Non-Parametric Exploration in Multi-Armed Bandits

Emilie Kaufmann,

based on joint works with
Dorian Baudry and Odalric-Ambrym Maillard

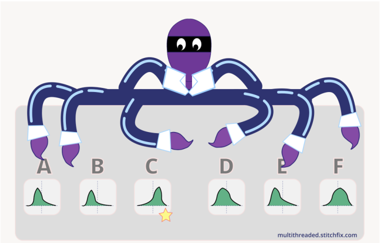
Université de Lille

Inria

Banff Workshop
December 1st, 2021

The stochastic MAB model

- K *unknown* reward distributions $\nu_1, \ldots, \nu_K$ called *arms*
- a each time t, select an arm A_t and observe a reward $X_t \sim \nu_{A_t}$



Objective: find a sequential sampling strategy $\mathcal{A} = (A_t)$ that maximizes the sum of rewards $\Leftrightarrow$ minimize the *regret*

$$\mathcal{R}_T(\mathcal{A}) = \mu^\star T - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$$

[Robbins, 52] [Lattimore and Csepesvari 20]

Outline

(Don't) Follow The Learder

A very simple algorithm exploiting the current knowledge:

$$A_{t+1} = \underset{a \in [K]}{\arg\max} \ \hat{\mu}_a(t)$$

where

- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$ is the number of selections of arm a
- $\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a)$ is the empirical mean of the rewards collected from arm a

(Don't) Follow The Lerarder

A very simple algorithm exploiting the current knowledge:

$$A_{t+1} = \underset{a \in [K]}{\arg\max}\ \hat{\mu}_a(t)$$

where

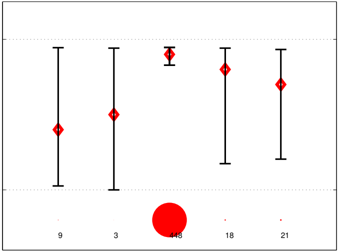
- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$ is the number of selections of arm a
- $\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a)$ is the empirical mean of the rewards collected from arm a

Properties:

- a simple, non-parametric algorithm
- achieves linear regret
- ➜ need for an exploration/exploitation trade-off

Smarter algorithms: Two dominant families

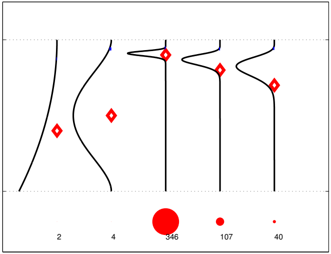
Upper Confidence Bound
(UCB)



$$A_{t+1} = \underset{a \in [K]}{\text{argmax}} \ \mathrm{UCB}_a(t)$$

where $\mathrm{UCB}_a(t)$ is an UCB on the unknown mean μ_a

Thompson Sampling
(TS)

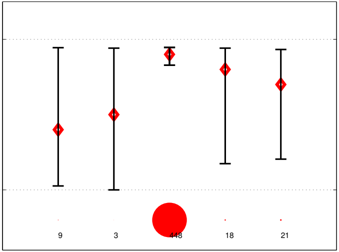


$$A_{t+1} = \underset{a \in [K]}{\text{argmax}} \ \tilde{\mu}_a(t)$$

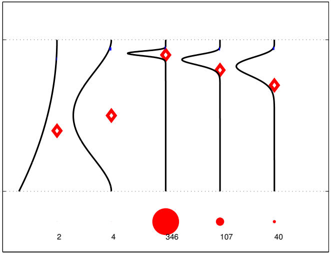
where $\tilde{\mu}_a(t)$ is a sample from a posterior distribution on μ_a

Smarter algorithms: Two dominant families

Upper Confidence Bound
(UCB)



Thompson Sampling
(TS)



$$A_{t+1} = \underset{a\in[K]}{\text{argmax}}\ \mathrm{UCB}_a(t)$$

$$A_{t+1} = \underset{a\in[K]}{\text{argmax}}\ \tilde{\mu}_a(t)$$

where $\mathrm{UCB}_a(t)$ is an UCB on the unknown mean μ_a

where $\tilde{\mu}_a(t)$ is a sample from a posterior distribution on μ_a

➜ both approaches can be tuned to achieve *optimality*

(Problem dependent, asymptotic) optimality

[Lai and Robbins 1985]: for simple* parametric arms distributions

$$\mathcal{R}_T(\mathcal{A}) \geq \left(\sum_{a:\mu_a<\mu_\star} \frac{\mu_\star - \mu_a}{\mathrm{kl}(\mu_a, \mu_\star)} \right) \log(T)$$

for T large enough.

Observation: UCB and TS need to know which distributions they are facing in order to match the lower bound

Wanted: a single algorithm that can be simultaneously asymptotically optimal for different classes of distributions

* distribution continuously parameterized by their means, typically one-parameter exponential family (Bernoulli, Gaussian with known variances, Poisson...)

Outline

Recent work on non-parameteric methods

- Perturbed History Exploration [Kveton et al. 19]
  - ➜ standard non-parametric bootstrap does not work
  - ➜ a fix by *adding fake samples in the history* of rewards
  - ➜ logarithmic regret for bounded distribution (not optimal)
- Non Parametric Thompson Sampling [Riou and Honda 20]
  - ➜ instead of the empirical mean, compute a *random reweighting of the history* (+ an upper bound on the support)
  - ➜ optimal regret for bounded distribution

Recent work on non-paramateric methods

- Perturbed History Exploration [Kveton et al. 19]
  - ➜ standard non-parametric bootstrap does not work
  - ➜ a fix by *adding fake samples in the history* of rewards
  - ➜ logarithmic regret for bounded distribution (not optimal)
- Non Parametric Thompson Sampling [Riou and Honda 20]
  - ➜ instead of the empirical mean, compute a *random reweighting of the history* (+ an upper bound on the support)
  - ➜ optimal regret for bounded distribution

From re-sampling to sub-sampling

[Baransi et al. 14], [Chan 20]

Subsampling Duelling Algorithms

A *round-based* approach

1. Find the *leader*: arm with largest number of observations
2. Organize $K - 1$ *duels*: *leader* vs *challengers*.
3. Draw a set of arms: *winning challengers* xor *leader*.

Subsampling Duelling Algorithms

A *round-based* approach

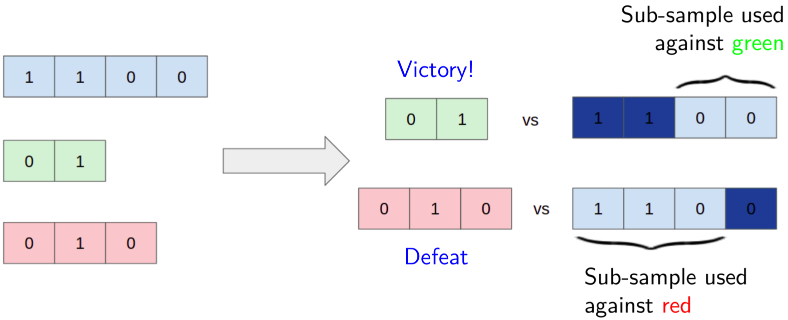
1. Find the *leader*: arm with largest number of observations
2. Organize $K-1$ *duels*: *leader* vs *challengers*.
3. Draw a set of arms: *winning challengers* xor *leader*.

How do duels work?

Idea: a *fair comparison* of two arms with different history size

- challenger: compute $\hat{\mu}_c$, the empirical mean
- leader: compute $\tilde{\mu}_\ell$, the mean of a *sub-sample* of the same size as the history of the challenger.
- challenger wins if $\hat{\mu}_c \geq \tilde{\mu}_\ell$

Illustration of a round



In this example the leader is *blue*: *green* wins against *blue*, *red* loses
$\Rightarrow$ only *green* is drawn at the end of the round.

Possible Sub-Sampling Schemes

Input of SDA: how to sub-sample n elements from N?

- Sampling Without Replacement (SW-SDA): pick a random subset of size n in $[1, N]$

  (as in BESA [Baransi et al. 14], analyzed for 2 arms)

- Random-Block Sampling (RB-SDA): return a block of size n starting from random $n_0 \sim \mathcal{U}([1, N-n])$

7.6	-4	0.7	1.4	3.1	0.1	-1.2

- Last Block Sampling (LB-SDA): return $\{N-n, \ldots, N\}$

7.6	-4	0.7	1.4	3.1	0.1	-1.2

Remark: SSMC [Chan 20] uses data-dependent sub-sampling

Outline

Regret of SDA algorithms

SDA algorithms are round-based

- $\mathcal{A}_r$: set of arms that are sampled in round r
- r_T (random) number of rounds before T samples are collected

$$
\begin{aligned}
\mathcal{R}_T(\mathcal{A}) &= \mathbb{E}\left[\sum_{t=1}^{T}(\mu_\star - \mu_{A_t})\right] \leq \mathbb{E}\left[\sum_{s=1}^{r_T}\sum_{k=1}^{K}(\mu_\star - \mu_k)\mathbb{1}(k \in \mathcal{A}_s)\right] \\
&\leq \mathbb{E}\left[\sum_{s=1}^{T}\sum_{k=1}^{K}(\mu_\star - \mu_k)\mathbb{1}(k \in \mathcal{A}_s)\right] \\
&= \sum_{k=1}^{K}(\mu_\star - \mu_k)\mathbb{E}\left[N_k(T)\right]
\end{aligned}
$$

$N_k(t) = \sum_{s=1}^{t} \mathbb{1}(k \in \mathcal{A}_s)$: number of draws of k in t rounds

First ingredient: Concentration

- $Y_{k,n}$: n-th observation from arm k
- $\overline{Y}_{k,\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} Y_{k,i}$ for $\mathcal{S} \subseteq [m]$
- $\mathcal{S}_k^r(m,n) \subseteq [m]$ sub-sample used in round r if arm k is the challenger and $N_k(r) = n$, with $n \leq m$

Definition (Block Sampler)

A *block sampler* always outputs a sequence of *consecutive observations* in the rewards history.

$\hookrightarrow$ Random Block and Last Block are block samplers, not SWR.

Lemma (concentration of a sub-sample)

Let $s \leq r$ and $\mathcal{M}_s = \{n_0 \leq N_b(s) \leq N_a(s) \leq r\}$. Under a block sampler, for any $\xi \in (\mu_a, \mu_b)$ it holds that

$$\sum_{s=1}^{r} \mathbb{P}\left(\bar{Y}_{a,N_a(s)} \geq \bar{Y}_{b,\mathcal{S}_b^s(N_b(s),N_a(s))}, \mathcal{M}_s\right) \leq \sum_{j=n_0}^{r} \mathbb{P}\left(\bar{Y}_{a,j} \geq \xi\right) + r\sum_{j=n_0}^{r} \mathbb{P}\left(\bar{Y}_{b,j} \leq \xi\right)$$

First ingredient: Concentration

Assumption 1: (*arm concentration*)

$$\begin{aligned} \forall x > \mu_k, \quad \mathbb{P}\left(\bar{Y}_{k,n} \geq x\right) &\leq e^{-nI_k(x)} \\ \forall x < \mu_k, \quad \mathbb{P}\left(\bar{Y}_{k,n} \leq x\right) &\leq e^{-nI_k(x)} . \end{aligned}$$

for some rate function $I_k(x)$ (1-d exp. families: $I_k(x) = \mathrm{kl}(x, \mu_k)$)

Lemma (for SDA using a block sampler)

Under Assumption 1, for every $\varepsilon > 0$, there exists a constant $C_k(\boldsymbol{\nu}, \epsilon)$ with $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)$ such that

$$\mathbb{E}[N_k(T)] \leq \frac{1+\epsilon}{I_1(\mu_k)} \log(T) + 32 \sum_{r=1}^{T} \mathbb{P}\left(N_1(r) \leq (\log(r))^2\right) + C_k(\boldsymbol{\nu}, \epsilon)$$

Proof: exploits only concentration (and how the algorithm works)

First ingredient: Concentration

Assumption 1: (*arm concentration*)

$$\begin{aligned} \forall x > \mu_k, \quad \mathbb{P}\left(\bar{Y}_{k,n} \geq x\right) & \leq & e^{-nI_k(x)} \\ \forall x < \mu_k, \quad \mathbb{P}\left(\bar{Y}_{k,n} \leq x\right) & \leq & e^{-nI_k(x)} \ . \end{aligned}$$

for some rate function $I_k(x)$ (1-d exp. families: $I_k(x) = \mathrm{kl}(x, \mu_k)$)

Lemma (for SDA using a block sampler)

Under Assumption 1, for every $\varepsilon > 0$, there exists a constant $C_k(\boldsymbol{\nu}, \epsilon)$ with $\boldsymbol{\nu} = (\nu_1, \dots, \nu_k)$ such that

$$\mathbb{E}[N_k(T)] \leq \frac{1+\epsilon}{I_1(\mu_k)} \log(T) + 32 \sum_{r=1}^{T} \mathbb{P}\left(N_1(r) \leq (\log(r))^2\right) + C_k(\boldsymbol{\nu}, \epsilon)$$

Proof: exploits only concentration (and how the algorithm works)

Two extra ingredients

To upper bound $\sum_{r=1}^{T} \mathbb{P}\left(N_1(r) \leq (\log(r))^2\right)$, we further need:

① Diversity: the sub-sampler produces a variety of *independent* sub-samples when being called a lot of time

$X_{m,H,j}$:= number of mutually non-overlapping sets when we draw m sub-samples of size j in a history of size H. Under RB sampling,

$$\sum_{r=1}^{T} \sum_{j=1}^{(\log r)^2} \mathbb{P}\left(X_{N_r,N_r,j} < \gamma \frac{r}{(\log r)^2}\right) = o(\log T).$$

for $N_r = O(r/\log^2(r))$ and some $\gamma \in (0,1)$

Two extra ingredients

To upper bound $\sum_{r=1}^{T} \mathbb{P}\left(N_1(r) \leq (\log(r))^2\right)$, we further need:

② a Balance condition: the optimal arm (arm 1) is not likely to loose many duels based on *independent* sub-samples

Introducing the balance function of arm k of cdf F_K,

$$\alpha_k(M,j) := \mathbb{E}_{X\sim\nu_{1,j}}\left[(1 - F_{\nu_{k,j}}(X))^M\right]$$

we need, that each arm $k \neq 1$ satisfy the balance condition :

$$\forall \beta \in (0,1), \quad \sum_{t=1}^{T} \sum_{j=1}^{\lfloor (\log t)^2 \rfloor} \alpha_k\left(\left\lfloor \beta t/(\log t)^2 \right\rfloor, j\right) = o(\log T)\ .$$

➜ an assumption on the arms' distributions

Two extra ingredients

To upper bound $\sum_{r=1}^{T} \mathbb{P}\left(N_1(r) \leq (\log(r))^2\right)$, we further need:

② a Balance condition: the optimal arm (arm 1) is not likely to loose many duels based on *independent* sub-samples

Introducing the balance function of arm k of cdf F_K,

$$\alpha_k(M, j) := \mathbb{E}_{X \sim \nu_{1,j}} \left[(1 - F_{\nu_{k,j}}(X))^M \right]$$

we need, that each arm $k \neq 1$ satisfy the balance condition :

$$\forall \beta \in (0,1), \quad \sum_{t=1}^{T} \sum_{j=f_t^*}^{\lfloor (\log t)^2 \rfloor} \alpha_k\left(\left\lfloor \beta t / (\log t)^2 \right\rfloor, j\right) = o(\log T) \ .$$

(* relaxed balance condition if the algorithm adds forced exploration of level f_r)

➜ an assumption on the arms' distributions

Final results

General Theorem [Baudry et al., 20]

If all arms satisfy assumption 1and the sub-optimal arms satisfy the balance condition, RB-SDA satisfies, for all sub-optimal arm k,

$$\mathbb{E}[N_k(T)] \leq \frac{1+\varepsilon}{I_1(\mu_k)} \log(T) + o_\varepsilon(\log T) \ .$$

One-parameter exponential families:

- satisfy Assumption 1 and $I_1(x) = \mathrm{kl}(x, \mu_k)$
- Bernoulli, Gaussian and Poisson distributions satisfy the balance condition (with $f_r = 1$, i.e. without forced exploration)
- any exponential family satisfy the relaxed balance condition with $f_r = \sqrt{\log(r)}$

➜ RB-SDA is asymptotically optimal for *different* exponential family bandit models (possibly with unbounded support)

Outline

Works very well in practice!

Average Regret on $N = 10000$ random instances with $K = 10$

- **Bernoulli arms**

T	TS	IMED	PHE	SSMC	RB-SDA
100	**13.8**	15.1	16.7	16.5	**14.8**
1000	**27.8**	31.9	39.5	34.2	**31.8**
10000	**45.8**	51.2	72.3	55.0	**51.1**
20000	**52.2**	57.6	85.6	61.9	**57.7**

- **Gaussian arms**

T	TS	IMED	SSMC	RB-SDA
100	41.2	45.1	40.6	**38.1**
1000	76.4	82.1	76.2	**70.4**
10000	118.5	124.0	120.1	**111.8**
20000	132.6	138.1	135.1	**125.7**

many more experiments in [Baudry et al. 20]

Conclusion

Subsampling Duelling Algorithms

An alternative to UCB or Thompson Sampling that can be asymptotically optimal without prior knowledge on the type of distributions of the arms

Follow-up works:

- an analysis of LB-SDA and its potential for non-stationary bandits [Baudry et al., AISTATS 21]
- Dirichlet Sampling, a non-parametric algorithm under weaker assumptions on the arms [Baudry et al., NeurIPS 21]

Future works:

- precisely characterize the class of distributions for which SDA algorithms can be used
- extensions to more complex models (e.g., linear bandits, reinforcement learning)?

References

- Baransi et al., *Sub-sampling for multi-armed bandits*, ECML 2014
- Baudry et al., *Sub-sampling for Efficient Non-Parametric Bandit Exploration*, NeurIPS 2020
- Baudry et al., *On limited-memory subsampling strategies for Bandits*, ICML 2021
- Baudry et al., *From Optimality to Robustness: Dirichlet Sampling Strategies in Stochastic Bandits*, NeurIPS 2021
- Hock Peng Chan. *The multi-armed bandit problem: An efficient nonparametric solution.* The Annals of Statistics, 2020
- Kveton et al. *Garbage in, reward out: Bootstrapping exploration in multi-armed bandits.* ICML, 2019
- Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019
- Lai and Robbins. *Asymptotically efficient adaptive allocation rules*. Advances in Applied Mathematics, 1985
- Riou and Honda. *Bandit algorithms based on Thompson Sampling for bounded reward distributions.* ALT, 2020.
- Robbins. *Some aspects of the sequential design of experiments*. Bulletin of the American Mathematical Society, 1952