Sub-Sampling for Stationary and Non-Stationary Bandits

Dorian Baudry

Ecole Polytechnique & ENSAE Paris

Research conducted during Ph.D. at Inria Lille:

On Limited-Memory Sub-Sampling Strategies for Bandits,

by **DB**, Yoan Russac & Olivier Cappé, published at **ICML 2021.**

Outline

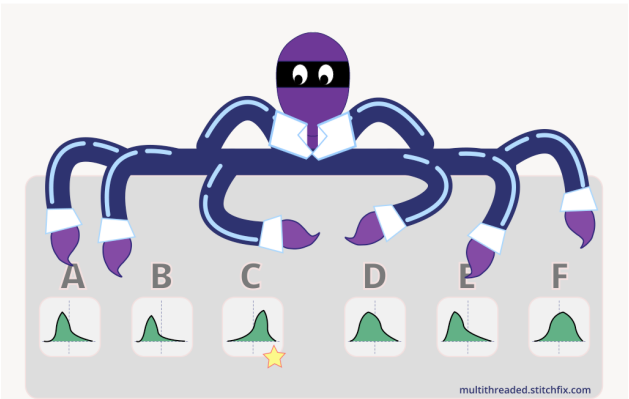
- Introduction
  - Multi-Armed Bandits (MAB)
  - Motivation for non-parametric algorithms
- Last-Block Sub-Sampling Dueling Algorithm (LB-SDA)
- From LB-SDA to SW-LB-SDA
- Ideas for future works

Table of Contents

Multi-Armed Bandits (MAB)

- K unknown reward distributions $(\nu_1, \ldots, \nu_K)$ called arms.
- At each time t a learner selects an arm and observe a (random) reward.
- **Objective**: maximize the expected sum of rewards.

  $\hookrightarrow$ Exploration/Exploitation trade-off.



Definitions

Maximizing the expected sum of rewards $\equiv$ minimizing the *regret*.

Regret: If expectations $= (\mu_1, \ldots, \mu_K)$ and $\mu^\star = \max_k \mu_k$,

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}(\mu^\star - \mu_{A_t})\right] = \sum_{k=1}^{K} \Delta_k \mathbb{E}[N_k(T)]\ ,$$

with

- $\Delta_k = \mu^\star - \mu_k$: "sub-optimality gap" of arm k.
- $N_k(T) = \sum_{t=1} \mathbb{I}\{A_t = k\}$: Number of pulls of arm k.

Achievable guarantees: $\mathcal{R}_T = \Omega\left(\sum_{k:\Delta_k>0} \frac{\log(T)}{\Delta_k} \vee \sqrt{KT}\right)$.

Non-Stationary MAB (NS-MAB)

- **Evolving distributions**: $(\nu_{1,t}, \ldots, \nu_{K,t})_{t \in \mathbb{N}}$.
- Time-dependent best arm, $k_t^\star = \operatorname{argmax} \mu_{k,t}$.
- **Models**: variation budget, smoothly-evolving environment, finite number of abrupt changes.

$\hookrightarrow$ Additional exploration is required!

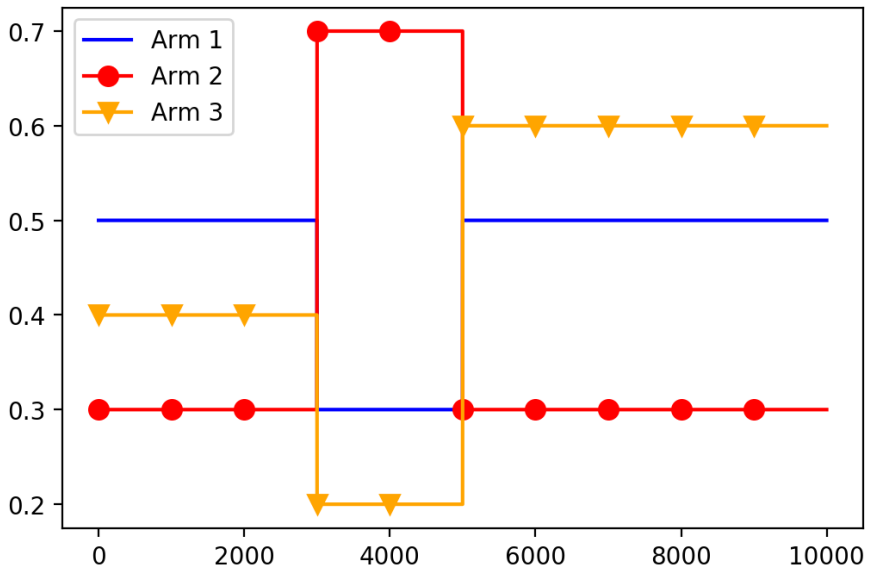


Figure: Evolving mean for 3 Bernoulli distributions, experiment from the paper.

Abruptly changing environments

Over T rounds, at most Γ_T **breakpoints**,

$$\Gamma_T = \sum_{t=1}^{T-1} \mathbb{1}\left(\exists k : \nu_{k,t+1} \neq \nu_{k,t}\right) \ .$$

$\hookrightarrow$ **Stationary phases** between breakpoints.

Dynamic Regret:

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}(\mu_t^\star - \mu_{A_t,t})\right] \leq \sum_{\phi=1}^{\Gamma_T}\sum_{k=1}^{K} \Delta_k^\phi \mathbb{E}\left[N_k^\phi\right] \ .$$

$\hookrightarrow$ Goal $= \forall\phi$, upper bound $(N_k^\phi)_{k:\Delta_k^\phi>0}$.

$\hookrightarrow$ $\mathcal{R}_T = \Omega\left(\sqrt{\Gamma_T K T}\right)$.

Motivation: Stationary case

Usual Bandit algorithms:

- Upper Confidence Bound (UCB)
- Thompson Sampling (TS)
- Index Minimized Empirical Divergence (IMED)

Require initial ***knowledge*** on the distributions: Conjugate prior/posterior, KL div...

$\hookrightarrow$ Can we obtain good guarantees without using this information?

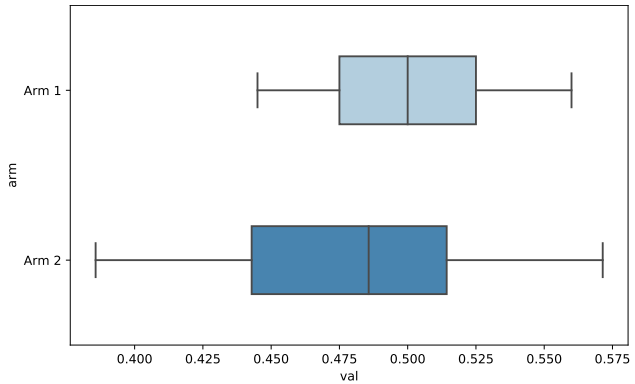


Figure: $5-95\%$ confidence intervals for empirical means, Bernoulli distrib., $(p_1 = 0.5, N_1 = 200,\ p_2 = 0.48, N_2 = 60)$

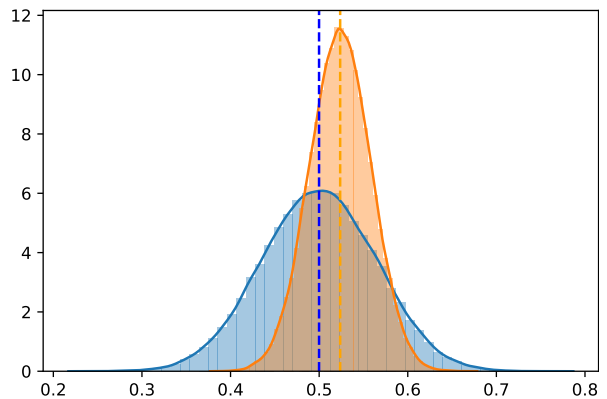


Figure: Densities of two Beta distrib.: $\mathrm{Beta}(30, 30)$ and $\mathrm{Beta}(110, 100)$

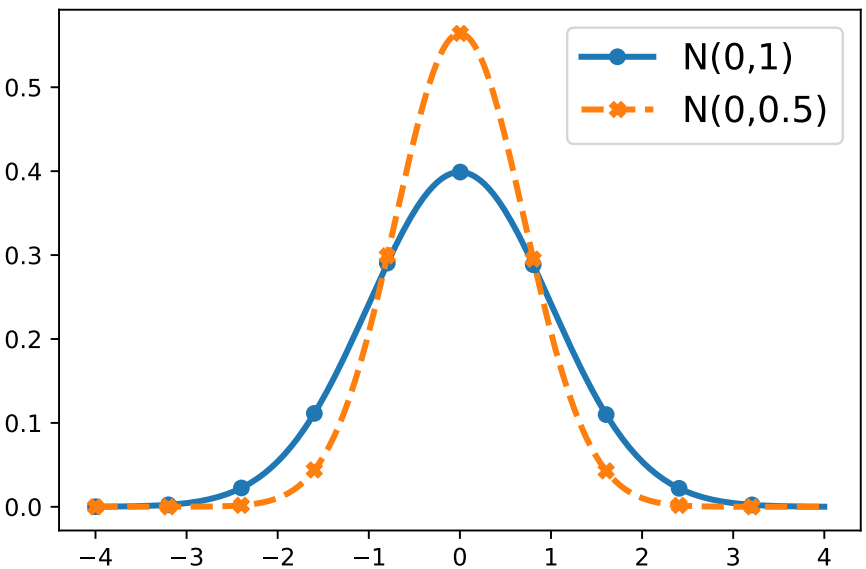
Example: additive noise model

Model: $\forall k, n,\ X_{k,n} = \mu_k + \xi_{k,n}$, where $(\xi_{k,n})_{n\in\mathbb{N}}$ are i.i.d. noise.

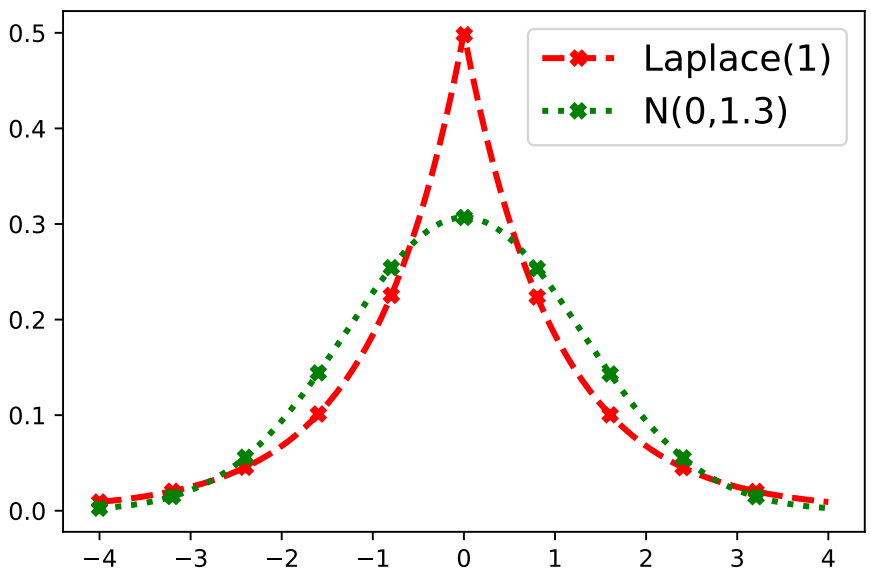
Usual assumption: $\xi_{k,n} \sim \mathcal{N}(0,1)$ (or sub-Gaussian).

What if the model is misspecified?

Gaussian Bandit works ✓



Linear regret ✗

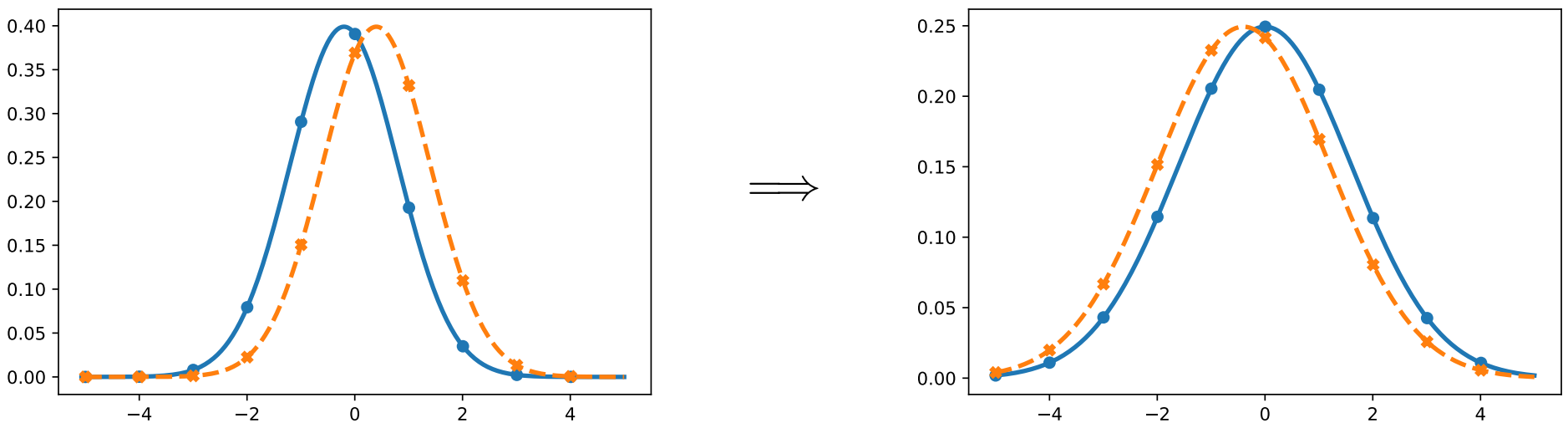


Motivation: NS-MAB

Standard Assumption: **Bounded** rewards/**same family** of distribution.

$\hookrightarrow$ What if the shape of the distributions change?

Simple example: Gaussian distributions with time-varying variance.



$\hookrightarrow$ Existing algorithms must use a conservative model, we seek adaptation.

Table of Contents

Why Sub-Sampling?

Intuition: Greedy ($A_t = \text{argmax}\, \widehat{\mu}_k(t)$) $\rightarrow$ fixed prob. of a *bad* scenario.

Example:

1. Best arm 1 collects bad samples $\Rightarrow \widehat{\mu}_1(t) \leq \mu_2 - \epsilon$ under-estimates μ_1.
2. Arm 2 is pulled a lot $\Rightarrow \widehat{\mu}_2(t) \approx \mu_2$, good estimation for 2.
3. Arm 1 is **never pulled again**, $\widehat{\mu}_2(t) \leq \mu_2 - \epsilon$ becomes too unlikely!

Idea: Comparing the means of **sub-samples** of the **same size** is a "fair" comparison between two arms!

$\hookrightarrow$ Sub-Sampling is the right way to be **Greedy**!

Idea 2: Making **diverse** comparisons *should* induce exploration.

Subsampling Dueling Algorithms (SDA)

Objective: framework for pairwise comparisons between arms.

A **round-based** approach [Chan, 2020]:

1. Choose **leader**: $\text{argmax}_{k\in[K]} N_k(t) \Rightarrow$ large history to sample from!
2. Perform $K-1$ ***duels***: *leader* vs each *challenger*.
3. Draw a set of arms: *winning challengers* (if any) or *leader* (if none).

Duel:

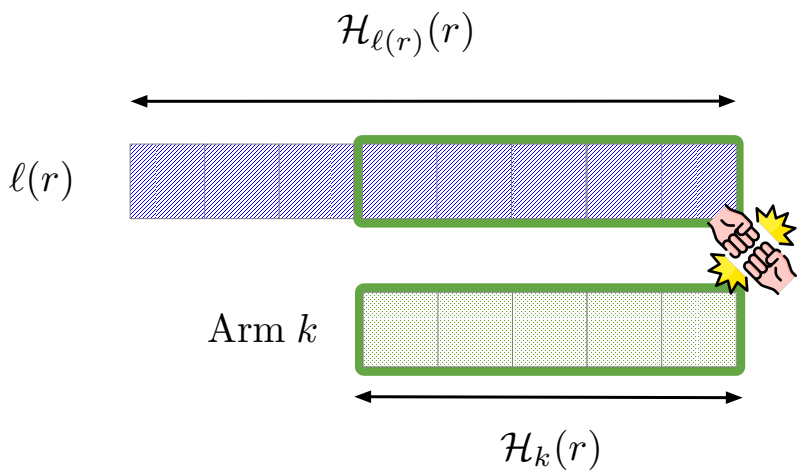
- Challenger $\rightarrow$ **empirical mean** $\widehat{\mu}_{k,N_k}$ (full sample size N_k).
- Leader $\rightarrow$ **mean** $\widehat{\mu}_{\ell,S(N_k,N_\ell)}$ of a ***subsample*** $S(N_k, N_\ell)$ of size N_k.

$$\text{Winner} = \begin{cases} k \text{ if } \widehat{\mu}_{k,n} \geq \widehat{\mu}_{\ell,S(N_k,N_\ell)} \\ \ell \text{ otherwise.} \end{cases}$$

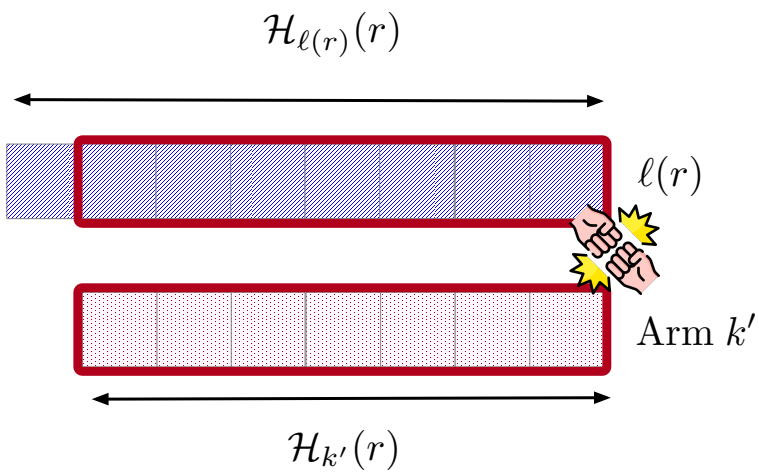
Last Block SDA (LB-SDA)

Sub-sampling: Sampling without replacement, Random Block, Last Block,. . .

↪ Last Block is **cheap**: sequential update when leader doesn't change.



Duel between k and $\ell(r)$ at round r



Duel between k' and $\ell(r)$ at round r

When/Why does LB-SDA work?

3 ingredients: *concentration*, *diversity* and *balance*.

1. **Concentration**:

$$\exists \text{ function } I_k : \; I_k(\mu_k) = 0 \text{ and } \begin{cases} \forall x \geq \mu_k, \; \mathbb{P}(\widehat{\mu}_{k,n} \geq x) \leq e^{-nI_k(x)} \\ \forall x \leq \mu_k, \; \mathbb{P}(\widehat{\mu}_{k,n} \leq x) \leq e^{-nI_k(x)} \end{cases}.$$

   $\hookrightarrow$ Mild assumption: ✓ with *light-tailed* distributions.

2. **Diversity**: LB-SDA offers a variety of sub-samples for duels.

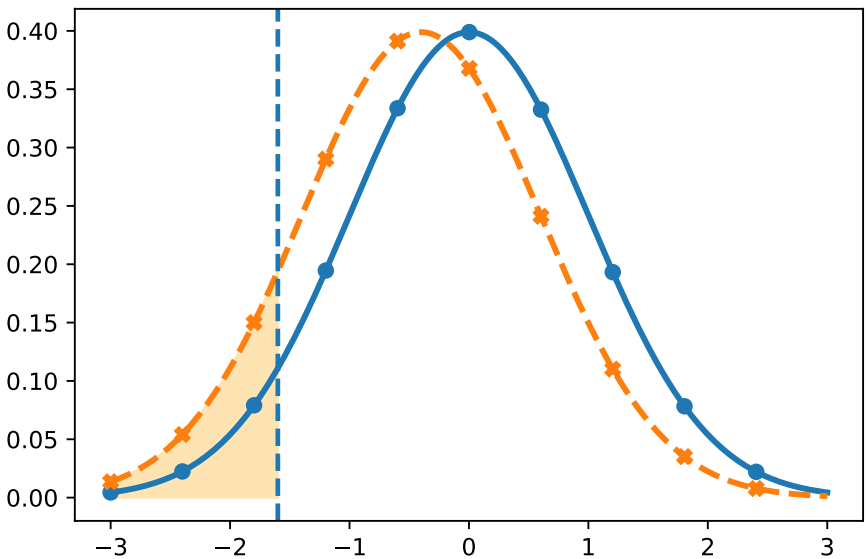
   $\hookrightarrow$ ✓ the leader is pulled most of the time $\Rightarrow$ "sliding window".

3. **Balance**: playing diverse duels guarantees a "good arm" to win in a **reasonable time** $\longrightarrow$ more restrictive.

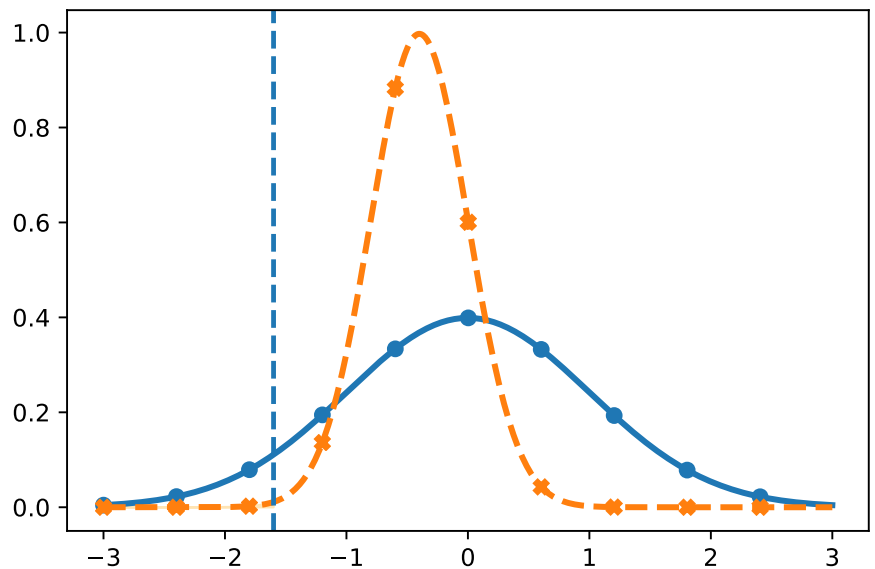
Illustration of the balance condition

$$\nu_1 = \mathcal{N}(0,1) \quad , \quad \begin{cases} \nu_2 = \mathcal{N}(-0.4, 1) & \text{(left)} \\ \nu_2 = \mathcal{N}(-0.4, 0.4) & \text{(right)} \end{cases} .$$

Good instance ✓



Bad instance ✗



$$\widehat{\mu}_{11} \leq -1.6 \ \ (\approx 10\% \text{ prob.}) \Longrightarrow \begin{cases} 1 \text{ pulled after } \approx 10 \text{ duels. (first case)} \\ 1 \text{ pulled after } \approx 750 \text{ duels. (second case)} \end{cases} .$$

Theoretical Guarantees

Theorem (Logarithmic regret of LB-SDA)

If $\nu_1, \dots, \nu_K) \in \mathcal{F}^K$ *are light-tailed and satisfy a* ***balance condition****, LB-SDA implemented with forced exploration* $f_t = \sqrt{\log(t)}$ *satisfy*

$$\forall \epsilon > 0 : \quad \mathbb{E}[N_k(T)] \leq \frac{1+\epsilon}{I_1(\mu_k)} \log(T) + \mathcal{O}_\epsilon(1) \, .$$

Corollary:

- **LB-SDA is optimal** for any Single-Parameter Exp. Family (SPEF)

  $\hookrightarrow$ satisfy the condition and $I_1(\mu_k) = \mathrm{kl}(\mu_k, \mu_1)$.

- **Logarithmic regret** in the additive noise model under mild assumptions.

$\hookrightarrow$ while using no information on the families of distributions!

Table of Contents

Motivation

Two ways to adapt standard bandit algorithms for NS-MAB:

- Passively forgetting strategies (sliding window, discount, . . .)
- Active restarts (change-point detection)

Sliding window τ = play with observations collected *within the last* τ rounds.

Equipping LB-SDA with a sliding window is natural:

- LB-SDA is "already" a sliding-window algorithm **for the leader**.
  $\hookrightarrow$ needs to be one for challengers too
- Passive forgetting is better-suited for non-parametric assumptions.

Sliding Window LB-SDA (SW-LB-SDA)

- LB-SDA with a sliding window of size τ
- Additional mechanisms to ensure diversity/balance in the NS-MAB:
  - Sampling obligation $\rightarrow \sqrt{\log(\tau)}$.
  - Diversity check: play each challenger k that played the same duel for a "long time" ($\Omega((\log(\tau)^2))$).
  - Modified leader: can become leader only by defeating the current leader.

Theorem (Regret Guarantees of SW-LB-SDA)

*If T and Γ_T (nb. of breakpoints) are known, and that during a stationary phase the **balance condition** is satisfied, then*

$$\tau = \mathcal{O}\left(\sqrt{T \log(T) \Gamma_T^{-1}}\right) \Longrightarrow \mathcal{R}_T = \mathcal{O}\left(\sqrt{T \Gamma_T \log T}\right) .$$

Experiment: Gaussian arms with evolving variance

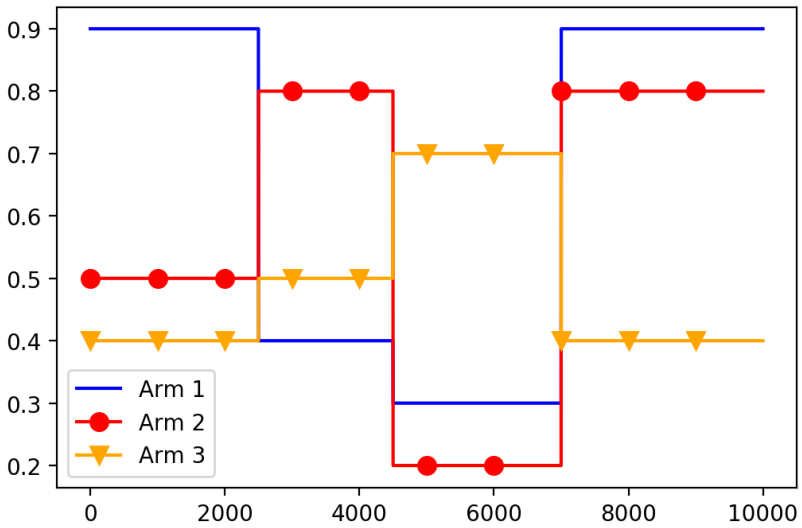


Figure: Time-dependent means, with standard deviations $\sigma = \{0.25, 0.5, 1, 0.25\}$

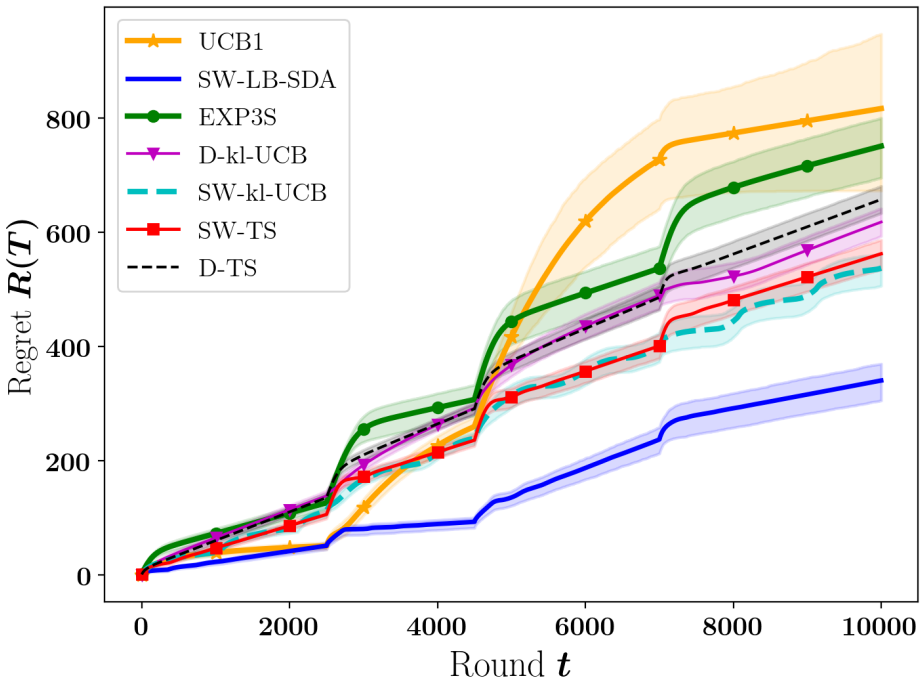


Figure: Dynamic regret averaged on 2000 independent replications.

$\hookrightarrow$ SW-LB-SDA naturally adapts to the variance changes!

Experiment: Bernoulli arms

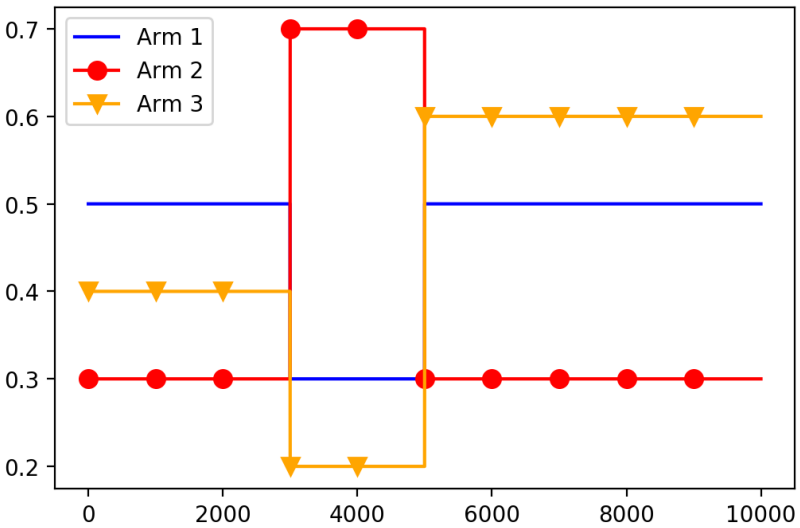


Figure: Time-dependent means

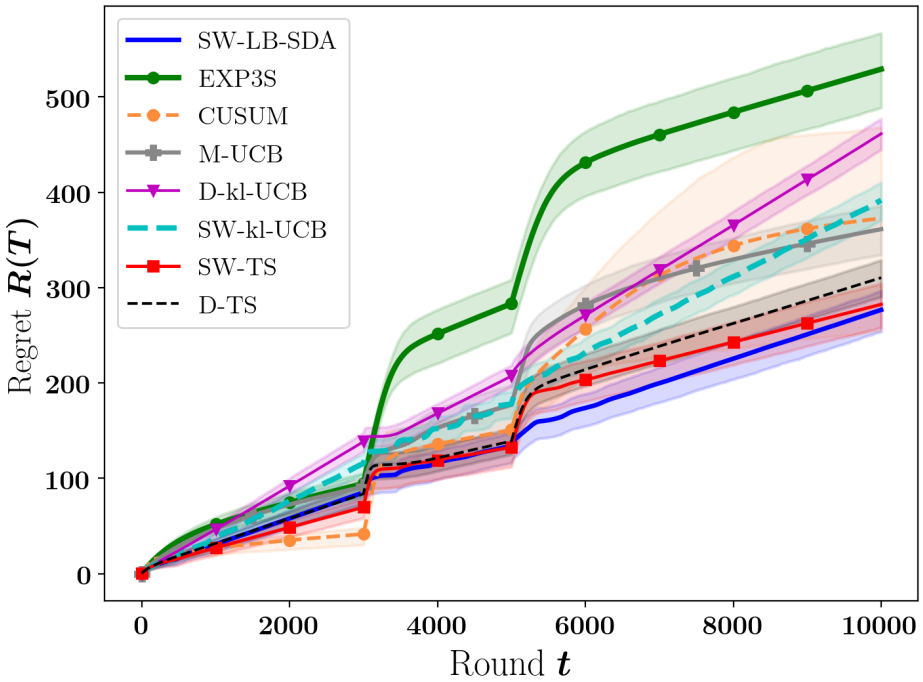


Figure: Dynamic regret averaged on 2000 independent replications.

$\hookrightarrow$ Less severe jump at the beginning of phase 3.

Table of Contents

Ideas for future works

- LB-SDA for structured problems (e.g Linear Bandits)

  $\hookrightarrow$ how to equalize the quantity of information collected?

- Analyze SW-LB-SDA under different types of non-stationarity

  $\hookrightarrow$ variation budget, slowly drifting, rotting bandits, ...

- Adapt LB-SDA when the balance condition is not satisfied?

  $\hookrightarrow$ Idea 1: change the statistics used for comparison [Chan, 2020]

  $\hookrightarrow$ Idea 2: transform the empirical distribution while preserving their order (e.g. binarized rewards [Agrawal and Goyal, 2012]).

- LB-SDA for alternative bandit problems (e.g. risk-averse bandits).

  $\hookrightarrow$ QoMAX-SDA for Extreme Bandits [Baudry et al., 2022].

Thank you for your attention !






Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem.
In Conference on learning theory, pages 39–1. JMLR Workshop and Conference Proceedings.


Baudry, D., Kaufmann, E., and Maillard, O.-A. (2020).
Sub-sampling for efficient non-parametric bandit exploration.
Advances in Neural Information Processing Systems, 33.


Baudry, D., Russac, Y., and Cappé, O. (2021).
On limited-memory subsampling strategies for bandits.
In Meila, M. and Zhang, T., editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 727–737. PMLR.


Baudry, D., Russac, Y., and Kaufmann, E. (2022).
Efficient algorithms for extreme bandits.
In The 25th International Conference on Artificial Intelligence and Statistics, AISTATS 2022.


Chan, H. P. (2020).
The multi-armed bandit problem: An efficient nonparametric solution.
The Annals of Statistics, 48(1):346–373.

Balance function

Definition (Balance function of two distributions)

For two distributions of cdf F_1 and F_2, the balance function is defined for any $(M, j) \in \mathbb{N}^2$ as

$$\alpha_{12}(M, j) := \mathbb{E}_{X \sim F_1}\left((1 - F_2(X))^M\right).$$

Interpretation:

1. Draw a sample of size j from F_1, compute $\widehat{\mu}_1$.
2. Play M duels against independent samples of size j drawn from F_2: $\widehat{\mu}_1$ vs $\widehat{\mu}_2^1, \ldots, \mu_2^M$.
3. $\alpha_{12}(M, j) = \mathbb{P}(\widehat{\mu}_1 \leq \min_j \widehat{\mu}_2^j)$: prob. that 1 does not win a single duel!

Balance condition for a bandit problem

Definition (Balance Condition)

Consider $\nu_1, \dots, \nu_K$ of resp. means $\mu_1, \dots, \mu_K$, where $\mu_1 = \max \mu_k$, and some function f_t. Let $M_t = \mathcal{O}(t/(\log t)^2)$, $n_t = \mathcal{O}((\log t)^2)$. The balance condition holds if

$$\forall k > 1 : \sum_{t=1}^{T} \sum_{j=f_t}^{n_t} \alpha_{1k}(M_t, j) = o(\log T) \ .$$

- $\hookrightarrow$ M_t is the number of "diverse" duels that we are sure to obtain with LB sub-sampling.
- $\hookrightarrow$ f_t is an amount of *forced exploration* introduced in SDA, i.e: if some arm satisfies $N_k(t) < f(t)$ it is automatically pulled.
- $\hookrightarrow$ this is the property that restrains the family of distributions for which SDA works.

Properties of the balance condition

Lemma (Some families satisfying the balance condition)

The balance condition holds for:

- ✓ *any SPEF if* $f(t) = \sqrt{\log t}$ *(See Lemma 4 in [Baudry et al., 2021]).*
- ✓ *Bernoulli, Poisson and Gaussian with shared variance with* $f(t) = 1$.
- ✓ *if the densities* f_1 *and* f_2 *satisfy* $f_1(x) \leq c f_2(x)$ *for some* $c < 1$ *and any* $x < q$ *for some* $q \in \mathbb{R}$, *with* $f(t) = \sqrt{\log t}$.

$\hookrightarrow$ For $T = 10^4$, $f_t = \sqrt{\log 10^4} \approx 3 \rightarrow$ this is not harmful in practice!

- ✖ Unfortunately there are some counter-examples: Multinomial distributions, Gaussian with different variances.

  Intuitively when the better arm has a "worse" left tail.

Empirical results for SDA

Table: Average Regret on 10000 random experiments with Bernoulli Arms (means sampled uniformly)

Horizon	TS	IMED	PHE	SSMC	RB	WR
10^2	13.8	15.1	16.7	16.5	14.8	14.3
10^3	27.8	31.9	39.5	34.2	31.8	30.9
10^4	45.8	51.2	72.3	55.0	51.1	50.6
2.10^4	52.2	57.6	85.6	61.9	57.7	57.3

Table: Average Regret on 10000 random experiments with Gaussian Arms ($\mu_i \sim \mathcal{N}(0,1)$ for each arm i)

Horizon	TS	IMED	WR	RB	WR
10^2	41.2	45.1	40.6	38.1	38.3
10^3	76.4	82.1	76.2	70.4	72.7
10^4	118.5	124.0	120.1	111.8	115.8
2.10^4	132.6	138.1	135.1	125.7	130.2

Our experiments in [Baudry et al., 2020] show that SDA perform as well as the best competitors using less knowledge!

From LB-SDA to LB-SDA-LM

Practical advantages of LB-SDA

- Fully non-parametric: same algorithm for all distributions
- Fast to compute:
  - $\mathcal{O}(1)$ most often (sequential update of the means)
  - $\mathcal{O}(log T)$ when leader changes (re-computing sub-sample means)

Drawback (shared by all subsampling algorithms)

- Storage of all T observations is required.
  Is it necessary ? $\rightarrow$ In practice only $\mathcal{O}(\log T)$ are actually used.

We study: LB-SDA with **limited storage memory** (LB-SDA-LM)

- Store $m_t = \mathcal{O}((\log t)^2)$ rewards for each arm at round t.
- Replace oldest observations by newest when capacity exceeded.

Properties

Theorem (Asymptotic Optimality LB-SDA-LM)

Just as LB-SDA, LB-SDA-LM is asymptotically optimal when arms belong to the same Single-Parameter Exponential Family (SPEF).

Table: Storage/computational cost at round T for some subsampling algorithms.

Algorithm	Storage	Comp. cost: Best-Worst case
SSMC [Chan, 2020]	$O(T)$	$O(1)$-$O(T)$
RB-SDA	$O(T)$	$O(\log T)$
LB-SDA	$O(T)$	$O(1)$-$O(\log T)$
LB-SDA-LM	$O((\log T)^2)$	$O(1)$-$O(\log T)$

Example with Bernoulli arms

$\mu_1 = 0.05$
$\mu_2 = 0.15$

Memory:

$$m_r = \log(r)^2 + 50$$

$\hookrightarrow$ Between 50 and 150 samples kept for each arm.

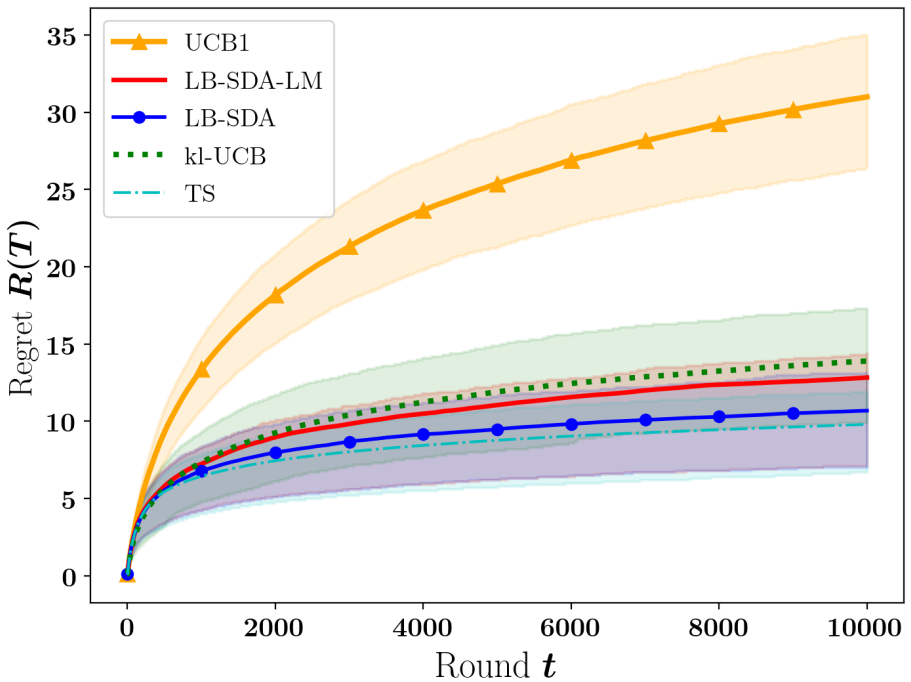


Figure: Cost of storage limitation on a Bernoulli instance. The reported regret are averaged over 2000 independent replications.

$\rightarrow$ Limiting memory does not have a significant cost in this example!