i = \mu^* - \mu_i$

$$\Delta_{min}\ \delta_T \leq \mathrm{E}[r_T] \leq \Delta_{\max}\ \delta_T$$

Fixed Confidence: the confidence level of quality of returned arms is fixed

-> To minimize the number of rounds needed

i.e. δ is given, minimize budget T s.t. $\mathrm{P}\,[r_T \geq \epsilon] \leq \delta$

Unified Methods: UGapE

SELECT-ARM (t)

Compute $B_k(t)$ for each arm $k \in A$

Identify the set of m arms $J(t) \in \arg \overset{1..m}{\min_{k \in A}} B_k(t)$

Pull the arm $I(t) = \arg\max_{k \in \{l_t, u_t\}} \beta_k(t-1)$

Observe $X_{I(t)}\big(T_{I(t)}(t-1)+1\big) \sim \nu_{I(t)}$

Update $\widehat{\mu}_{I(t)}(t)$ and $T_{I(t)}(t)$

High probability upper and lower bounds on the mean of arm k

$$\begin{cases} U_k(t) = \widehat{\mu}_k(t-1) + \beta_k(t-1) \\ L_k(t) = \widehat{\mu}_k(t-1) - \beta_k(t-1). \end{cases}$$

Upper bound on simple regret $B_k(t) = \overset{m}{\max_{i \neq k}} U_i(t) - L_k(t)$

Best possible arm left outside of J(t); worst possible arm among those in J(t);

$u_t = \arg\max_{j \notin J(t)} U_j(t)$ and $l_t = \arg\min_{i \in J(t)} L_i(t)$.

Represents: how bad the choice of J(t) could be

Fixed budget

UGapEb: $\beta_k(t-1) = b\sqrt{\frac{a}{T_k(t-1)}}$

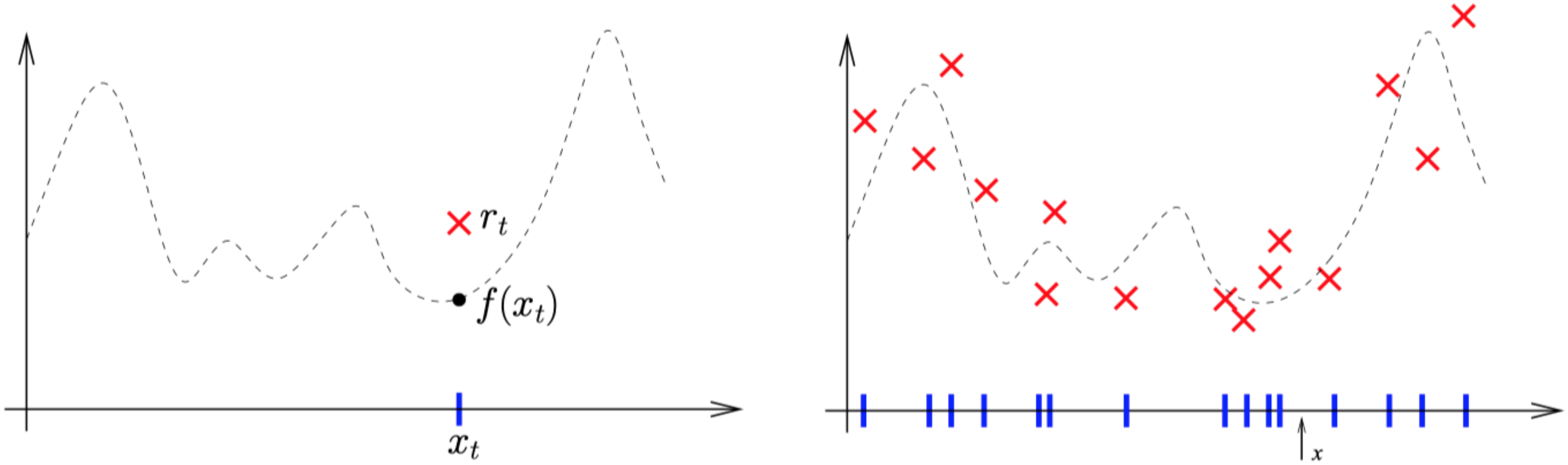
Fixed confidence

UGapEc: $\beta_k(t-1) = b\sqrt{\frac{c\log\frac{4K(t-1)^3}{\delta}}{T_k(t-1)}}$

Gabillon, Victor, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence. In *Advances in Neural Information Processing Systems 25.* 2012.

Black-box function optimisation

Optimism in the face of uncertainty for black-box optimisation of a function f given stochastic evaluation of the function.



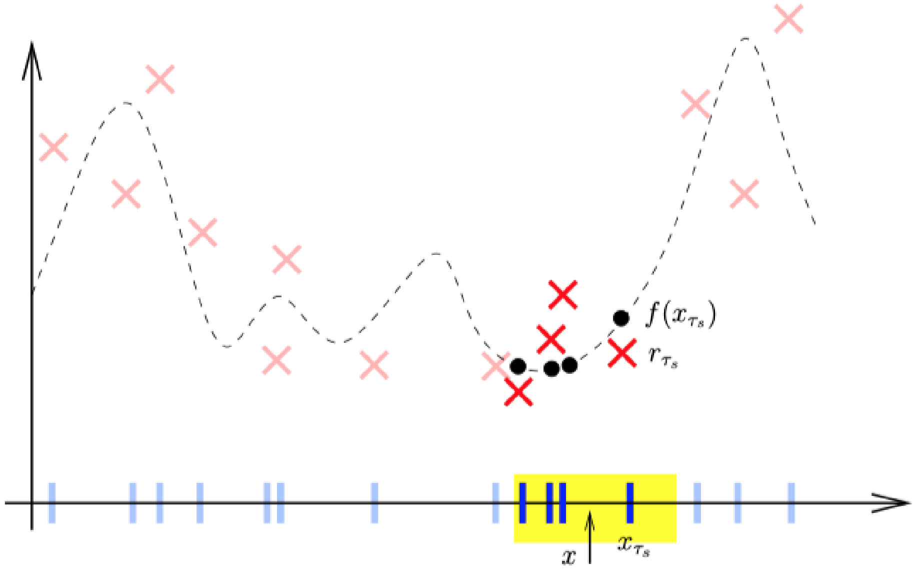
$$y_t = f(x_t) + \epsilon_t$$

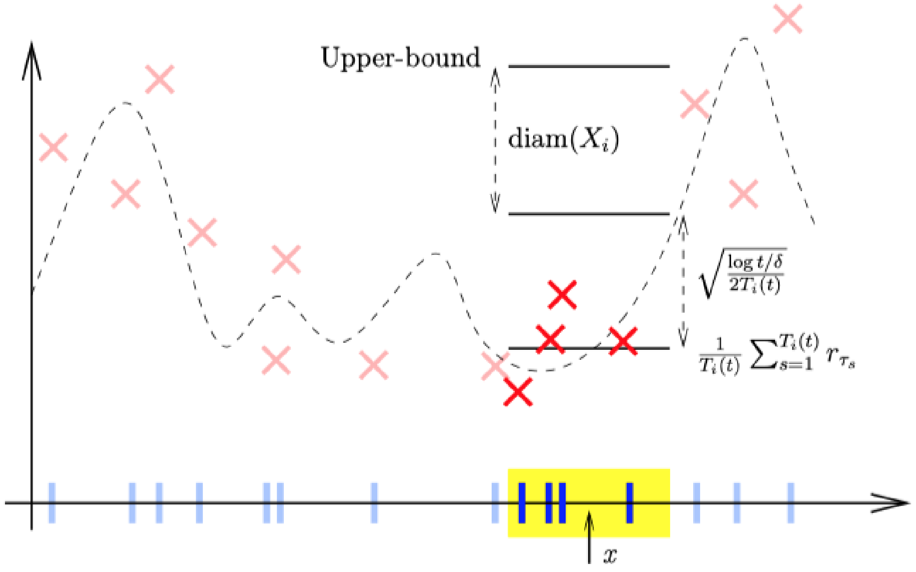
$$E[\epsilon_t | x_t] = 0$$

Goal: Search for a good approximation of
the maximum of a function $\chi \rightarrow \mathbb{R}$
with a fixed number of function evaluations

Munos, Rémi. “From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning,” 2014.

Optimism in the face of uncertainty





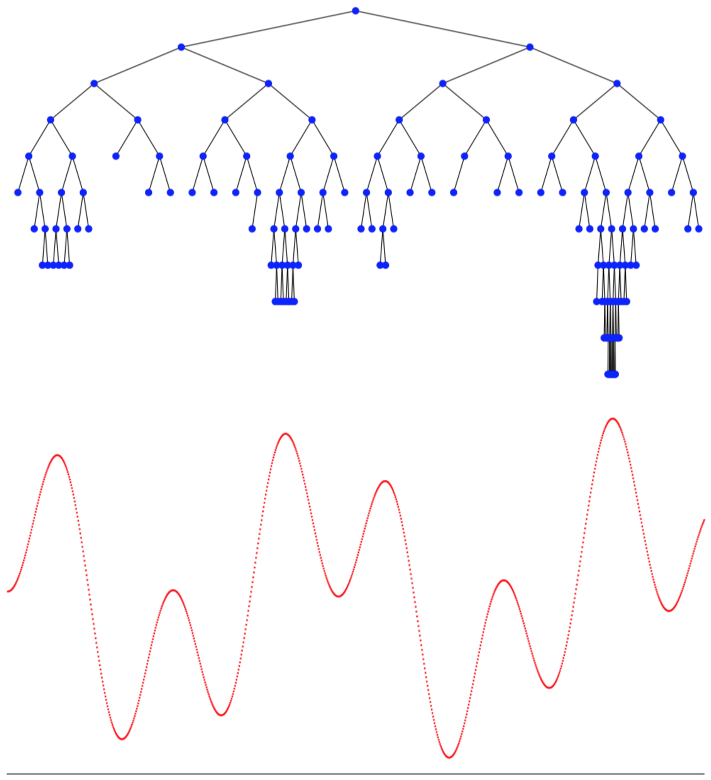
Given a subset $X_i \in \chi$

$$B_{t,T_i(t)}(X_i) \stackrel{\text{def}}{=} \frac{1}{T_i(t)} \sum_{s=1}^{T_i(t)} r_{\tau_s} + \sqrt{\frac{\log t/\eta}{2T_i(t)}} + \mathrm{diam}(X_i) \geq \max_{x \in X_i} f(x).$$

A trade-off in the choice of size of X_i

Munos, Rémi. “From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning,” 2014.

For many arms or continuous arms: Hierarchical Partition



For each round: $$b_{h,j}(t) \stackrel{\text{def}}{=} \hat{\mu}_{h,j}(t) + \sqrt{\frac{\log(n^2/\eta)}{2T_{h,j}(t)}} + \delta(h),$$

- Select the leaf node with highest <u>b-value</u>
- sample the center point and collect rewards
- Expand node: when # draws > some <u>threshold</u>

(depends on T, smoothness of f)

Extend: model f as a sample from Guassian process, e.g. GPOO

Munos, Rémi. “From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning,” 2014.
Mengyan Zhang, Russell Tsuchida, and Cheng Soon Ong. “Gaussian Process Bandits with Aggregated Feedback”. In: 36th AAAI Conference on Artificial Intelligence (2022).

Contextual BAI with linear rewards – Generalized Successive Elimination

$$y_i = f(x_i) + \epsilon$$

Linear model $\hat{\mu}_{i,t} = x_i^T \hat{\theta}_t$

Generalized linear model $\hat{\mu}_{i,t} = g(x_i^T \hat{\theta}_t)$

e.g. logistic regression $\mathrm{g}(x) = (1 + \exp(-x))^{-1}$

Repeat

Explore: split budget evenly to each stage, then allocate the budget in each stage according to some allocation policy

Estimate: estimate each arm's mean $\hat{\mu}_{i,t}$

Elimination: sort arms in descending order of $\hat{\mu}_{i,t}$, keep the first $1/\eta$ of them while eliminating the rest

Until only one arm is remained

Uniform allocation

Optimal allocation

G-optimal design: minimize the maximum variance uniformly along all directions

Azizi, MohammadJavad, Branislav Kveton, and Mohammad Ghavamzadeh. "Fixed-Budget Best-Arm Identification in Contextual Bandits: A Static-Adaptive Algorithm." *ArXiv:2106.04763 [Cs]*, June 9, 2021.

Conclusion

Applications
- Vaccine allocation
- DNA design
- Recommendation system

Best Arm Identification
- Fixed Budget:
  - Multi-armed bandits
    - UCB-based
    - Successive Rejects Type
    - Gap based – a unified algorithm
  - Black-box function optimization
  - Contextual bandits: Linear rewards
- Fixed Confidence

Regret minimization: maximize the cumulative reward
(i.e. minimize cumulative regret)

Collections of classical papers https://homepages.cwi.nl/~wmkoolen/PureExploration18/

References

- Even-Dar, Eyal, Shie Mannor, and Yishay Mansour. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. 2006.
- Audibert, Jean-Yves, and Sébastien Bubeck. Best Arm Identification in Multi-Armed Bandits. COLT2010.
- Gabillon, Victor, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence. In *Advances in Neural Information Processing Systems 25.* 2012.
- Bubeck, Sébastien, Tengyao Wang, and Nitin Viswanathan. Multiple Identifications in Multi-Armed Bandits. ICML2013.
- Munos, Rémi. “From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning,” 2014.
- **Mengyan Zhang**, Russell Tsuchida, and Cheng Soon Ong.. Opportunities and Challenges in Designing Genomic Sequences. ICML Workshop on Computational Biology 2021.
- **Mengyan Zhang**, and Cheng Soon Ong. Quantile Bandits for Best Arms Identification. ICML2021.
- Azizi, MohammadJavad, Branislav Kveton, and Mohammad Ghavamzadeh. “Fixed-Budget Best-Arm Identification in Contextual Bandits: A Static-Adaptive Algorithm.” *ArXiv:2106.04763 [Cs]*, June 9, 2021.
- **Mengyan Zhang**, Russell Tsuchida, and Cheng Soon Ong. Gaussian Process Bandits with Aggregated Feedback. AAAI2022.