# On Limited-Memory Subsampling Strategies for Bandits

**Dorian Baudry**[*,1], Yoan Russac [*,2], Olivier Cappé[2]

[*] Equal contribution
[1] CNRS, Univ. Lille, CNRS, Inria & Université de Lille, France
[2] DI ENS, CNRS, Inria

**June, 2021**

## Stochastic $K$-armed Bandits

- $K$ *unknown* reward distributions called *arms*
- Learner sequentially collects rewards and update her policy
- Objective: minimize the *regret* $\Longleftrightarrow$ maximize the expected sum of rewards

Settings considered in this paper

- *Stationary* arms (fixed at the start)
- *Abruptly changing* environments: arms are stationary between *breakpoints*.

We study the **Last-Block Subsampling Dueling Algorithm (LB-SDA)** proposed in [Baudry et al., 2020] in these two settings.

# Last Block Subsampling Dueling Algorithm (LB-SDA)

Main idea: different number of rewards collected for each arm, comparing the means is sub-optimal (greedy).

$\rightarrow$ Comparing means of sub-sample of same size = fair comparison!

A *round-based* approach

1. Choose a *leader*: arm with largest number of observations!
2. Perform $K-1$ *duels*: *leader* vs each *challenger*.
3. Draw a set of arms: *winning challengers* (if any) or *leader* (if none).

Index used for an arm in a duel

- Challenger $\rightarrow$ empirical mean (full sample size $N_k$).
- Leader $\rightarrow$ mean of the *subsample* of its $N_k$ last observation (last block).
- Winner: arm with the largest index!

# Limiting the memory with 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)
  - $\mathcal{O}(logT)$ when leader changes (re-computing the 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.

## Idea

Store $m_t = O((\log t)^2)$ rewards for each arm at round $t \rightarrow$ LB-SDA-LM.

# Properties

**Theorem (Asymptotic Optimality of LB-SDA and LB-SDA-LM)**

*LB-SDA and LB-SDA-LM are both asymptotically optimal (see [Lai and Robbins, 1985]) when arms belong to the same Single-Parameter Exponential Family*

$\rightarrow$ for any Single-Parameter Exponential Family , unknown by the learner!

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 [Baudry et al., 2020] | $O(T)$ | $O(\log T)$ |
| LB-SDA | $O(T)$ | $O(1)$-$O(\log T)$ |
| LB-SDA-LM | $O((\log T)^2)$ | $O(1)$-$O(\log T)$ |

# Abruptly Changing Environments: SW-LB-SDA

Sliding Window LB-SDA

- Natural adaptation of LB-SDA with a sliding window of size $\tau$
- Additional mechanism to ensure sufficient exploration
- Non-parametric nature $\Rightarrow$ potential for new settings

**Theorem (Asymptotic optimality of SW-LB-SDA)**

*If the time horizon $T$ and the number of breakpoints $\Gamma_T$ are known, choosing $\tau = O(\sqrt{T \log(T)/\Gamma_T})$ ensures that the dynamic regret of SW-LB-SDA satisfies*

$$\mathcal{R}_T = O(\sqrt{T \Gamma_T \log T}) \,.$$

# Example of application with Gaussian arms

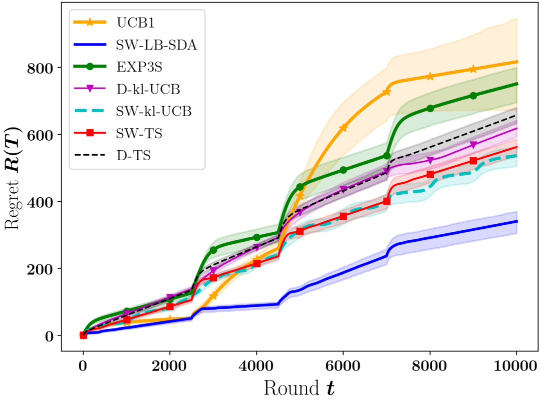


Figure: Performance on a Gaussian instance with time-dependent means and standard deviations averaged on 2000 independent replications.

→ SW-LB-SDA naturally adapts to the variance changes!

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

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

Lai, T. L. and Robbins, H. (1985).
Asymptotically efficient adaptive allocation rules.
Advances in applied mathematics, 6(1):4–22.

Thank you !