Globally Nonstationary Multi-Armed Bandits

Pratham Gupta and Sahil Chaudhary

27 October, 2025

1 Introduction

Multi Armed Bandits(MAB) has been used to solve numerous problems, such as online recommendations, online advertising, stream monitoring etc. The standard assumption is the reward generating process remains stationary, as in the distribution of the rewards doesn't change. It has been observed that in many applications like online recommendations, consumer reacts to the recommendations and other environment affecting the product and thus assuming the reward remains stationary over the time is not **fair**.

There are various algorithms which has been introduced to address this problem. Most of the non stationary MAB (NS-MAB) algorithms rely on passive forgetting methods based on a sliding window or fixed time resetting. While they address the issue of non-static environment, they fail to achieve same rates as popular algorithms like UCB, Thompson in the case stationary environment. This can be addressed by using adaptive windowing techniques like ADWIN which avoids unnecessary reset or fixed sized windows by automatically adjusting the window length based on changepoint detection. [6] propose ADS/ADR-Bandit algorithm which utilizes ADWIN to solve the problem of NS-MAB without loss of performance in the case stationary environment.

2 Problem Statement

We consider a nonstationary multi-armed bandit environment with $'K'$ arms, where the reward distribution of each arm $'i'$ at time $'t'$ has a time-varying mean $\mu_{i,t}$. The objective is to minimize cumulative regret as these means evolve.

A data stream refers to a sequence of observations $S = (x_1, x_2, \ldots, x_T)$ such that the observation i is drawn from distribution with mean μ_t. In data stream literature [4], following types of data streams evolutions are generally studied:

- **Static Streams:** if $\mu_t = \mu$ for all $t \in [T]$ and some $\mu \in [0, 1]$
- **Abruptly Changing:** if $\mu_t = \mu_{t+1}$ for all $t \in [T]$ except for a set of change points $\mathcal{T}_C = \{c_1, c_2, \ldots, c_N\}$.
- **Gradually Changing:** if $|\mu_{t+1} - \mu_t| \leq b$ for all $t \in [T]$ and some constant $b \in (0, 1)$

Definition 2.1. *(Drift Tolerant Regret)*
Assume a Nonstationary environment that is abrupt or gradual. Let $\Delta_i = \mu_{1,1} - \mu_{i,1}$ *be the initial gap between the optimal arm and arm* i*. Let* $\epsilon(t) = \max_{s \leq t} \max_{i \in [K]} |\mu_{i,t} - \mu_{i,1}|$ *be the maximum drift of the arms by time step* t*. For* $c > 0$*, let* $Reg_{tr}(T, c) := \sum_t (reg(t) - c\epsilon(t))_+$ *where* $(x)_+ = \max(x, 0)$*.* $Reg_{tr}(T, c)$ *is the regret where the regret proportional to*

drift is tolerated.
A bandit algorithm has logarithmic drift-tolerant regret if a factor $C^{dt} = O(1)$ exists such that,

$$\mathbb{E}[Reg_{tr}(T, C^{dt})] \leq C^{dt} \sum_{i \neq 1} \frac{\log T}{\Delta_i} \tag{1}$$

Note: Assume that rewards $x_{1,t}, x_{2,t}, \cdots$ are binary (i.e., 0 or 1). Then, TS and KL-UCB has logarithmic drift tolerant regret.
Definition 2.2 (Globally gradual changes)**.** *A set of K streams has globally gradual changes if a constant $C^{gr} \in (0, 1]$ exists such that, $|\mu_{i,t} - \mu_{i,s}| \geq C^{gr}|\mu_{j,t} - \mu_{j,s}|$ holds for any two arms $i, j \in [K]$ and any two time steps $t, s \in [T]$.*
Definition 2.3 (Globally abrupt changes)**.** *Let M be the number of change points $\mathcal{T}_c = \{T_1, T_2, \ldots, T_M\}$ and $(T_0, T_{M+1}) = (0, T)$. The m-th changepoint is a global change with C^{ab} if $\max_{i,j \in [K], t \in \mathcal{T}_c} \frac{|\mu_{j,t} - \mu_{j,t+1}|}{|\mu_{i,t} - \mu_{i,t+1}|} \leq C^{ab}$ is finite.*
Definition 2.4 (Detectability)**.** *For the m-th changepoint, let $\epsilon_m = \min_i |\mu_{i,m} - \mu_{i,m+1}|$. We let $U(\epsilon) = (\log(T^3))/(2\epsilon^2)$. The m-th changepoint is detectable if $T_m - T_{m-1}, T_{m+1} - T_m \geq 48KU(\epsilon_m)$.*

3 Related Work

Strategies for non stationary multi armed bandits can be broadly classified as **passive** or **active**, based on how they handle changes in reward distributions.

Passive Approaches: These algorithms adapt by continuously forgetting historical data, rather than explicitly detecting when a change occurs. Discounted UCB , sliding TS, Sliding UCB [5] are few examples of passive approaches, which use a discount factor or a fixed size window respectively. While effective, their optimal tuning often requires priori knowledge of the number of breakpoints (Υ_T), achieving $O(\sqrt{T\Upsilon_T \log T})$ regret. A different passive model uses temporal variation(V) instead of discrete breakpoints to quantify non-static. REXP-3 adapts Exp3 algorithm with periodic restarts to achieve the minimax regret of $O(KV^{1/3}T^{2/3})$ [1].

Active Approaches: Active approaches uses a change point detection(CPD) mechanism. When a change is detected, the algorithm typically resets its statistics. M-UCB[3] , for instance combines UCB with a simple CPD component that compares sample means over a sliding window, achieving $O(\sqrt{MKT \log T})$ regret, where M is the number of segments and K is the number of arms. [7] introuces UCBL-CPD and ImpCPD which achieve regret of $\tilde{O}(\sqrt{MT}), O\left(\sqrt{MT}\right)$ where M is the number of change points and T is the horizon.

The algorithm we present in this report also falls into the active-detection category. Their approach builds on the insight that by focusing on detecting global changes (a concept also explored in [7]), it is possible to adapt to non-stationarity without resorting to the forced exploration common in many active methods. This allows our algorithm to adapt efficiently while preserving optimal performance in stationary environments. In the following sections, we present an algorithm to solve the case non-stationary.

4 How to Solve it?

The key idea is to maintain a variable length window of recent data items, and check on arrival of new data item, if there exists a partition such that the average of the data items differ in these partition by a certain threshold ϵ_{cut}^{δ} dependent on the size of the sub-windows and a confidence parameter δ. We present ADS-bandit algorithm below which mixes the earlier idea with any bandit algorithm. We only present ADS-bandit algorithm in this report. [1]

Algorithm 1 ADS-bandit

Require: Set of arms $[K]$, confidence level δ, base-bandit algorithm

1: Initialize the base-bandit algorithm.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $(I(t), X(t)) =$ BASE-BANDIT$(W(t))$ ▷ Do one time step of the base-bandit algorithm
4: **ADWIN():**
5: Define, $\hat{\mu}_W = \frac{1}{|W|}\sum_{x_i \in W} x_i, \quad \epsilon_{\text{cut}}^{\delta} = \sqrt{\frac{1}{2|W_1|}\log\frac{1}{\delta}} + \sqrt{\frac{1}{2|W_2|}\log\frac{1}{\delta}}$
6: **if** $\exists$ a split $W(t+1) = W_1 \cup W_2$ such that $|\hat{\mu}_{i,W_1} - \hat{\mu}_{i,W_2}| \geq \epsilon_{\text{cut}}^{\delta}$ **then**
7: Update the window $W(t+1) = W_2$ of the base-bandit algorithm.
8: **end if**
9: **end for**

We define the error in identification of mean at time t and provide the error incurred in computation of the same due to ADWIN in next subsection:

4.1 ADWIN

Definition 4.1. *(Total Error) The total error of estimator $\hat{\mu}_W$ is defined as $Err(T) = \sum_{t=1}^{T} |\hat{\mu}_{W(t)} - \mu_t|$.*

Theorem 4.1 (ADWIN Error Bounds)**.** *Let the ADWIN algorithm be run with a confidence parameter $\delta = 1/T^3$. The expected total error, $\mathbb{E}[Err(T)]$, is bounded according to the environmental properties(under certain assumptions) as summarized in Table below:*

Environment	**Assumptions**	**Error Bound ($\mathbb{E}[\mathbf{Err}(T)]$)**
Static	Stationary stream	$\tilde{O}(\sqrt{T})$
Abrupt	M changepoints	$\tilde{O}(\sqrt{MT})$
Gradual	Change parameter $b = O(T^{-d})$ for $d \in (0, 3/2)$	$\tilde{O}(T^{1-d/3})$

Table 1: Summary of ADWIN Expected Total Error Bounds ($\delta = 1/T^3$).

Note that: [6] provides finite anlaysis of the error term which wasn't discussed in the earlier set of literature[2]. This bound on Error term helps later in regret analysis of adapting bandit algorithms for non-stationary environment.

[1] **Remark:** Due to space constraints, we omit presenting the ADR-bandit algorithm in this report, which is a modified algorithm inspired from ADS and is engineered to facilitate the analysis of the algorithm. However, we present the corresponding bound in this report.

4.2 Regret bounds of ADR-bandit algorithms

Definition 4.2 (Monitoring consistency)**.** *ADR bandit with any base bandit algorithm has the following two properties:*

1. *Let $t \geq KN + 1$. Then, at least one of the arms $\{i^{(l-1)}, i^{(l)}\}$ satisfies the following:*

   This arm was drawn at least N times before round t and will be drawn at least N times within the next KN rounds.

2. *For any block $l = 1, 2, \cdots$ there exists at least one arm that is drawn at least N times for each subblock in l.*

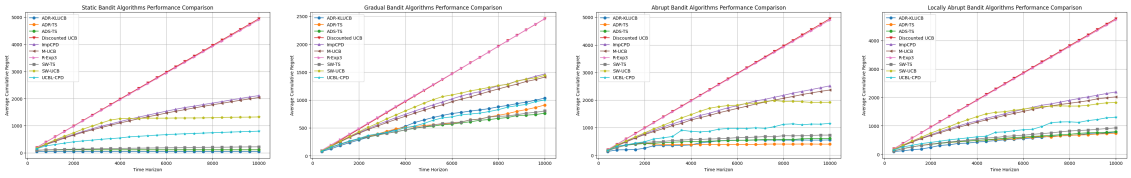
Theorem 4.2 (Regret Bounds of ADR-bandit)**.** *The expected regret of the **ADR-bandit algorithm** using a base-bandit algorithm with logarithmic drift-tolerant regret and confidence $\delta = 1/T^3$ is bounded across different environments as summarized in Table 2.*

Environment	Assumptions	Regret Bound
Stationary	Base algorithm has logarithmic stationary regret.	$\mathbb{E}[\text{Reg}(T)] \leq C^{st} \sum_i \frac{\log T}{\Delta_i} + O(1)$
Abrupt	M detectable (definition 2.4) & globally abrupt changes (definition 2.3). $N \geq 16U(\epsilon_m)$, $KN \leq (T_{m+1} - T_m)/3$.	$\mathbb{E}[\text{Reg}(T)] = \tilde{O}(\sqrt{MKT})$
Gradual	Globally gradual changes (definition 2.2) Monitoring consistency (definition 4.2) Change speed $b = T^{-d}$, $N = \tilde{\Theta}((bK)^{-2/3})$.	$\mathbb{E}[\text{Reg}(T)] = \tilde{O}(\sqrt{K}T^{1-d/3})$

Table 2: Summary of ADR-bandit Regret Bounds

5 Experiments

We tried reproducing the experiments done in [6] with same hyperparameters and we observe no performance drop of ADS/ADR-bandit algorithm in static environment and its excellence in abrupt/global against popular passive and active algorithms.



(a) Static (b) Gradual (c) Abrupt (d) Abrupt local

Figure 1: Regret using various bandit algorithms in non stationary environments

The codes for the same is available at Github Repo Link.

6 Conclusion

In this report, we explored nonstationary multi-armed bandit problems and presented ADS/ADR-bandit algorithms[6] which demonstrate significant improvement in regret performance under nonstationary environments without suffering degradation in stationary settings unlike existing algorithms.

References

[1] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards, 2019.

[2] Albert Bifet and Ricard Gavaldà. *Learning from Time-Changing Data with Adaptive Windowing*, pages 443–448.

[3] Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit, 2019.

[4] João Gama, Indrundefined Žliobaitundefined, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4), March 2014.

[5] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems, 2008.

[6] Junpei Komiyama, Edouard Fouché, and Junya Honda. Finite-time analysis of globally nonstationary multi-armed bandits. *Journal of Machine Learning Research*, 25(112):1–56, 2024.

[7] Subhojyoti Mukherjee and Odalric-Ambrym Maillard. Distribution-dependent and time-uniform bounds for piecewise i.i.d bandits, 2019.