# On Limited-Memory Subsampling Strategies for Bandits

Dorian Baudry*,[1,2], Yoan Russac*,[1,3], Olivier Cappé[1,3]

[1]CNRS, Inria, [2] Université de Lille, [3]ENS, Université PSL, * Equal Contribution

## Motivations

- Bandit algorithms using subsampling have strong empirical performance for a broad family of distributions but existing works require storing the entire history of rewards [1, 2].
  ↪ We **reduce the storage constraint** by limiting the memory while maintaining the theoretical guarantees.
- **Non-stationary** environments: ubiquitous in real-world applications, and subsampling algorithms have never been applied to this setting.

## Setting

- $K$ unknown reward distributions called *arms*.
- The learner sequentially collects rewards and update her policy.
- Objective: **Minimizing the regret**

Two different settings considered:

1. **Stationary** setting: the reward distributions are fixed.
2. **Abruptly changing environment**: the rewards distributions are stationary between *breakpoints*. We consider the **dynamic regret**

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}(\mu_t^\star - \mu_{A_t})\right].$$

## Subsampling Ideas: LB-SDA Algorithm

Different number of rewards collected for each arm, a simple comparison of the means is sub-optimal (greedy algorithms)

↪ Comparing means of subsamples of the same size = fair comparison !

A *round-based* approach:

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

Subsampling index

- For a challenger: empirical mean (full sample size $N_k$ for challenger $k$).
- For the leader: mean of the subsample of its last $N_k$ observation in the duel with challenger $k$: simple and efficient subsampling method!

↪ The winner is the arm with the largest index.

## Limiting the memory: from LB-SDA to LB-SDA-LM

LB-SDA has some advantages:

- Fully non-parametric algorithm: the same algorithm can be used for different reward distributions.
- Computationally efficient: $\mathcal{O}(1)$ most often (for the sequential update of the means), $\mathcal{O}(\log T)$ when the leader changes.

Drawback: Storage of all $T$ observations required

↪ We design LB-SDA-LM (Limited Memory) to solve this issue.

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

**Theorem 1.** In any stationary environments, LB-SDA and LB-SDA-LM are both asymptotically optimal (their regret matches the Lai & Robbins Lower Bound) when the $K$ arms belong to the same Single-Parameter Exponential Family.

Comparison with existing works:

| Algorithm | Storage | Comp. cost (Best-Worst case) |
|---|---|---|
| BESA [1] | $O(T)$ | $O((\log T)^2)$ |
| SSMC [3] | $O(T)$ | $O(1)$-$O(T)$ |
| RB-SDA [2] | $O(T)$ | $O(\log T)$ |
| LB-SDA (**this paper**) | $O(T)$ | $O(1)$-$O(\log T)$ |
| LB-SDA-LM (**this paper**) | $O((\log T)^2)$ | $O(1)$-$O(\log T)$ |

Tab. 1: Storage and computational cost at round $T$ for existing subsampling algorithms

## Empirical validation

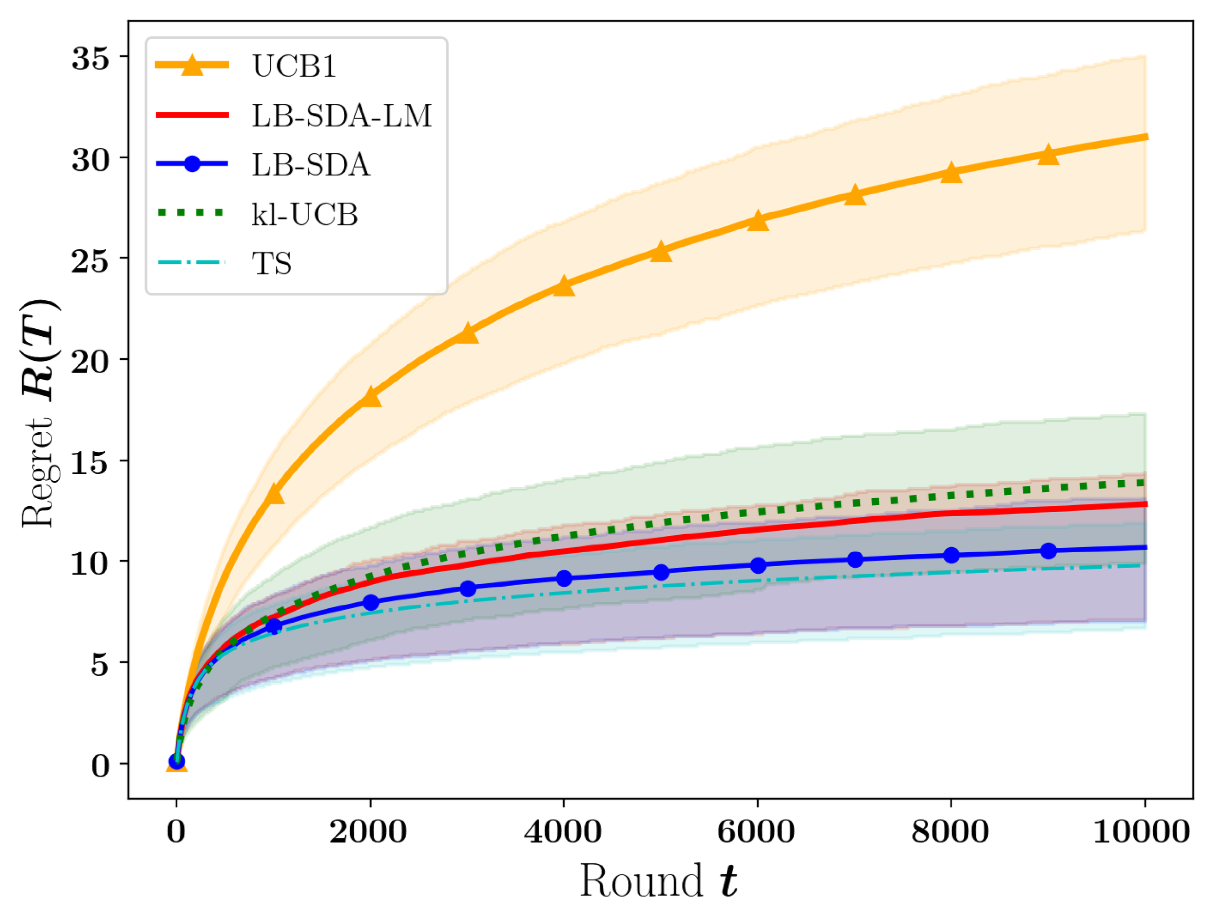


Fig. 1: Cost of storage limitation on a Bernoulli instance.

## Non-Stationary Environments: Additional Challenges

General Idea: **Combining** subsampling ideas with a **sliding-window** technique.

- Additional mechanisms to ensure sufficient exploration.
- Non-parametric nature of the algorithm: new settings for non-stationary environments with **evolving variances** and evolving means.

**Theorem 2.** If the time horizon and the number of breakpoints $\Gamma_T$ are known, for any abruptly changing environment where for each stationary period the arms comes from the same Single-Parameter Exponential Family, by choosing $\tau = \mathcal{O}(\sqrt{T\log(T)/\Gamma_T})$ the dynamic regret of SW-LB-SDA satisfies:

$$\mathcal{R}_T = \mathcal{O}\left(\sqrt{T\Gamma_T \log T}\right)$$

## Experiments in Abruptly Changing Environments

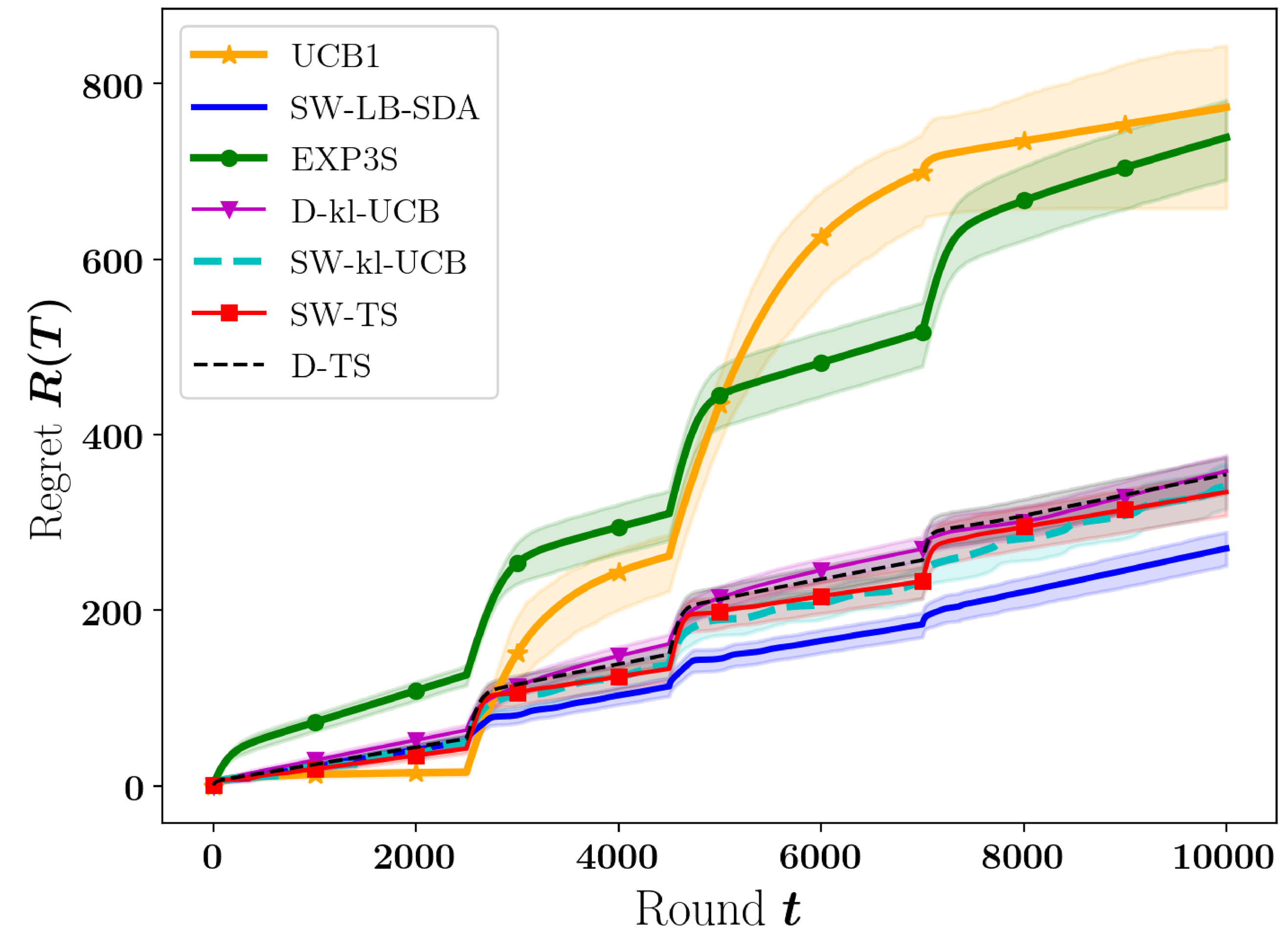


Fig. 2: Performance on a Gaussian instance with a constant standard deviation of $\sigma = 0.5$ averaged on 2000 independent runs.

### References

[1] A. Baransi, O.-A. Maillard, and S. Mannor. Sub-sampling for multi-armed bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 115–131. Springer, 2014.

[2] D. Baudry, E. Kaufmann, and O.-A. Maillard. Sub-sampling for efficient non-parametric bandit exploration. In *NeurIPS 2020*, 2020.

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