DAL: A PRACTICAL PRIOR-FREE BLACK-BOX FRAMEWORK FOR NON-STATIONARY BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a practical, black-box framework termed Detection Augmented Learning (DAL) for the problem of non-stationary bandits without prior knowledge of the underlying non-stationarity. DAL accepts any stationary bandit algorithm as input and augments it with a change detector, enabling applicability to all common bandit variants. Extensive experimentation demonstrates that DAL consistently surpasses current state-of-the-art methods across diverse non-stationary scenarios, including synthetic benchmarks and real-world datasets, underscoring its versatility and scalability. We provide theoretical insights into DAL's strong empirical performance, complemented by thorough experimental validation.

1 INTRODUCTION

Bandit models underpin a wide range of engineering systems, from recommendation and ads to dynamic pricing and real-time bidding (Lefortier et al., 2014; Li et al., 2010; Schwartz et al., 2017; Sertan et al., 2012; Tajik et al., 2024; Flajolet & Jaillet, 2017). Many variants of multi-armed bandits (MABs) have emerged since the work of (Robbins, 1952), which fall into parametric bandits (PB) (Auer, 2002; Faury et al., 2020; Filippi et al., 2010), non-parametric bandits (NPB) (Srinivas et al., 2010) and contextual bandits (CB) (Woodroofe, 1979; Langford & Zhang, 2007). In the general bandit problem, in each round, an agent receives a context C_t randomly sampled from a set $\mathcal{C}$, and selects a policy π_t from a policy set Π—a set of mappings from $\mathcal{C}$ to a compact action set $\mathcal{A} \subseteq \mathbb{R}^d$. Then, the agent chooses action $A_t = \pi_t(C_t)$ and receives reward

$$X_t = f_t(C_t, A_t) + \varepsilon_t,$$

where $f_t : \mathcal{C} \times \mathcal{A} \to \mathbb{R}$ is the reward function and ε_t is the zero-mean sub-Gaussian noise. The goal is to minimize the dynamic regret, using a causal policy π_t based on past interactions:

$$R_T := \mathbb{E}_{\substack{A_t \sim \pi_t \\ C_t \sim \mathcal{P}_t}} \left[\sum_{t=1}^{T} \max_{\pi \in \Pi} f_t(C_t, \pi(C_t)) - f_t(C_t, A_t) \right].$$

CBs follow the general formulation above, where the context C_t is independently sampled from $\mathcal{P}_t$ and $|\mathcal{A}|$ is finite. In PB and NPB settings, the context is fixed across time and $|\mathcal{A}|$ can be infinite. With slight abuse of notation, we write $f_t(C_t, A_t) = f_t(A_t)$ in PBs and NPBs. For PBs, $f_t(A_t) = \mu(\langle \theta_t, A_t \rangle)$, where θ_t is a bounded unknown parameter and $\mu : \mathbb{R} \to \mathbb{R}$ is injective. These include linear bandits (LBs), with μ as identity, generalized linear bandits (GLBs), and self-concordant bandits (SCBs), where μ is self-concordant and the noise variance may depend on the mean (Russac et al., 2021). For NPB, we consider kernelized bandits (KBs), where $f_t \in H_k$, a reproducing kernel Hilbert space (RKHS) induced by a continuous positive semi-definite kernel $k : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ with $k(x, x) \leq 1$ and $\|f_t\|_{H_k} \leq B$. In KBs, a central complexity measure is the maximum information gain γ_T (worst-case mutual information between f and T noisy evaluations). For compact $\mathcal{A} \subset \mathbb{R}^d$: $\gamma_T = \mathcal{O}((\log T)^{d+1})$ for the Squared Exponential (SE) kernel, and $\gamma_T = \mathcal{O}(T^{\beta} \log T)$ with $\beta = d(d+1)/[2\nu + d(d+1)]$ for Matérn(ν) kernels.

Bandits remain practically relevant today: recent deployments span A/B testing (Zhang et al., 2025), clinical trials (Varatharajah & Berry, 2022), large language models (Shin et al., 2025), diffusion models (Aouali, 2024), and computer architecture (Gerogiannis & Torrellas, 2023), which even leverage the canonical formulations as the core decision engine. Accordingly, the key challenge is

developing bandit methods that perform reliably under real-world constraints—aimed at practical effectiveness, not just analysis. The lion's share of the literature on bandits assumes *stationarity*—i.e., fixed f_t, θ_t, $\mathcal{P}_t$—but this rarely holds in practice due to evolving conditions (Agrawal & Jia, 2019; Cai et al., 2017; Chen et al., 2020; Lu et al., 2019). Non-stationary (NS) settings are often categorized into two types–*gradual drifts* and *abrupt changes*. In the drifting model, f_t and $\mathcal{P}_t$ evolve slowly under a variation budget constraint (Besbes et al., 2014; Wei & Luo, 2021). In contrast, piecewise stationary (PS) models assume abrupt shifts at unknown change-points:

$$1 =: \nu_0 < \nu_1 < \cdots < \nu_{N_T} < \nu_{N_T+1} := T + 1, \quad N_T : \text{total number of changes}$$

with $f_t = f_{t'}$ and $\mathcal{P}_t = \mathcal{P}_{t'}$ for $t, t' \in \{\nu_k, \ldots, \nu_{k+1} - 1\}$ and different across change-points.

NS bandit algorithms are typically either *adaptive*—adjusting continuously, or *restarting*—choosing to unlearn and kickstart the learning process at certain times. They may also be *prior-based* (assuming knowledge of the non-stationarity) or *prior-free*. Prior-based adaptive methods (discounting/sliding window) weigh recent observations more heavily: NS-MABs (Garivier & Moulines, 2011; Kocsis & Szepesvári, 2006), NS-LBs (Cheung et al., 2019; Russac et al., 2019), NS-GLBs (Faury et al., 2021; Russac et al., 2020), NS-SCBs (Russac et al., 2021; Wang et al., 2023), NS-KBs (Deng et al., 2022; Zhou & Shroff, 2021). Prior-based restarting approaches use budgeted restarts: NS-MABs (Besbes et al., 2014), NS-LBs/GLBs (Zhao et al., 2020), NS-KBs (Zhou & Shroff, 2021). Detection-based restarting methods exist in both flavors: prior-based for NS-MABs (Cao et al., 2019b; Liu et al., 2018) and NS-CBs (Luo et al., 2018); prior-free for NS-MABs (Auer et al., 2019; Besson et al., 2022; Huang et al., 2025), for NS-LBs/KBs (Hong et al., 2023) and for NS-CBs (Wu et al., 2018; Chen et al., 2019). The most closely related work is that of Huang et al. (2025), which addresses PS-MABs and introduces techniques that we build upon in establishing our theory. However, our setting is much more complex, as it extends to *general NS bandits*.

Among prior-free methods, *black-box* approaches are particularly appealing: they equip *any* stationary bandit algorithm with non-stationarity handling capabilities. MASTER (Wei & Luo, 2021) is the only known order-optimal black-box method for general bandit and reinforcement learning settings. Importantly, although MASTER is order-optimal, it is not practically applicable (Gerogiannis et al., 2025). More broadly, the literature emphasizes theory over evidence, as empirical validation of order-optimal methods is scarce: NS-NPBs and NS-PBs are evaluated almost exclusively on synthetic data (Wang et al., 2023; Hong et al., 2023; Gerogiannis et al., 2025), and NS-CBs lack experiments altogether (Chen et al., 2019). We close these gaps with a theoretically grounded, practical black-box framework and comprehensive real-world evaluation in standard benchmarks.

Contributions. We present (to our knowledge) the first *practical* prior-free, black-box detection-based framework for general NS bandits. The design is motivated by three pragmatic insights: (i) prior knowledge of non-stationarity is rarely available, (ii) restart-style methods can have lower worst-case complexity than fully adaptive schemes (Peng & Papadimitriou, 2024), and (iii) a black-box reduction simplifies NS algorithm design to specifying when to restart a stationary learner. Our method is simple—combining a change detector with any stationary bandit algorithm—modular, and easy to implement. Empirically, extensive synthetic and real-world evaluations in standard datasets show consistent gains over both prior-free and prior-based baselines, and (to our knowledge) provide the first comprehensive real-world assessment of order-optimal baselines previously lacking empirical study. Theoretically, under mild assumptions, our regret matches the state-of-the art for PS-LBs, PS-GLBs and PS-CBs and *improves* the best known bounds for PS-SCBs and PS-KBs; for drifting regimes we identify conditions for good performance and validate them empirically.

2 The DAL Framework

The DAL framework is a black-box characterized by a modular structure of three components: a non-stationarity detector, a forced exploration scheme, and a bandit algorithm. We provide high-level ideas of the structure of our approach and formally present our framework in Alg. 1.

Non-Stationarity Detector To identify changes in the environment, DAL uses a general-purpose detector $\mathcal{D}$ for monitoring shifts in the distribution of judiciously chosen reward observation sequences obtained through forced exploration. This distinguishes our approach from methods like

MASTER, which rely on detecting violations of stationary regret guarantees. We adopt a detector aligned with Besson et al. (2022); Huang et al. (2025), grounded in the well-established theory of quickest change detection (Veeravalli & Banerjee, 2013; Xie et al., 2021). Given any arbitrary context, DAL samples rewards from actions within a carefully selected finite subset, and detects changes in the mean reward associated with the context-action pair.

Alg. 1 Detection **A**ugmented **L**earning (**DAL**)

Input: bandit $\mathcal{B}$, detector $\mathcal{D}$, covering set $\mathcal{A}_{\mathrm{e}}$, context set $\mathcal{C}$, horizon T, frequencies $\{\alpha_k\}_{k=1}^{T}$
Initialize: histories $\mathcal{H}_{(c,a)} \leftarrow \emptyset \ \forall (c,a) \in \mathcal{C} \times \mathcal{A}_{\mathrm{e}}$, detection $\tau \leftarrow 0$, counter $k \leftarrow 1$

for $t = 1, 2, \ldots, T$ **do**
  Observe context C_t
  if $(t - \tau + 1 \mod \lceil N_{\mathrm{e}}/\alpha_k \rceil) + 1 = i \in [N_e]$ **then**
    Play action $a^{(i)}$ and receive reward X_t
    Add sample X_t into history $\mathcal{H}_{(C_t, a^{(i)})}$
    if $\mathcal{D}\left(\mathcal{H}_{(C_t, a^{(i)})}\right) = \text{detection}$ **then**
      Reset the bandit algorithm $\mathcal{B}$
      Clear all $\mathcal{H}_{(c,a)} \ \forall (c,a) \in \mathcal{C} \times \mathcal{A}_{\mathrm{e}}$,
      $\tau \leftarrow t, \quad k \leftarrow k + 1$
    end if
  else
    Run the stationary bandit algorithm $\mathcal{B}$
  end if
end for

Forced Exploration In stationary bandit settings, effective algorithms quickly concentrate on (near-)optimal actions for each context, rarely exploring suboptimal actions. In NS environments, however, this behavior may lead to missed changes on these rarely sampled actions, and thus, *forced exploration* on these actions is essential. When the action space is large or infinite, exploring all actions becomes infeasible. Therefore, DAL only does extra exploration on a finite *covering set*, $\mathcal{A}_{\mathrm{e}} = \{a^{(i)} : i \in [N_{\mathrm{e}}]\} \subseteq \mathcal{A}$, in which $a^{(i)}$ denotes the i-th action in $\mathcal{A}_{\mathrm{e}}$. $\mathcal{A}_{\mathrm{e}}$ is designed such that the mean reward of at least one context-action pair in $\mathcal{C} \times \mathcal{A}_{\mathrm{e}}$ changes whenever a change occurs. In particular, after the $(k-1)^{\mathrm{th}}$ restart, DAL is forced to play each action in $\mathcal{A}_{\mathrm{e}}$ once for N_{e} steps, followed by the bandit algorithm for the next $\lceil N_{\mathrm{e}}/\alpha_k \rceil - N_{\mathrm{e}}$ steps, repeatedly, until the k^{th} restart. Here, $\alpha_k \in (0,1)$ is the exploration frequency, striking a balance between detection delay and regret from extra exploration.

Bandit Algorithm With a detector $\mathcal{D}$ and forced exploration, DAL augments a (stationary) bandit algorithm $\mathcal{B}$: It resets $\mathcal{B}$ entirely whenever $\mathcal{D}$ detects changes in a reward distribution associated with any context-action pair in $\mathcal{C} \times \mathcal{A}_{\mathrm{e}}$, and runs $\mathcal{B}$ with periodic forced exploration otherwise. A key advantage of DAL is its ability to translate strong stationary performance into robust performance under NS conditions. Therefore, by selecting a well-performing bandit algorithm, the DAL framework inherently achieves effective adaptation to NS environments. In fact, the only requirement for DAL's input stationary algorithm is to attain optimal stationary regret performance bounds.

3 Practical Performance

We evaluate DAL on piecewise-stationarity and drifting non-stationarity across synthetic and real-world benchmarks. Baselines include: MASTER (Wei & Luo, 2021), the only other black-box algorithm with order-optimal regret. While MASTER has no formal guarantees for SCBs, empirical evidence (Wang et al., 2023) supports pairing it with Logistic-UCB-1 (Faury et al., 2020) for NS-SCBs. We also include two state-of-the-art prior-free, order-optimal methods: ADA-OPKB (Hong et al., 2023) for NS-LBs/NS-KBs and ADA-ILCTB+ (Chen et al., 2019) for NS-CBs. ADA-OPKB requires extensive parameter tuning for strong performance (7 hyper-parameters), posing a challenge in a truly prior-free environment. For fair comparison, however, we tune its parameters to optimize its performance in our evaluation. We also tune MASTER's single hyperparameter (n) for best performance. Finally, two prior-based discounted methods are used for comparison: WeighTUCB (Wang et al., 2023) (drifting PBs and PS-SCBs) and WGP-UCB (Deng et al., 2022) (drifting KBs). Unless noted otherwise, we adopt the hyperparameters in the prior works in our experiments.

Across all settings, DAL uses the *Generalized Likelihood Ratio (GLR)* test (Huang & Veeravalli, 2025) as the detector $\mathcal{D}$. Concretely: In NS-LBs, LinUCB (Abbasi-yadkori et al., 2011) pairs with Gaussian GLR. In NS-GLBs, GLM-UCB (Filippi et al., 2010) pairs with Gaussian GLR. In NS-SCBs, OFUGLB (Lee et al., 2024) pairs with Bernoulli GLR. In NS-KBs, REDS (Salgia et al., 2024) pairs with Gaussian GLR. In NS-CBs, SquareCB (Foster & Rakhlin, 2020) pairs with Bernoulli GLR. The GLR test implementation follows Huang et al. (2025), and we provide their structure in

the Appendix. For all settings, we set $\alpha_k = \sqrt{k||\mathcal{C}|N_e}/(2\sqrt{T}\log^2 T)$ as per Theorem 4.4. A crucial advantage of DAL is that it is *hyperparameter-free*, guided entirely by our theoretical principles. Due to space constraints, the selection of $\mathcal{A}_e$ is discussed more analytically in the Appendix.

3.1 Synthetic Experiments

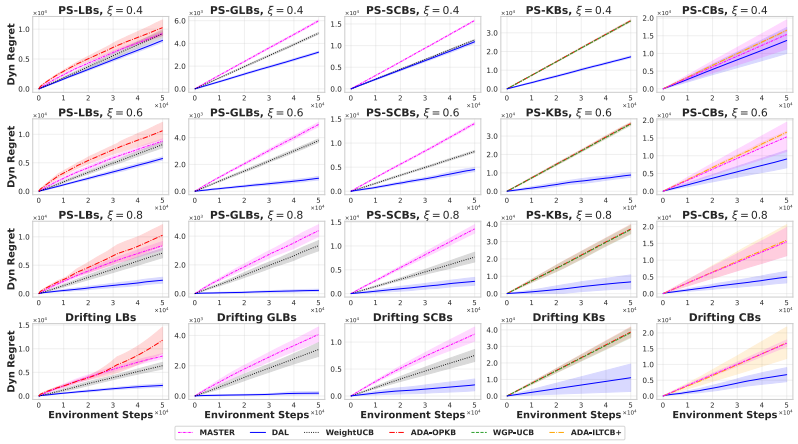

Figure 1: Dynamic regret vs. environment steps for synthetic experiments (lower=better). First three rows correspond to the geometric change-points and the final one to the drifting case.

3.1.1 Experimental Parameters

Common parameters In all synthetic experiments, the action space comprises 100 unique actions with dimension $d = 10$. These actions are sampled independently from $\mathcal{N}(0, I)$. The horizon is fixed to $T = 50000$ and we average the results over 15 independent trials.

Remark 3.1. *In Alg. 1, when $|\mathcal{A}|$ is finite, change-detection can be performed on the actions selected by $\mathcal{B}$ that are not in $\mathcal{A}_e$, which improves performance. This does not affect the theoretical properties of the algorithm, and we employ this variation for our experiments.*

NS-PBs The actions are scaled to lie within an L-ball and the underlying parameters θ_t belong to an S-ball. Specifically, for NS-LBs and NS-GLBs we have that $S = L = 1$, while for NS-SCBs, we have that $L = 1$ but $S = 3$. Every time a θ_t is initialized or changed, its elements are chosen independently and uniformly in $[-1, 1]$, and then are scaled to the S-ball. For both NS-GLBs and NS-SCBs, we select $\mu(x) := \sigma(x) = (1 + e^{-x})^{-1}$ (sigmoid). The additive noise ε_t is sampled according to $\mathcal{N}(0, 0.01)$ at each time-step, while for NS-SCBs, we sample the random reward according to $\text{Bernoulli}(\mu(\langle\theta_t, A_t\rangle))$ at time t. To set $\mathcal{A}_e$ in NS-PBs, we use Corollary 4.5.

NS-KBs Actions are scaled in the $\sqrt{d}$-ball and $\varepsilon_t \sim \mathcal{N}(0, 0.01)$. Regarding the choice of kernel for NS-KBs, we employ the SE kernel with $\ell = 0.2$. We follow a procedure similar to Chowdhury & Gopalan (2017); Deng et al. (2022). Specifically, every time we initialize or change the reward function, f_t is generated from the RKHS obtained by a discretization of $[-1, 1]$ into 200 evenly spaced points. The reward function is generated as $f(\cdot) = \sum_{i=1}^{M} \alpha_i k(\cdot, x_i)$ with α_i uniformly distributed in $[-1, 1]$ and $M = 200$. To identify $\mathcal{A}_e$, we use Corollary 4.5.

NS-CBs The context $C_t \in \mathbb{R}^{d_c}$ is drawn at each round from a fixed pool of 1000 normalized vectors with $d_c = 10$, according to a categorical distribution. At every initialization or change, at least one of the f_t or $\mathcal{P}_t$ changes. For $a \in \mathcal{A}$ and context C_t, f_t is clipped in $[0, 1]$, and is given by

$$f_t(C_t, a) = \left[b_a + z^{(\text{sig})}\,\sigma(u_a^\top C_t) + z^{(\text{sin})}\,\sin(v_a^\top C_t) + z^{(\text{xpr})}\,C_{t,2}C_{t,3}\right]_{[0,1]},$$

where $u_a, v_a \sim \mathcal{N}(0, I)$, $b_a \sim \text{Unif}[0.3, 0.7]$, and $z^{(\text{sig})}, z^{(\text{sin})}, z^{(\text{xpr})}$ are drawn uniformly from $[0.25, 0.45]$, $[0.15, 0.35]$, $[0.10, 0.25]$, respectively. Rewards are sampled as $\text{Bernoulli}(f_t(C_t, A_t))$. Since the reward function lacks any arm-related structure, here we set $\mathcal{A} = \mathcal{A}_e$ (see Remark 4.3).

3.1.2 EXPERIMENTAL BENCHMARKS

Piecewise Stationarity In the PS setting, we adopt the geometric change-point model proposed in Gerogiannis et al. (2025), and independently sample the intervals between the change-points according to a geometric distribution with parameter $\rho = T^{-\xi}$, for $\xi \in \{0.4, 0.6, 0.8\}$. We do not impose any restriction on the lengths of the intervals between change-points in our experiments.

Drifting Non-Stationarity Regarding comparisons in drifting non-stationarity, we adopt the following drift model: in each run, the reward structure changes linearly over T rounds from an initial value to a final value, where the end-points are chosen as in the beginning of the section. Specifically,

$$\text{PBs: } \theta_t = (1 - t/T)\,\theta_{\text{init}} + (t/T)\,\theta_{\text{final}}, \quad \text{KBs: } f_t = (1 - t/T)\,f_{\text{init}} + (t/T)\,f_{\text{final}},$$

$$\text{CBs: } \phi_t = (1 - t/T)\,\phi_{\text{init}} + (t/T)\,\phi_{\text{final}},\ \phi_t := (u_{a,t}, v_{a,t}, b_{a,t}, \mathbf{z}_t),\ \mathbf{z}_t := (z_t^{(\text{sig})}, z_t^{(\text{sin})}, z_t^{(\text{xpr})}).$$

Experimental Results Per the results in Figure 1, DAL outperforms the current state-of-art methods in every synthetic experiment. DAL only abandons the actions chosen by the stationary bandit algorithm and restarts learning when an efficient change detector flags a mean-shift in rewards; hence, it avoids unnecessary restarts, especially when the intervals between the change-points are long enough for such detectors to correctly flag said changes without false alarms. Regarding drifting non-stationarity, DAL significantly outperforms all other methods. In fact, it fares better than both WeightUCB and ADA-OPKB, which not only are known to attain the optimal regret bound in the drift setup, but have also been shown to perform well in practice.

3.2 REAL-WORLD EXPERIMENTS

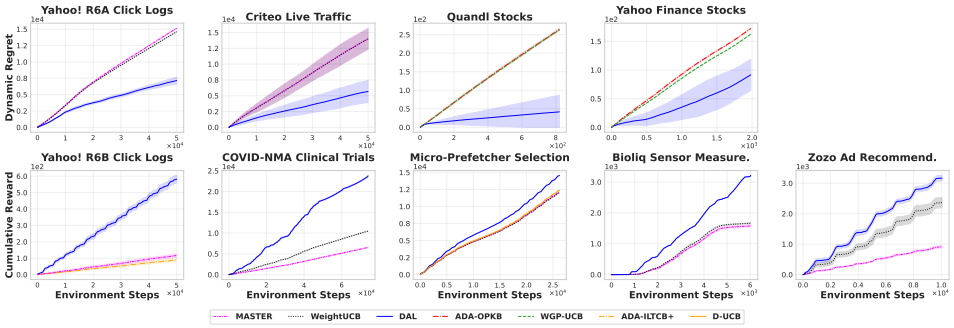

Figure 2: Results for real-world experiments of Section 3.2, averaged over 15 independent runs. Top: dynamic regret (lower=better); Bottom: cumulative reward (higher=better).

Microarchitecture Prefetcher Selection Benchmark. We introduce a novel dataset for NS bandit evaluation using the data of Gerogiannis & Torrellas (2023). The dataset is derived from the SPEC06/17 benchmark suites, the gold standard in evaluating computer microarchitectural models. The dataset includes 11 prefetcher configurations (actions) that trade aggressiveness against efficiency. At each time-step, the reward is the normalized instructions per cycle in $[0, 1]$, and the horizon is $T = 26224$. We obtained the data directly from the authors.[1] Following Gerogiannis & Torrellas (2023), we also evaluate D-UCB (Kocsis & Szepesvári, 2006) in its native form; while for our baselines we model the task as an NS-SCB. For reproducibility, D-UCB hyperparameters follow its original paper and Gerogiannis & Torrellas (2023). Evaluation is by cumulative reward.

Stock Market Benchmarks. NS-KBs have been applied to stock market prediction, and we follow the procedure of (Deng et al., 2022) to simulate two environments: one using their original data (Quandl stocks) and one constructed from NASDAQ-100 stocks retrieved via the `yfinance` Python package.[2] In the Yahoo-based dataset, we retain stocks with sufficient history (T=2000,

[1] We aim to release the dataset to facilitate real-world experimentation by the bandit research community.

[2] Data retrieved from Yahoo Finance using the publicly available `yfinance` package. Used solely for non-commercial, academic research purposes.

approx. 5.5 years) and select the 50 most volatile as actions. Daily closing prices define the reward function, and the empirical price covariance matrix is used as the kernel. To increase difficulty, we add Gaussian noise $\mathcal{N}(0, 0.01)$ to the reward at each time-step. Evaluation is by dynamic regret.

COVID-NMA Clinical Benchmark. We construct an NS-SCB benchmark from the open COVID-NMA database (Boutron et al., 2025). To maximize coverage while retaining clinical meaning, we form a *UNION* endpoint: for each bucketed-treatment arm and month, we include both Clinical Improvement at Day 28 (when reported) and Survival at Day 28 (1–mortality) as separate contributions, leading to binary rewards (1=success). Treatments (actions) are mapped into 13 classes and month counts are expanded exactly (s successes and $n-s$ failures per bin) and concatenated in a fixed chronological order (month, `clinD28` then `mortD28`, then bucket) to yield a long non-stationary sequence with $T \approx 7.4\times10^4$. Evaluation is based on cumulative reward.

Click Log Benchmarks. We use the Yahoo! R6A click log dataset.[3] Following prior works (Cao et al., 2019b; Seznec et al., 2020), we compute smoothed click-through rates (CTRs) via rolling averages over 2000 rounds, average CTRs within each subperiod, and suppress fluctuations below 0.005. To increase difficulty, we combine actions across 5 days, leading to 64 actions, compress the horizon to 50000, and multiply final CTRs by 10. We model the resulting environment as an NS-SCB problem, reflecting the logistic reward structure typical in such settings (Russac et al., 2021). Evaluation is by dynamic regret.

Alongside the first benchmark, we build a fixed-arm replay benchmark from the additional Yahoo! R6B click logs and cast it as an NS-CB problem.[3] We select the highest-CTR articles to form an action set of 51 actions. To improve coverage at a fixed horizon, we round-robin interleave days and then select $T = 50000$. For each visit, we intersect the candidate set with this vocabulary, keep rounds where the displayed item remains, and record the binary click as the raw reward ($X_t \in \{0, 1\}$). We rely on R6B's uniform-random logging for unbiased replay/IPS evaluation (Li et al., 2011). Our metric is (replay) cumulative reward.

Live Traffic Benchmark. We construct an NS bandit environment based on the Criteo live traffic dataset (Diemert et al., 2017), following the preprocessing approach of Russac et al. (2019) but modeling the problem as an NS-GLB rather than an NS-LB. We estimate the underlying parameter θ^* using logistic regression. Unlike Russac et al. (2019), in which the authors employ a single change, we introduce shifts in θ^* via a geometric change-point model with parameter $\xi = 0.8$ and extend the horizon to $T = 50000$. The metric here is the dynamic regret.

Sensor Correlation Benchmark. We use the Bioliq dataset provided by Komiyama et al. (2024), which contains a week of measurements from 20 sensors in a power plant. We process the reward as Komiyama et al. (2024) and construct an NS-SCB environment with 190 actions. At each time-step, the reward is 1 if the last 1000 measurements exceed a threshold of 2.04, and 0 otherwise. Evaluation is based on cumulative reward.

Ad Recommendation Benchmark. We evaluate on the Zozo environment, a real-world ad recommender system deployed on an e-commerce platform, introduced by Saito et al. (2021). Using the dataset preprocessed by Komiyama et al. (2024), we construct an NS-GLB environment that captures the dynamics of online ad recommendation. Unlike Komiyama et al. (2024), in which the authors limit the setup to 10 actions due to sparsity, we keep all 80 ads as actions. Following their setup, we assign a reward of 1 to any ad that received at least one user click within a one-second window, and 0 to ads with no clicks. Here, evaluation is based on cumulative reward.

Based on the results in Figure 2, DAL consistently outperforms all state-of-the-art baselines across real-world benchmarks, in both dynamic regret and cumulative reward. We attribute this strong performance to the robustness DAL demonstrates in the synthetic settings, which captured a range of challenging NS scenarios. These findings underscore DAL's practical effectiveness. In what follows, we provide a theoretical explanation for its performance.

[3] Yahoo! Front Page Today Module User Click Log Datasets: https://webscope.sandbox.yahoo.com.

4 THEORETICAL INSIGHTS

4.1 ON EFFECTIVE DETECTION

When selecting a non-stationarity detector, accuracy and efficiency are essential for ensuring optimal regret growth. Any detector aiming to identify distribution shifts inherently requires a certain number of samples, both before and after the change. Ideally, this sample complexity should scale appropriately to avoid negatively impacting the total regret. To this end, GLR tests have been shown to achieve a pre- and post-change sample complexity of the order $\log T$ (Huang & Veeravalli, 2025). Since logarithmic terms are disregarded in dynamic regret analyses, it suggests that integrating this detection mechanism can achieve optimal regret growth.

To select which samples should be fed into the detector, one needs to properly select the covering set $\mathcal{A}_{\mathrm{e}}$, so that it contains actions that can capture changes in the reward function for any context. However, changes cannot be arbitrarily small, as no change detector may be able to identify them. Hence, $\mathcal{A}_{\mathrm{e}}$ should be designed such that whenever a change occurs, reward sequences associated with at least one context-action pair in $\mathcal{C} \times \mathcal{A}_{\mathrm{e}}$ exhibit an *appreciable* mean-shift. Define

$$\Delta_{\mathrm{c}} \coloneqq \inf_{f \neq f'} \max_{(c,a) \in \mathcal{C} \times \mathcal{A}_{\mathrm{e}}} |f(c,a) - f'(c,a)|.$$

Then, Δ_{c} captures how well the context-action pairs in $\mathcal{C} \times \mathcal{A}_{\mathrm{e}}$ can discern between candidate reward functions. According to Huang et al. (2025), Δ_{c} crucially affects the performance of the GLR test, as its pre- and post- change sample complexity grows with $1/\Delta_{\mathrm{c}}^2$. The more discernible the changes are, the easier the detection becomes. Since forced exploration incurs regret, $\mathcal{A}_{\mathrm{e}}$ should ideally be chosen to minimize N_{e} while maximizing Δ_{c}. However, this cannot be done since the function f_t is unknown. Hence, we provide the ways with which one can ensure appreciable mean-shift (i.e., $\Delta_{\mathrm{c}} > 0$) in settings where the reward function has a certain *structure* (e.g., linear dependence on the arms or prescribed smoothness). Specifically, the NS-PB and NS-KB settings satisfy such conditions. Using these choices of $\mathcal{A}_{\mathrm{e}}$, one can guarantee order-optimal regret in certain cases, as shown in the next section. The proofs of the following propositions are given in the Appendix.

Proposition 4.1. *In NS-PBs, $\mathcal{A}_{\mathrm{e}}$ can be any arbitrary maximal linearly independent subset of $\mathcal{A}$.*

Proposition 4.2. *In NS-KBs, assume that $\mathcal{A} \subseteq [0, R]^d$ w.l.o.g., and that there exists an $\tilde{a} \in \mathcal{A}$ s.t.*

$$\inf_{f \neq f'} |f(\tilde{a}) - f'(\tilde{a})| > 2BL_u\delta_T,$$

where BL_u is the Lipschitz constant of all $f \in H_k(\mathcal{A})$. Let $\mathcal{V}_T \subset \mathcal{A}$ be the set of the centers of the balls of an arbitrary δ_T-cover. Then, $\mathcal{A}_{\mathrm{e}}$ can be taken as $\mathcal{V}_T$, with $|\mathcal{V}_T| \leq \lceil \sqrt{d}R/2\delta_T \rceil^d$.

Remark 4.3. *In NS-CBs, if f_t and $\mathcal{A}$ satisfy the structural assumptions of the preceding propositions for any fixed context, we can set $\mathcal{A}_{\mathrm{e}}$ similarly. Without such structure, we set $\mathcal{A}_{\mathrm{e}} = \mathcal{A}$, as $\mathcal{A}$ is finite.*

4.2 ON ORDER-OPTIMALITY IN PIECEWISE STATIONARY ENVIRONMENTS

In the PS setting, the minimax regret lower bound under bandit feedback is $\tilde{\Omega}(\sqrt{N_T T})$ (Garivier & Moulines, 2011),[4] which applies across all settings considered in this work, differing only in problem-dependent constants. Under certain conditions on the minimum spacing between change-points stated in Huang et al. (2025), our algorithm matches this bound with state-of-the-art dependence on these constants. Specifically, the assumption states that $\nu_k - \nu_{k-1}$ should be large enough for the change detector to acquire enough samples for triggering restarts.

Let $\ell_{\mathcal{D}}$ and $m_{\mathcal{D}}$ be the latency and the pre-change window length of the detector $\mathcal{D}$ (Huang et al., 2025), respectively. To characterize DAL's performance under piecewise stationarity, we employ the methodology of Huang et al. (2025), incorporating the regret analysis of the stationary bandit algorithm and that of the change detector. Theorem 4.4 is based on Theorem 1 from Huang et al. (2025), however, since we are studying general bandits, additional novel analysis is required. Thus, a proof sketch and the new analysis are provided in the Appendix.

Theorem 4.4. *For the PS setting in Section 1, consider Alg. 1 using the GLR test with parameters δ_{F} and δ_{D}, a stationary bandit algorithm $\mathcal{B}$ with regret upper bound $R_{\mathcal{B}}$, a covering set $\mathcal{A}_{\mathrm{e}}$ and*

[4] We use the $\sim$ in $\tilde{\Omega}(\cdot)$ to hide polylogarithmic factors.

forced exploration frequencies $(\alpha_k)_{k=1}^T$. *Assume* $\nu_1 \geq m_1$ *and* $\nu_k - \nu_{k-1} \geq \ell_{k-1} + m_k$ *for* $k \in \{2, \ldots, N_T\}$, *where* $m_k = \lceil N_e/\alpha_k \rceil m_{\mathcal{D}}$ *and* $\ell_k = \lceil N_e/\alpha_k \rceil \ell_{\mathcal{D}}$ *for* $k \in [N_T]$ *in PS-PBs and PS-KBs, and* $m_k = \lceil N_e/\alpha_k \rceil [\log(T) + m_{\mathcal{D}}/s]$ *and* $\ell_k = \lceil N_e/\alpha_k \rceil [\log(T) + \ell_{\mathcal{D}}/s]$ *for* $k \in [N_T]$ *with* $s = \min_{c \in \mathcal{C}, t \in [T]: \mathcal{P}_t(c) > 0} \mathcal{P}_t(c)$ *in PS-CBs. If* $R_{\mathcal{B}}(T) = \tilde{\mathcal{O}}(d^p \gamma_T^q (|\mathcal{A}| \log |\Pi|)^r \sqrt{T})$ *with* $p, q, r \geq 0$, $\delta_F = \delta_D = T^{-\gamma}$ *for* $\gamma > 1$, *and* $\alpha_k = \sqrt{k|\mathcal{C}|N_e}/(2\sqrt{T}\log^2 T)$, *then DAL's regret satisfies,* $R_T = \tilde{\mathcal{O}}(d^p \gamma_T^q (|\mathcal{A}| \log |\Pi|)^r \sqrt{N_T T} + \sqrt{|\mathcal{C}| N_e N_T T})$.

In PS-PBs and PS-KBs, since the context is fixed, the minimum change-point separation $\ell_{k-1} + m_k$ is smaller. However, in PS-CBs, a context may appear with a low probability, and thus, the minimum change-point separation requires an extra $\log T$ term and an extra s factor. Using Theorem 4.4 and Propositions 4.1 and 4.2, we now present the optimal regret DAL attains.

Corollary 4.5. *Assume that the conditions of Theorem 4.4 hold. In PS-PBs, select* $\mathcal{A}_e$ *as in Proposition 4.1. In PS-KBs, select* $\mathcal{A}_e$ *as in Proposition 4.2 with* $\delta_T := \frac{Rd^{1/2-2p/d}}{2(C\gamma_T^{2q})^{1/d}}$ *for some* $C > 0$. *In PS-CBs, set* $\mathcal{A}_e$ *as in Remark 4.3. Then, DAL attains*

$$R_T = \tilde{\mathcal{O}}(d^p \gamma_T^q (|\mathcal{A}| \log |\Pi|)^r \sqrt{N_T T}).$$

If the base stationary algorithm has order-optimal regret, DAL retains optimality in PS-PBs, PS-KBs and PS-CBs. This also holds when $N_e < d$ *or* $|\mathcal{A}| < d$ *in PS-PBs, when* $N_e < \gamma_T$ *or* $|\mathcal{A}| < \gamma_T$ *in PS-KBs, and when* Π *is the universal set of all mappings from* $\mathcal{C}$ *to* $\mathcal{A}$.

Remark 4.6. *The assumption on the change-points is necessary to prove the order-optimality,* ***but it is not for practical performance****. None of our experiments enforced this assumption, and DAL dominated in both the synthetic and the real-world simulations as shown in Section 3.*

The assumption on the minimum separation between change-points essentially requires scaling as $\tilde{\mathcal{O}}(\sqrt{T/k})$. However, this condition primarily emerges from a conservative proof technique used in Huang et al. (2025), where missed detections are aggregated into a single adverse event. Practically, and as corroborated by our experiments, this assumption is often violated without negatively impacting the regret performance—even under scenarios with frequent and arbitrarily placed change-points (e.g. $\xi = 0.4$). We suspect that this resilience arises because the GLR test, while potentially missing isolated short intervals, reliably detects subsequent changes when stationary segments exceed the threshold length. Even if a change is entirely missed during a segment shorter than $\tilde{\mathcal{O}}(\sqrt{T/k})$, the resulting regret remains under that order. Conversely, when the assumption holds, optimal regret is provably guaranteed. Thus, the required separation threshold acts as a practical "sweet spot": segments longer than this threshold are detected reliably, ensuring optimal performance, while shorter segments incur minimal regret, thereby preserving overall optimal regret guarantees.

State-of-the-art Regret. In line with the black-box design philosophy, Corollary 4.5 enables regret upper bounds across all settings considered, with flexibility in the choice of stationary bandit algorithms. When using specific stationary algorithms from Section 3, DAL matches the state-of-the-art regret bounds in PS-LBs and PS-GLBs at $\tilde{\mathcal{O}}(d\sqrt{N_T T})$. In PS-CBs, DAL achieves the state-of-the-art regret bound of $\tilde{\mathcal{O}}(\sqrt{|\mathcal{A}| N_T T \log |\Pi|})$. More notably, DAL improves the best known bounds in the PS-SCB and PS-KB settings. For PS-SCBs, the strongest, prior-based, bound is due to WeightUCB (Wang et al., 2023), which achieves $\tilde{\mathcal{O}}(d^{2/3} T^{2/3} N_T^{1/3})$.[5] DAL improves this to $\tilde{\mathcal{O}}(d\sqrt{N_T T})$ with our algorithmic choices. Although this matches the bound in Russac et al. (2021), their analysis relies on substantially stronger assumptions than those in Huang et al. (2025). For PS-KBs, the prior-free ADA-OPKB (Hong et al., 2023) achieves $\tilde{\mathcal{O}}(\sqrt{d\gamma_T N_T T})$, while DAL improves this to $\tilde{\mathcal{O}}(\sqrt{\gamma_T N_T T})$. This highlights the interesting feature of DAL: the order-wise dependence on problem parameters from the stationary setting seamlessly transfers to the PS setting without degradation. A more detailed comparison of regret bounds is provided in the Appendix.

4.3 On Drifting Environments

Based on the previous section, at first glance, one can expect that DAL is not able to handle drifting non-stationarity. Our results in Section 3 naturally lead us to ask when and why DAL performs well

[5] While MASTER may be extendable to PS-SCBs, no regret bound is currently known.

in drifting environments. As a first step to study this, we perform another experiment with LBs. Specifically, the parameter θ_t in each time-step t evolves randomly as follows,

$$\theta_{t+1} := \theta_t + \zeta_{t+1}$$

where $\zeta_{t+1} \in \mathbb{R}^d$ is chosen uniformly over a δ-ball. If the resulting θ_{t+1} violates the norm-bound S, we disregard that choice of ζ_{t+1} and sample again. We sample $\varepsilon_t \sim \mathcal{N}(0, 0.1)$ at each t. We compare the cumulative dynamic regret up to time T of DAL+LinUCB and WeightUCB over a range of δ's in Figure 3. The remaining parameters are chosen to be the same as those in Section 3, with the exception of $d = 5$. The DAL algorithm performs better than WeightUCB for smaller values of δ, but the conclusion reverses upon increasing δ.

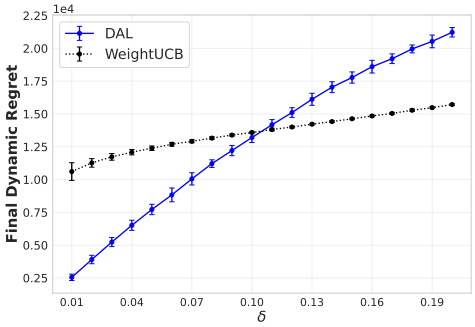

Figure 3: Final dynamic regret vs. radius of change δ: Drifting LBs.

We now shed light into our hypothesis behind the observations from Figure 3. For playing an action $a \in \mathcal{A}$ at time $t + 1$, we get the random reward,

$$X_{t+1} = \langle \theta_t, a \rangle + \langle \zeta_{t+1}, a \rangle + \varepsilon_{t+1}.$$

If the governing parameter does not change, then the corresponding reward from playing action a would have been $X'_{t+1} = \langle \theta_t, a \rangle + \varepsilon'_{t+1}$, where ε'_{t+1} is another realization of the noise. Statistically, a specific instance of $\langle \zeta_{t+1}, a \rangle + \varepsilon_{t+1}$ and ε'_{t+1} are close to each other, when δ is small, albeit the resulting (small) mean-shift due to the drift in the governing parameter. For practical purposes, the impact of the drift can be absorbed into the noise term ε_{t+1} when δ is small. As a result, one expects an algorithm tailored to handle piecewise stationarity to perform reasonably well for slowly drifting environments. On the other hand, if δ is large, the bias induced by ζ_{t+1} is large enough to disallow absorbing it into the zero-mean noise term. Over a few time-steps, the cumulative effect of this compounding bias is then large enough to completely violate the stationarity assumption. With large enough δ, the change in θ_t over a few time-steps can be considered large enough to trigger a restart.

5 Summary and Outlook

We introduced DAL, a practical, prior-free black-box framework for general non-stationary bandits. Its plug-and-play design integrates seamlessly with a wide range of stationary bandit algorithms. Through extensive experiments in both PS and drifting settings—spanning synthetic and real-world benchmarks, DAL consistently outperforms all prior-free baselines, including the black-box gold standard MASTER and the state-of-the-art methods ADA-OPKB and ADA-ILCTB+, and even surpasses leading prior-based methods like WeightUCB and WGP-UCB. Its leading performance in real-world scenarios highlights its value as a practical and effective solution.

On the theoretical side, using existing results and providing novel techniques, we showed that DAL inherits and adapts the regret guarantees of its stationary input algorithm, achieving order-optimal regret under piecewise stationarity, with mild change-point separation. As a result, it matches the best existing bounds in PS-LBs, PS-GLBs and PS-CBs while improving the best known bounds for PS-SCBs and PS-KBs. Regarding drifting non-stationarity, we hypothesized key conditions under which DAL excels–an insight further validated through additional experiments under drifting settings. Our results suggest that a well-designed algorithm for the PS setting can extend to a broad range of drifting scenarios, bridging the gap between these two regimes.

While DAL advances both theory and practice, it opens new directions. First, regret guarantees for detection-based methods in drifting environments remain unexplored. Second, the current regret bounds for DAL rely on a separation condition between change-points—a standard assumption in the detection-based literature (see e.g., (Auer et al., 2019; Besson et al., 2022; Huang et al., 2025))—which nonetheless limits the extent to which DAL achieves fully prior-free theoretical optimality. Addressing these gaps would deepen our understanding of detection-based methods in more continuous forms of non-stationarity. Finally, DAL's modular nature invites extensions to broader settings, including general non-stationary reinforcement learning. We believe that deepening the study of piecewise stationarity may be the key to tackling these broader challenges and DAL can serve as a solid foundation towards that goal.

REFERENCES

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1638–1646, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/agarwalb14.html.

Shipra Agrawal and Navin Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 127–135, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/agrawal13.html.

Shipra Agrawal and Randy Jia. Learning in Structured MDPs with Convex Cost Functions: Improved Regret Bounds for Inventory Management. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, 2019.

Imad Aouali. Diffusion models meet contextual bandits with large action spaces, 2024. URL https://arxiv.org/abs/2402.10028.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively Tracking the Best Bandit Arm with an Unknown Number of Distribution Changes. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 138–158. PMLR, 25–28 Jun 2019.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/903ce9225fca3e988c2af215d4e544d3-Paper.pdf.

Lilian Besson, Emilie Kaufmann, Odalric-Ambrym Maillard, and Julien Seznec. Efficient Change-Point Detection for Tackling Piecewise-Stationary Bandits. *Journal of Machine Learning Research*, 23(77):1–40, 2022.

Isabelle Boutron, Anna Chaimani, Lina Ghosn, Theodoros Evrenoglou, Gabriel Ferrand, Alexander Jarde, Hillary Bonnet, Carolina Riveros, Ruba Assi, Carolina Graña, Nicholas Henschke, Declan Devane, Jörg J. Meerpohl, Gabriel Rada, Giacomo Grasselli, Asbjørn Hróbjartsson, David Tovey, and Philippe Ravaud. Covid-nma initiative: Living systematic review on covid-19 treatments and vaccines — dataset. Zenodo, March 2025. URL https://doi.org/10.5281/zenodo.14965887. Dataset.

Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malialis, Jun Wang, Yong Yu, and Defeng Guo. Real-Time Bidding by Reinforcement Learning in Display Advertising. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM)*, 2017.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly Optimal Adaptive Procedure with Change Detection for Piecewise-Stationary Bandit. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 418–427. PMLR, 16–18 Apr 2019a. URL https://proceedings.mlr.press/v89/cao19a.html.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly Optimal Adaptive Procedure with Change Detection for Piecewise-Stationary Bandit. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 418–427. PMLR, 16–18 Apr 2019b. URL https://proceedings.mlr.press/v89/cao19a.html.

Chacha Chen, Hua Wei, Nan Xu, Guanjie Zheng, Ming Yang, Yuanhao Xiong, Kai Xu, and Zhenhui. Toward A Thousand Lights: Decentralized Deep Reinforcement Learning for Large-Scale Traffic Signal Control. In *AAAI Conference on Artificial Intelligence*, 2020.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 696–726. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/chen19b.html.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to Optimize under Non-Stationarity. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1079–1087. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/cheung19b.html.

Sayak Ray Chowdhury and Aditya Gopalan. On Kernelized Multi-armed Bandits. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 844–853. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/chowdhury17a.html.

Nando De Freitas, Alex J. Smola, and Masrour Zoghi. Exponential regret bounds for Gaussian process bandits with deterministic observations. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 955–962, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Yuntian Deng, Xingyu Zhou, Baekjin Kim, Ambuj Tewari, Abhishek Gupta, and Ness Shroff. Weighted Gaussian Process Bandits for Non-stationary Environments . In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6909–6932. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/deng22b.html.

Eustache Diemert, Julien Meynet, Pierre Galland, and Damien Lefortier. Attribution Modeling Increases Efficiency of Bidding in Display Advertising. In *Proceedings of the ADKDD'17*, ADKDD'17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450351942. doi: 10.1145/3124749.3124752. URL https://doi.org/10.1145/3124749.3124752.

Louis Faury, Marc Abeille, Clement Calauzenes, and Olivier Fercoq. Improved Optimistic Algorithms for Logistic Bandits. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3052–3060. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/faury20a.html.

Louis Faury, Yoan Russac, Marc Abeille, and Clément Calauzènes. Regret Bounds for Generalized Linear Bandits under Parameter Drift, 2021. URL https://arxiv.org/abs/2103.05750.

Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clement Calauzenes. Jointly Efficient and Optimal Algorithms for Logistic Bandits . In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 546–580. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/faury22a.html.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric Bandits: The Generalized Linear Case. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf.

Arthur Flajolet and Patrick Jaillet. Real-Time Bidding with Side Information. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/0bed45bd5774ffddc95ffe500024f628-Paper.pdf.

Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3199–3210. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/foster20a.html.

Aurélien Garivier and Eric Moulines. On Upper-Confidence Bound Policies for Switching Bandit Problems. In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann (eds.), *Algorithmic Learning Theory*, pp. 174–188, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-24412-4.

Argyrios Gerogiannis, Yu-Han Huang, and Venugopal Veeravalli. Is Prior-Free Black-Box Non-Stationary Reinforcement Learning Feasible? In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2692–2700. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/gerogiannis25a.html.

Gerasimos Gerogiannis and Josep Torrellas. Micro-armed bandit: Lightweight & reusable reinforcement learning for microarchitecture decision-making. In *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO '23, pp. 698–713, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400703294. doi: 10.1145/3613424.3623780. URL https://doi.org/10.1145/3613424.3623780.

Kihyuk Hong, Yuhang Li, and Ambuj Tewari. An Optimization-based Algorithm for Non-stationary Kernel Bandits without Prior Knowledge. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 3048–3085. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/hong23b.html.

Yu-Han Huang and Venugopal V. Veeravalli. Sequential Change Detection for Learning in Piecewise Stationary Bandit Environments, 2025. URL https://arxiv.org/abs/2501.10974.

Yu-Han Huang, Argyrios Gerogiannis, Subhonmesh Bose, and Venugopal V. Veeravalli. Change Detection-Based Procedures for Piecewise Stationary MABs: A Modular Approach, 2025. URL https://arxiv.org/abs/2501.01291.

Levente Kocsis and Csaba Szepesvári. Discounted ucb. In *2nd PASCAL Challenges Workshop*, volume 2, pp. 51–134, 2006.

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. URL http://jmlr.org/papers/v25/21-0916.html.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized Exploration in Generalized Linear Bandits. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2066–2076. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/kveton20a.html.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/4b04a686b0ad13dce35fa99fa4161c65-Paper.pdf.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Jungyhun Lee, Se-Young Yun, and Kwang-Sung Jun. A Unified Confidence Sequence for Generalized Linear Models, with Applications to Bandits. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 124640–124685. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e17afd5784c442d178744533c16d3c96-Paper-Conference.pdf.

Damien Lefortier, Pavel Serdyukov, and Maarten de Rijke. Online Exploration for Detecting Shifts in Fresh Intent. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pp. 589–598, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325981. doi: 10.1145/2661829.2661947. URL https://doi.org/10.1145/2661829.2661947.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 661–670, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772758. URL https://doi.org/10.1145/1772690.1772758.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pp. 297–306, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450304931. doi: 10.1145/1935826.1935878. URL https://doi.org/10.1145/1935826.1935878.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably Optimal Algorithms for Generalized Linear Contextual Bandits. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2071–2080. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/li17c.html.

Fang Liu, Joohyun Lee, and Ness Shroff. A Change-Detection Based Framework for Piecewise-Stationary Multi-Armed Bandit Problem. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11746. URL https://ojs.aaai.org/index.php/AAAI/article/view/11746.

Junwei Lu, Chaoqi Yang, Xiaofeng Gao, Liubin Wang, Changcheng Li, and Guihai Chen. Reinforcement Learning with Sequential Information Clustering in Real-Time Bidding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1739–1776. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/luo18a.html.

Binghui Peng and Christos Papadimitriou. The complexity of non-stationary reinforcement learning. In Claire Vernade and Daniel Hsu (eds.), *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pp. 972–996. PMLR, 25–28 Feb 2024.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535, 1952.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted Linear Bandits for Non-Stationary Environments. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/263fc48aae39f219b4c71d9d4bb4aed2-Paper.pdf.

Yoan Russac, Olivier Cappé, and Aurélien Garivier. Algorithms for Non-Stationary Generalized Linear Bandits, 2020. URL https://arxiv.org/abs/2003.10113.

Yoan Russac, Louis Faury, Olivier Cappé, and Aurélien Garivier. Self-Concordant Analysis of Generalized Linear Bandits with Forgetting . In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 658–666. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/russac21a.html.

Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=tyn3MYS_uDT.

Sudeep Salgia, Sattar Vakili, and Qing Zhao. Random Exploration in Bayesian Optimization: Order-Optimal Regret and Computational Efficiency. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43112–43141. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/salgia24a.html.

Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

Girgin Sertan, Mary Jérémie, Preux Philippe, and Nicol Olivier. Managing advertising campaigns—an approximate planning approach. *Frontiers of Computer Science*, 6(2):209, 2012. doi: 10.1007/s11704-012-2873-5. URL https://journal.hep.com.cn/fcs/EN/abstract/article_3688.shtml.

Julien Seznec, Pierre Menard, Alessandro Lazaric, and Michal Valko. A single algorithm for both restless and rested rotting bandits. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3784–3794. PMLR, 26–28 Aug 2020.

Suho Shin, Chenghao Yang, Haifeng Xu, and MohammadTaghi Hajiaghayi. Tokenized bandit for LLM decoding and alignment. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=TFXxarWZzv.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 1015–1022, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Mahmoud Tajik, Babak Mohamadpour Tosarkani, Ahmad Makui, and Rouzbeh Ghousi. A novel two-stage dynamic pricing model for logistics planning using an exploration–exploitation framework: A multi-armed bandit problem. *Expert Systems with Applications*, 246:123060, 2024. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.123060. URL https://www.sciencedirect.com/science/article/pii/S0957417423035625.

Yogatheesan Varatharajah and Brent Berry. A contextual-bandit-based approach for informed decision-making in clinical trials. *Life*, 12(8):1277, 2022.

V. V. Veeravalli and T. Banerjee. Quickest Change Detection. In *Academic press library in signal processing: Array and statistical signal processing*. Academic Press, Cambridge, MA, 2013.

Jing Wang, Peng Zhao, and Zhi-Hua Zhou. Revisiting Weighted Strategy for Non-stationary Parametric Bandits. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7913–7942. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/wang23k.html.

Chen-Yu Wei and Haipeng Luo. Non-stationary Reinforcement Learning without Prior Knowledge: an Optimal Black-box Approach. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4300–4354. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/wei21b.html.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979. doi: 10.1080/01621459.1979.10481033. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10481033.

Qingyun Wu, Naveen Iyer, and Hongning Wang. Learning contextual bandits in a non-stationary environment. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pp. 495–504, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210051. URL https://doi.org/10.1145/3209978.3210051.

Liyan Xie, Shaofeng Zou, Yao Xie, and Venugopal V. Veeravalli. Sequential (Quickest) Change Detection: Classical Results and New Directions. *IEEE Journal on Selected Areas in Information Theory*, 2(2):494–514, 2021. doi: 10.1109/JSAIT.2021.3072962.

Yu Zhang, Shanshan Zhao, Bokui Wan, Jinjuan Wang, and Xiaodong Yan. Strategic a/b testing via maximum probability-driven two-armed bandit. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=BwYQ1MTrCR.

Yu-Jie Zhang and Masashi Sugiyama. Online (Multinomial) Logistic Bandit: Improved Regret and Constant Computation Cost. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 29741–29782. Curran Associates, Inc., 2023.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A Simple Approach for Non-stationary Linear Bandits. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 746–755. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/zhao20a.html.

Huozhi Zhou, Lingda Wang, Lav Varshney, and Ee-Peng Lim. A Near-Optimal Change-Detection Based Algorithm for Piecewise-Stationary Combinatorial Semi-Bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6933–6940, Apr. 2020. doi: 10.1609/aaai.v34i04.6176. URL https://ojs.aaai.org/index.php/AAAI/article/view/6176.

Xingyu Zhou and Ness Shroff. No-Regret Algorithms for Time-Varying Bayesian Optimization. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, 2021. doi: 10.1109/CISS50987.2021.9400292.

A Experimental Details

A.1 Change Detectors

Below, we provide the *algorithmic implementation* of the Generalized Likelihood Ratio (GLR) test, including the necessary information for both its Gaussian and Bernoulli versions.

Alg. 2 Generalized **L**ikelihood **R**atio (GLR) Test

Input:observations $X_1, \ldots, X_n$, false alarm probabililty δ_{F}, missed detection probability δ_{D}
Output:`detection` if a change-point is detected, else `no-detection`

1: **for** $s = 1$ to $n - 1$ **do**
2: Compute empirical means $\hat{\mu}_{1:s}$, $\hat{\mu}_{s+1:n}$, and $\hat{\mu}_{1:n}$.
3: Compute test statistic:

$$\mathrm{GLR}_s \leftarrow s \cdot \mathrm{kl}(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}) + (n-s) \cdot \mathrm{kl}(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n})$$

4: **if** $\mathrm{GLR}_s \geq 6 \log(1 + \log(n)) + \frac{5}{2} \log\left(\frac{4n^{3/2}}{\delta_{\mathrm{F}}}\right) + 11.$ **then**
5: **return** `detection`
6: **end if**
7: **end for**
8: **return** `no-detection`

In the case of the Bernoulli GLR version we have that,

$$\mathrm{kl}(x, y) = x \ln\left(\frac{x}{y}\right) + (1-x) \ln\left(\frac{1-x}{1-y}\right).$$

The Bernoulli GLR test is used in the case of sub-Bernoulli distributions (Besson et al., 2022), e.g., distributions bounded in [0,1].

On the other hand, for the case of the Gaussian GLR version, assuming variance σ^2, we have that,

$$\mathrm{kl}(x, y) = \frac{(x-y)^2}{2\sigma^2}.$$

The Gaussian GLR test is employed in the case of σ^2-sub-Gaussian distributions, e.g., bounded distributions in $[a, a + 2\sigma]$ for some $a \in \mathbb{R}$.

For the practical tuning of the detector, we follow the information provided in Huang et al. (2025).

A.2 On Forced Exploration in Finite Action Spaces

Covering Set Construction. In practice, the covering set $\mathcal{A}_{\mathrm{e}}$ is selected according to Propositions 4.1, 4.2, and Remark 4.3 together with the specifications of Corollary 4.4. However, in finite-action settings, the full construction may not be feasible: the action set $\mathcal{A}$ may not contain enough elements to satisfy the required conditions. For instance, in the NS-PB setting, $\mathcal{A}$ may not include d linearly independent actions, while in the NS-KB case, it may lack a full δ_T-covering net for the chosen δ_T in Corollary 4.4. One might expect that when $|\mathcal{A}_{\mathrm{e}}| < d$ in PS-PBs or $|\mathcal{A}_{\mathrm{e}}| < \gamma_T$ in PS-KBs, the inability to detect all possible changes would degrade DAL's performance. In practice, however, DAL does not need to restart when changes in the reward function leave the mean reward of each action unchanged. Crucially, as discussed in Appendix B.5, DAL retains order-optimality even in these constrained regimes. Accordingly, whenever $|\mathcal{A}| < d$ or $|\mathcal{A}| < \gamma_T$, we simply set $\mathcal{A}_{\mathrm{e}} = \mathcal{A}$. In our experiments, the action set is finite (typically in the hundreds). For PS-PBs, the random generation of actions almost always guarantees d linearly independent vectors. For PS-KBs, since γ_T is typically large, we also use the full action set $\mathcal{A}$ as $\mathcal{A}_{\mathrm{e}}$ without impacting performance. On the other hand, since the regret bounds in PS-CBs include $|\mathcal{A}|$, as it is finite, in any PS-CB setting we can simply set $\mathcal{A}_e = \mathcal{A}$.

Practical Implementations. For NS-PBs, we construct $\mathcal{A}_{\mathrm{e}}$ by greedily selecting linearly independent actions until we obtain d such vectors, where d is the dimension of the action space. In the

NS-KB setting, $\mathcal{A}_e$ is formed by building a δ_T-cover over the bounded action space and choosing the centers of the covering balls. If the action space is continuous and bounded, these centers suffice to cover the space. If the action space is finite and $N_e < d^{2p}\gamma_T^{2q}\gamma_T$, then the entire set $\mathcal{A}$ serves as the covering set, as established in Corollary 4.4. Otherwise, if $N_e > d^{2p}\gamma_T^{2q}\gamma_T$, we select the $d^{2p}\gamma_T^{2q}\gamma_T$ actions closest to the covering-ball centers. Finally, in the NS-CB setting, selecting a smaller $\mathcal{A}_e$ compared to $\mathcal{A}$ does not affect regret, but improves practical performance due to less forced exploration. Thus, depending on the reward function and action set structures, it is recommended to decrease the cardinality of $\mathcal{A}_e$ as much as possible.

Experimental Choices. In our experiments, for NS-PBs the action set is sampled from a multivariate Gaussian distribution, which ensures the existence of d linearly independent actions. Thus, we always set $N_e = d$ using the greedy selection procedure described above. For NS-KBs, the regret bound for N_e obtained from Theorem 4.3 and Corollary 4.4 is extremely large for our horizons, implying that $|\mathcal{A}| < \gamma_T$. Consequently, in all NS-KB experiments we simply take $\mathcal{A}_e = \mathcal{A}$ and set N_e equal to the number of available actions, which yielded optimal performance. Finally, since the reward does not exhibit any structure with the arms in PS-CBs, we simply set $\mathcal{A} = \mathcal{A}_e$.

A.3 Real-World Data Preprocessing

Microarchitecture Prefetcher Selection Benchmark. We introduce a non–stationary bandit dataset derived from the MICRO'23 study of Gerogiannis & Torrellas (2023), built on the SPEC06/17 benchmark suites. Each action corresponds to one of 11 L2 prefetcher configurations (next–line on/off, stream degree, stride degree). The sequence spans T=26224 rounds; at round t, the reward is the trace–level normalized instructions–per–cycle in $[0, 1]$, computed from performance counters. We obtained the data directly from the original authors, and note that reproducing the exact series from scratch is not feasible without the same stack, microarchitectural parameters, and arm schedules described in the paper. We aim to release the dataset to facilitate real-world experimentation by the bandit research community.

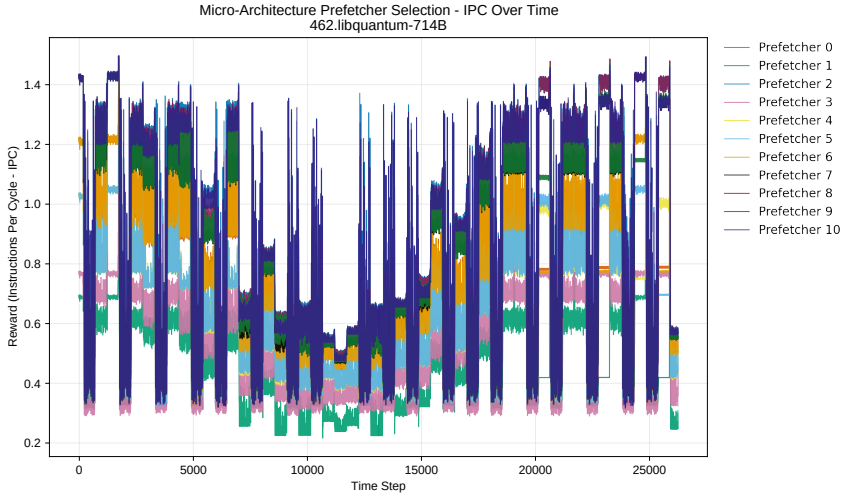

Figure 4: IPC of the prefetchers of the dataset over time.

Stock Market Data Construction. Regarding the stock market experiments we follow the procedure of Deng et al. (2022). For the first experiment, we use the data provided in Deng et al. (2022). For the other experiment, we collect daily closing prices of NASDAQ-100 companies using the Yahoo Finance API.[6] We filter out stocks with fewer than $T = 2000$ trading days and align all time series over the most recent T dates. From this pool, we remove stocks with extremely high volatility or mean price to make the problem non-trivial, then select the top K most volatile stocks from

[6]Data retrieved from Yahoo Finance using the publicly available `yfinance` package. Used solely for non-commercial, academic research purposes.

the remainder. In both cases, the stock prices are scaled accordingly to lie in $[0, 1]$. Each selected company's scaled closing price series defines the mean-reward sequence for one arm in a K-armed bandit problem. Finally, we corrupt the reward at each time step with $\mathcal{N}(0, 0.01)$ noise.

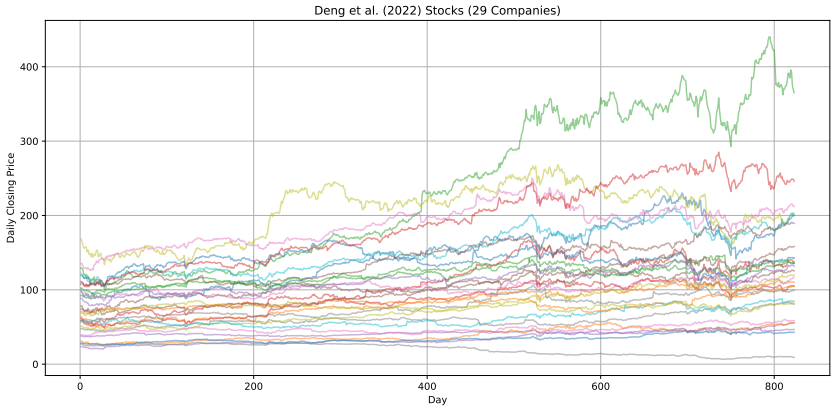

Figure 5: Daily closing prices from the dataset of Deng et al. (2022).

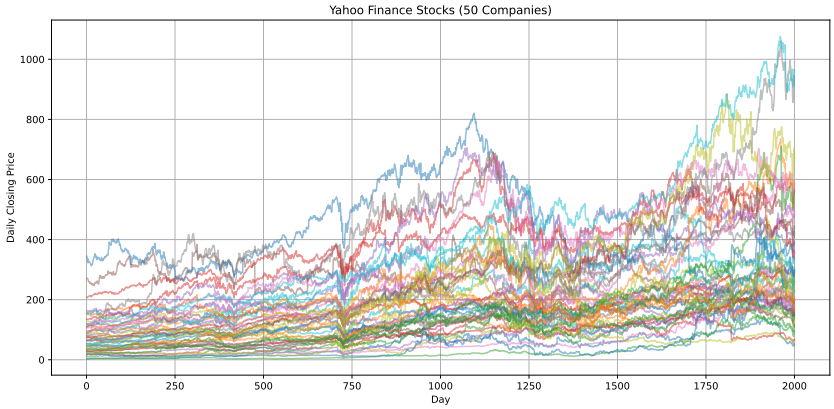

Figure 6: Daily closing prices obtained from Yahoo Finance.

COVID–NMA Clinical Dataset Construction. For the clinical benchmark based on the public COVID-NMA pharmacological RCT database (Boutron et al., 2025),[7]. we use only released arm-level counts and metadata and discretize time into calendar months, assigning each trial arm to its `Start_date` (falling back to `Pub_date_online`); rows with invalid or missing dates are discarded. We deterministically map case-insensitive rules on treatment type into 13 actions: *Antivirals (any)*, *Anti–inflammatory (steroids/NSAIDs)*, *Interleukin inhibitors*, *Monoclonal antibodies (other)*, *Immunoglobulins/Plasma*, *Antithrombotics*, *Antimicrobials*, *Immunomodulators (non–steroid)*, *Kinase inhibitors*, *Metabolic agents*, *Supportive care*, *Control/Standard care*, and *Other/Unknown*. At the bucket–month level we compute two endpoints: (i) *Clinical Improvement @ D28* (successes = number improved; trials = reported denominator, or baseline N if missing) and (ii) *Survival @ D28* derived from mortality (successes = denominator − deaths). To form a long non–stationary sequence, we adopt a union construction: for each $(k, t,$ endpoint) bin we emit exactly $s_{k,t}$ ones and $n_{k,t} - s_{k,t}$ zeros and concatenate all bins in a fixed order (month, `clinD28`, `mortD28`, bucket). The sequence is fully deterministic; in our run it comprises $T \approx 7.4 \times 10^4$ rounds with 13 actions.

[7]Data available at: https://doi.org/10.5281/zenodo.14965887

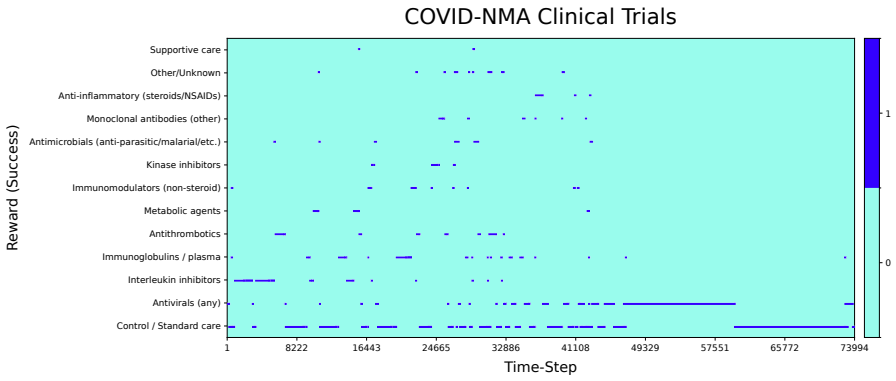

Figure 7: Raw rewards for COVID-NMA Clinical dataset (Boutron et al., 2025).

Yahoo! R6A Dataset Construction. For the NS bandit benchmark based on the Yahoo! R6A click log dataset[8], we follow the main procedure provided in Cao et al. (2019a); Zhou et al. (2020). We merge ten consecutive days of logs and we group the data by article ID and compute smoothed click-through rates (CTRs) using centered rolling averages over a 100-round window. This generates a time series of empirical CTRs for each article. We segment the dataset into ten distinct subperiods (each spanning half a day), filtering out actions with missing data or high noise. We further select a set of common actions present in all segments to ensure consistent tracking. We average CTRs within each subperiod and smoothing small deviations below a threshold 0.005. We stack selected actions across multiple days into a single $K \times T$ matrix, where K is the number of valid actions and T the compressed time horizon. To reduce spurious noise and compress the time scale, we apply local smoothing. Finally, we apply post-processing filters to remove (i) globally high-value actions (outliers with inflated CTRs), and (ii) actions that persist as best for too many segments.

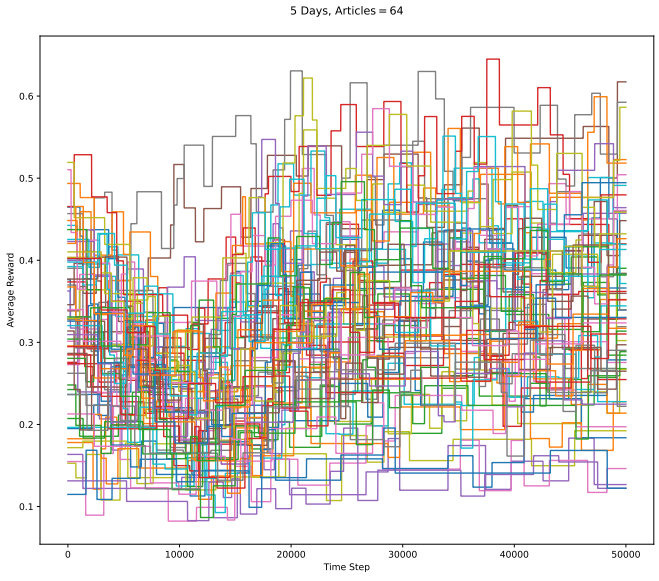

Figure 8: Mean rewards for the Yahoo! R6A dataset.

[8] Yahoo! Front Page Today Module User Click Log Dataset: https://webscope.sandbox.yahoo.com.

Yahoo! R6B Dataset Construction. We follow a two-stage pipeline tailored to the Yahoo! R6B logs.[8] *Stage 1 (action vocabulary):* we scan the logs to count displays and clicks per article and select the top items using the click-through rate with a minimum display threshold of 2, yielding a fixed action set with mapping id$\mapsto k \in \{0, \dots, K-1\}$ with $K = 51$, chosen on the same window as the evaluation files. *Stage 2 (replay log):* we reprocess the files and, for each round t, form a feature vector $\mathbf{x}_t$ from the given features, restrict the candidate set to the Top–K vocabulary to obtain $\mathcal{A}_t$, locate the displayed item's index $j_t^\star \in \{0, \dots, |\mathcal{A}_t| - 1\}$, and record the binary click $X_t \in \{0, 1\}$; we drop rounds where the displayed item lies outside Top–K or $|\mathcal{A}_t| < 2$. To increase coverage at a fixed horizon $T = 50000$, days are merged in a round-robin fashion before truncation. The resulting dataset stores $\{\mathbf{x}_t, \mathcal{A}_t, j_t^\star, r_t, t_t\}_{t=1}^T$. For offline *replay* evaluation, a policy π observes $(\mathbf{x}_t, \mathcal{A}_t)$ and proposes $A_t \in \{0, \dots, |\mathcal{A}_t| - 1\}$; we credit the outcome only when $a_t = j_t^\star$, and report cumulative reward $C_T = \sum_{t=1}^T \mathbf{1}\{a_t = j_t^\star\} r_t$.

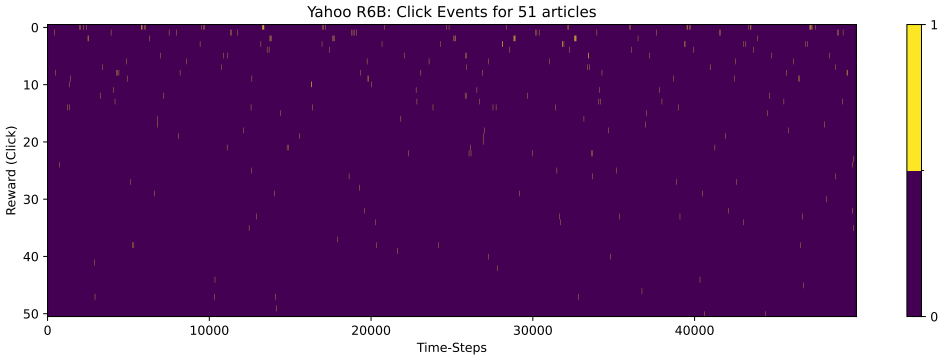

Figure 9: Rewards for the Yahoo! R6B dataset.

Sensor Correlation Data Construction. We use the Bioliq dataset from Komiyama et al. (2024), comprising a week of readings from 20 power plant sensors. Following their setup, we construct an NS-SCB environment with 190 actions: the reward is 1 if the last 1000 measurements exceed 2.04, and 0 otherwise. Evaluation is based on cumulative reward. Data available at `https://github.com/edouardfouche/G-NS-MAB/tree/master/data`.

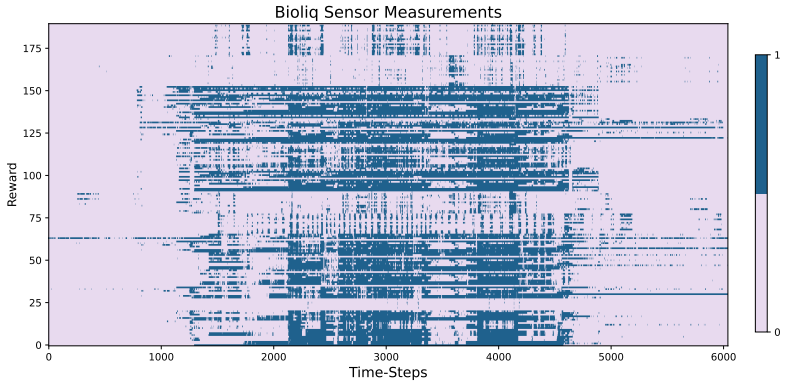

Figure 10: Raw rewards obtained from the Bioliq dataset (Komiyama et al., 2024).

Ad Recommendation Data Construction. We evaluate on the Zozo environment, a real-world ad recommender system from Saito et al. (2021), using the preprocessed dataset of Komiyama et al. (2024). We construct an NS-GLB environment with all 80 ads (unlike their 10-action setup), assigning reward 1 to any ad clicked within one second, and 0 otherwise. Evaluation is based on cumulative reward. Data available at `https://github.com/edouardfouche/G-NS-MAB/tree/master/data`.

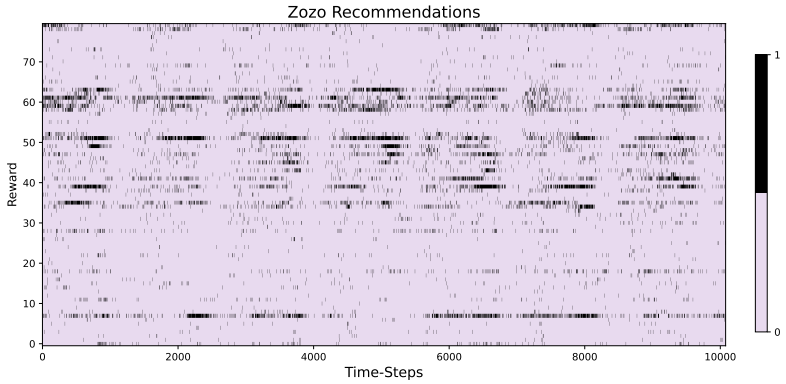

Figure 11: Raw rewards obtained from the Zozo dataset (Komiyama et al., 2024).

Live Traffic Data Construction. We construct a NS bandit environment based on the Criteo live traffic dataset (Diemert et al., 2017), following the preprocessing approach of Russac et al. (2019) but modeling the problem as an NS-GLB rather than an NS-LB. Specifically, the dataset includes banners shown to users, associated contextual variables, and whether each banner was clicked. We retain the categorical variables `cat1` through `cat9`, along with `campaign`, which uniquely identifies each campaign. These categorical features are one-hot encoded, and a dimensionality reduction via Singular Value Decomposition selects 50 resulting features. The parameter vector $\theta^\star$ is estimated using logistic regression. Rewards are then generated from this regression model with added Gaussian noise of variance $\sigma^2 = 0.01$. Unlike Russac et al. (2019), in which the authors employ a single change, we introduce shifts in θ^* via a geometric change-point model with parameter $\xi = 0.8$, chaging 60% of the θ^* coordinates at each time-step to $-\theta^*$ and extend the horizon to $T = 50000$.

A.4 Hardware Specifications

All experiments were employed on a desktop using an Intel(R) Xeon(R) W-2245 processor with 32 GB RAM. Each experiment had a total runtime below one hour.

B Theoretical Results

B.1 Regret Bounds of DAL in Piecewise Stationary Environments

As discussed in Section 4.2 of the paper, using Corollary 4.5, we can select different algorithms as input for DAL to attain or improve the state-of-the-art regret bounds in PS environments. Combining DAL with different bandit algorithms leads to the results in Table 1. It is evident that DAL matches the state-of-the-art regret bounds in PS-LBs, PS-GLBs and PS-CBs, and DAL improves the best known bounds in the PS-SCB and PS-KB settings. Note that for PS-SCBs, the strongest result corresponds to the prior-based WeightUCB Wang et al. (2023). As demonstrated in the final columns of the table, the order-wise dependence on problem parameters from the stationary setting seamlessly transfers to the PS setting without degradation.

B.2 Proof of Proposition 4.1

In the NS-PB setting, the reward at time t is given by $f_t(a) = \mu(\langle \theta_t, a \rangle)$ for all $a \in \mathcal{A}$, where μ is injective and $\theta_t \in \mathbb{R}^d$. To detect any changes in θ_t, it suffices to detect changes in the values $\langle \theta_t, a \rangle$ for a suitable set of actions.

Since μ is injective, each observation $y_{t,i} = \mu(\langle \theta_t, a_i \rangle)$ can be inverted to recover the inner product:

$$\langle \theta_t, a_i \rangle = \mu^{-1}(y_{t,i}).$$

Hence, observing $y_{t,i}$ is equivalent to observing $\langle \theta_t, a_i \rangle$.

Table 1: Regret bound comparison of algorithms for PS bandits, under the assumption of Huang et al. (2025). "†" denotes settings with finite number of actions, while MASTER, ADA-OPKB and SCB-WeightUCB also recover the appropriate bounds in this setting. "•" indicates prior-based algorithms.

PS Setting	Non-Stationary Algorithm	NS Algorithm Regret Bound in $\tilde{\mathcal{O}}(\cdot)$	DAL Input Regret Bound in $\tilde{\mathcal{O}}(\cdot)$
PS-LB	MASTER (Wei & Luo, 2021) + LinUCB	$d\sqrt{TN_T}$	-
	ADA-OPKB (Hong et al., 2023)	$d\sqrt{N_T T}$	-
	DAL (ours) + LinUCB (Abbasi-yadkori et al., 2011)	$d\sqrt{N_T T}$	$d\sqrt{T}$
	DAL (ours) + LinTS (Agrawal & Goyal, 2013)	$d^{3/2}\sqrt{N_T T}$	$d^{3/2}\sqrt{T}$
	DAL (ours) + PEGE† (Lattimore & Szepesvári, 2020)	$\sqrt{dN_T T}$	$\sqrt{dT}$
PS-GLB	MASTER (Wei & Luo, 2021) + GLM-UCB	$d\sqrt{N_T T}$	-
	DAL (ours) + GLM-UCB (Filippi et al., 2010)	$d\sqrt{N_T T}$	$d\sqrt{T}$
	DAL (ours) + GLM-TSL (Kveton et al., 2020)	$d^{3/2}\sqrt{N_T T}$	$d^{3/2}\sqrt{T}$
	DAL (ours) + SupCB-GLM† (Li et al., 2017)	$\sqrt{dN_T T}$	$\sqrt{dT}$
PS-SCB	SCB-WeightUCB• (Wang et al., 2023)	$d^{2/3}T^{2/3}N_T^{1/3}$	–
	DAL (ours) + OFU-ECOLog (Faury et al., 2022)	$d\sqrt{N_T T}$	$d\sqrt{T}$
	DAL (ours) + OFUL-MLogB (Zhang & Sugiyama, 2023)	$d\sqrt{N_T T}$	$d\sqrt{T}$
	DAL (ours) + OFUGLB (Lee et al., 2024)	$d\sqrt{N_T T}$	$d\sqrt{T}$
PS-KB	MASTER (Wei & Luo, 2021) + GPUCB	$\gamma_T\sqrt{N_T T}$	-
	ADA-OPKB (Hong et al., 2023)	$\sqrt{d\gamma_T N_T T}$	-
	DAL (ours) + GPUCB (Chowdhury & Gopalan, 2017)	$\gamma_T\sqrt{N_T T}$	$\gamma_T\sqrt{T}$
	DAL (ours) + REDS (Salgia et al., 2024)	$\sqrt{\gamma_T N_T T}$	$\sqrt{\gamma_T T}$
PS-CB	MASTER (Wei & Luo, 2021) + ILTCB	$\sqrt{\lvert\mathcal{A}\rvert N_T T\log\lvert\Pi\rvert}$	-
	ADA-ILTCB+ (Chen et al., 2019)	$\sqrt{\lvert\mathcal{A}\rvert N_T T\log\lvert\Pi\rvert}$	-
	DAL (ours) + ILTCB (Agarwal et al., 2014)	$\sqrt{\lvert\mathcal{A}\rvert N_T T\log\lvert\Pi\rvert}$	$\sqrt{\lvert\mathcal{A}\rvert T\log\lvert\Pi\rvert}$
	DAL (ours) + SquareCB (Foster & Rakhlin, 2020)	$\sqrt{\lvert\mathcal{A}\rvert N_T T\log\lvert\Pi\rvert}$	$\sqrt{\lvert\mathcal{A}\rvert T\log\lvert\Pi\rvert}$

Suppose that $\mathcal{A}_{\mathrm{e}} \subseteq \mathcal{A}$ is the maximal linearly independent subset of $\mathcal{A}$. Then, the vector θ_t is uniquely determined by the inner products $\langle \theta_t, a\rangle$ for $a \in \mathcal{A}_{\mathrm{e}}$. Therefore, any change in θ_t results in a detectable change in the vector of observations $(y_{t,i})_{a_i\in\mathcal{A}_{\mathrm{e}}}$, meaning that $\mathcal{A}_{\mathrm{e}}$ can be taken to be any maximal linearly independent subset of $\mathcal{A}$, with $|\mathcal{A}_e| \le d$.

B.3 Proof of Proposition 4.2

In this subsection, we establish the construction of $\mathcal{A}_{\mathrm{e}}$ in the NS-KB setting. According to Lemma 5 from De Freitas et al. (2012), we have that every $f \in H_k$ with $\|f\|_{H_k} \le B$ is Lipschitz continuous, satisfying the following,

$$|f(x) - f(y)| \;\le\; B\,L_u\,\|x-y\|_2,\ \forall\, x, y \in \mathcal{A}, \quad \text{where } L_u := \sup_{z\in D}\max_{i,j\le d}\Big[\frac{\partial^2 k(p,q)}{\partial p_i\,\partial q_j}\Big]^{1/2}_{p=q=z}.$$

Recall that $\mathcal{V}_T$ corresponds to the set of centers of the balls of an arbitrary δ_T-cover of $\mathcal{A} \subseteq [0,R]^d$. Let $[a]_{\mathrm{e}}$ denote the action in $\mathcal{V}_T$ that is the closest to $a \in \mathcal{A}$, i.e., $[a]_{\mathrm{e}} = \mathrm{argmin}_{x\in\mathcal{P}_T}\|a-x\|_2$. Then, we can leverage the Lipschitz property of functions in the RKHS to obtain the following upper bound: For any $a \in \mathcal{A}$ and $f \in H_k$ with $\|f\|_{H_k} \le B$,

$$|f(a) - f([a]_{\mathrm{e}})| \overset{(a)}{\le} BL_u\|a-[a]_{\mathrm{e}}\|_2 \overset{(b)}{\le} BL_u\delta_T. \tag{1}$$

Step (a) follows from the Lipschitz property in Lemma 5 of De Freitas et al. (2012), and step (b) results from the definition of a δ_T-cover. Then, for any arbitrary functions f and f' in H_k with $\|f\|_{H_k}, \|f'\|_{H_k} \le B$ and action $\tilde{a} \in \mathcal{A}$, we have

$$\begin{aligned}|f([\tilde{a}]_e) - f'([\tilde{a}]_e)| &\ge |f(\tilde{a}) - f'(\tilde{a})| - |f(\tilde{a}) - f([\tilde{a}]_{\mathrm{e}})| - |f'(\tilde{a}) - f'([\tilde{a}]_{\mathrm{e}})| \\ &\overset{(a)}{\ge} |f(\tilde{a}) - f'(\tilde{a})| - 2BL_u\delta_T \overset{(b)}{>} 0\end{aligned}$$

where step (a) is due to equation 1, and step (b) is due to the assumption in Proposition 4.2. This indicates that the value of the reward function at $[\tilde{a}]_{\mathrm{e}}$ must change by a non-zero amount. Thus, one

can use observations from action $[\tilde{a}]_{\mathrm{e}}$ in order to deduce whether the reward function has changed its value in action $\tilde{a}$. In addition, by the upper bound on the covering number, the cardinality of $\mathcal{V}_T$ is upper bounded by $\lceil\sqrt{d}R/2\delta_T\rceil^d$.

B.4 Sketch of Proof of Theorem 4.3

For PS-PBs and PS-KBs, the proof of Theorem 4.3 follows exactly the same as those of Theorem 1 and Corollary 1 in Huang et al. (2025), with the number of arms replaced by N_{e}, due to the different number of actions in the covering set. For completeness, we provide a proof sketch of Theorem 4.3: First, we partition the regret into two cases. If no false alarm occurs and all changes are detected within a short delay, we can separate the regret into three components: the regret due to forced exploration, the regret during the short detection (restart) delay after changes, and the regret incurred by the stationary bandit algorithm after the change is detected. If not, we use a crude linear bound to bound the regret and show that the probability of false alarm and that of late detection are low, which ensures that the regret due to detection failure is small.

For PS-CBs, the proof of Theorem 4.3 follows similar to those of NS-KBs and NS-PBs, except that the definition of successful detection events should be modified as follows: For any $k \in [N_T]$,

$$\mathcal{M}_k \coloneqq \left\{\forall l \in [k-1], c \in \mathcal{C}, i \in [N_{\mathrm{e}}] : \sum_{t=\tau_{l-1}}^{\nu_l-1} \mathbf{1}\left\{C_t = c, (t-\tau_{l-1}-1) \bmod \left\lceil \frac{N_{\mathrm{e}}}{\alpha_l}\right\rceil = i-1\right\} \geq m_{\mathcal{D}}\right\}, \quad (2)$$

$$\mathcal{L}_k \coloneqq \left\{\forall l \in [k-1], c \in \mathcal{C}, i \in [N_{\mathrm{e}}] : \sum_{t=\nu_l}^{\nu_l+\ell_l-1} \mathbf{1}\left\{C_t = c, (t-\tau_{l-1}-1) \bmod \left\lceil \frac{N_{\mathrm{e}}}{\alpha_l}\right\rceil = i-1\right\} \geq \ell_{\mathcal{D}}\right\}, \quad (3)$$

$$\mathcal{E}_k \coloneqq \{\forall\, l \in [k-1] : \tau_l \in \{\nu_l, \ldots, \nu_l + \ell_l - 1\}\} \cap \mathcal{M}_k \cap \mathcal{L}_k, \text{ and} \quad (4)$$

$$\mathcal{G}_k \coloneqq \mathcal{E}_k \cap \{\tau_k \geq \nu_k\}, \quad (5)$$

where τ_l denotes the l^{th} detection point, and recall that $\ell_l = \lceil N_{\mathrm{e}}/\alpha_l\rceil\lceil 2\log(T) + \ell_{\mathcal{D}}/s\rceil$ with $s = \min_{c\in\mathcal{C}, t\in[T]:\mathcal{P}_t(c)>0} \mathcal{P}_t(c)$ for $l \in [N_T]$. For convenience in notations, we denote $\tau_0 = 0$ and $[0] = \emptyset$. In the new definition of the detection success event, the number of reward samples for each context-action pair obtained from forced exploration is lower bounded, ensuring enough pre- and post-change samples for change detection. With these new successful detection events, the proof follows exactly the same as those of Theorem 1 in Huang et al. (2025), except for a new modified union bound on the detection failure event $\mathcal{G}_k^c$ derived as follows: For any $k \in \{2, \ldots, N_T\}$,

$$\begin{aligned}
&\mathbb{P}(\mathcal{G}_k^c) \\
&= \mathbb{P}\left(\{\exists\, l \in [k-1],\ \tau_l \notin \{\nu_l, \ldots, \nu_l + \ell_l - 1\}\} \cup \{\tau_k \leq \nu_k\} \cup \mathcal{M}_k^c \cup \mathcal{L}_k^c\right) \\
&\overset{(a)}{\leq} \sum_{l=1}^{k-1} \mathbb{P}\left(\tau_l \notin \{\nu_s, \ldots, \nu_l + \ell_l - 1\}, \mathcal{E}_{l-1}\right) + \mathbb{P}\left(\tau_k \leq \nu_k, \mathcal{E}_{k-1}\right) + \mathbb{P}\left(\mathcal{M}_k^c, \mathcal{E}_{k-1}\right) + \mathbb{P}\left(\mathcal{L}_k^c, \mathcal{E}_{k-1}\right) \\
&= \sum_{l=1}^{k-1} \mathbb{P}\left(\mathcal{E}_{l-1}\right) \mathbb{P}\left(\tau_l \notin \{\nu_l, \ldots, \nu_l + \ell_l - 1\} \middle| \mathcal{E}_{l-1}\right) \\
&\quad + \mathbb{P}\left(\mathcal{E}_{k-1}\right) \mathbb{P}\left(\tau_k \leq \nu_k \middle| \mathcal{E}_{k-1}\right) + \mathbb{P}\left(\mathcal{E}_{k-1}\right) \mathbb{P}\left(\mathcal{M}_k^c \middle| \mathcal{E}_{k-1}\right) + \mathbb{P}\left(\mathcal{E}_{k-1}\right) \mathbb{P}\left(\mathcal{L}_k^c \middle| \mathcal{E}_{k-1}\right) \\
&\overset{(b)}{\leq} \sum_{l=1}^{k-1} \mathbb{P}\left(\tau_l \notin \{\nu_l, \ldots, \nu_l + \ell_l - 1\} \middle| \mathcal{E}_{l-1}\right) + \mathbb{P}\left(\tau_k \leq \nu_k \middle| \mathcal{E}_{k-1}\right) + \mathbb{P}\left(\mathcal{M}_k^c \middle| \mathcal{E}_{k-1}\right) + \mathbb{P}\left(\mathcal{L}_k^c \middle| \mathcal{E}_{k-1}\right) \\
&= \sum_{l=1}^{k} \underbrace{\mathbb{P}\left(\tau_l < \nu_l \middle| \mathcal{E}_{l-1}\right)}_{\Phi_1} + \sum_{l=1}^{k-1} \underbrace{\mathbb{P}\left(\tau_l \geq \nu_l + \ell_l \middle| \mathcal{E}_{l-1}\right)}_{\Phi_2} + \underbrace{\mathbb{P}\left(\mathcal{M}_k^c \middle| \mathcal{E}_{k-1}\right)}_{\Phi_3} + \underbrace{\mathbb{P}\left(\mathcal{L}_k^c \middle| \mathcal{E}_{k-1}\right)}_{\Phi_4} \qquad (6)
\end{aligned}$$

where step (a) is due to the union bound, and step (b) is owing to the fact that $\mathbb{P}\left(\mathcal{E}_{k-1}\right) \leq 1$. The upper bounds on Φ_1 and Φ_2 follows the same as those in Huang et al. (2025). For terms Φ_3 and Φ_4,

we can upper bound them as follows: For any $k \in \{2, \dots, N_T\}$,

$$
\begin{aligned}
&\mathbb{P}\left(\mathcal{M}_k^c \middle| \mathcal{E}_{k-1}\right) = \\
&\mathbb{P}\left(\exists\, l<k, c \in \mathcal{C}, i \in [N_{\mathrm{e}}] : \sum_{t=\tau_{l-1}}^{\nu_l - 1} \mathbf{1}\left\{C_t = c, (t-\tau_{l-1}-1) \bmod \left\lceil \frac{N_{\mathrm{e}}}{\alpha_l} \right\rceil = i-1\right\} < m_{\mathcal{D}} \middle| \mathcal{E}_{k-1}\right) \\
&= \mathbb{P}\left(\exists c \in \mathcal{C}, i \in [N_{\mathrm{e}}] : \sum_{t=\tau_{k-2}}^{\nu_{k-1} - 1} \mathbf{1}\{C_t = c, (t-\tau_{l-1}-1) \mod \lceil N_{\mathrm{e}}/\alpha_l \rceil = i-1\} < m_{\mathcal{D}} \middle| \mathcal{E}_{k-1}\right) \\
&\overset{(a)}{\leq} \sum_{c \in \mathcal{C}} \sum_{i \in [N_e]} \mathbb{P}\left(\sum_{t=\tau_{k-2}}^{\nu_{k-1} - 1} \mathbf{1}\{C_t = c, (t-\tau_{l-1}-1) \mod \lceil N_{\mathrm{e}}/\alpha_l \rceil = i-1\} < m_{\mathcal{D}} \middle| \mathcal{E}_{k-1}\right) \\
&\overset{(b)}{\leq} \sum_{c \in \mathcal{C}} \sum_{i \in [N_e]} \exp\left(-2 m_{\mathcal{D}} (\log T)^2\right) \\
&\overset{(c)}{\leq} |\mathcal{C}| N_{\mathrm{e}} T^{-2}
\end{aligned} \tag{7}
$$

where step (a) stems from the union bound. In step (b), since $\mathbf{1}\{C_t = c\}$ are i.i.d. $1/4$-sub-Gaussian for $t > \tau_{k-2}$ given $\mathcal{E}_{k-1}$, we can apply Chernoff bound to obtain the inequality. In step (c), we simply use the fact that $m_{\mathcal{D}} > 0$ and $\log T > 1$ for $T \geq 3$. Note that $\mathcal{M}_1^c = \emptyset$ and thus $\Pr(\mathcal{M}_1^c | \mathcal{E}_0) = 0$. The derivation of the upper bound of Φ_4 follows the exact same steps as in the ones in equation 7. The rest of the proof follows exactly the same as those in Huang et al. (2025).

B.5 Proof of Corollary 4.4

In PS-PBs, $N_{\mathrm{e}} = d$, $p \geq 1/2$, and $q = r = 0$. Thus, $R_T = \tilde{\mathcal{O}}(\sqrt{dN_T T} + d^p \sqrt{N_T T}) = \tilde{\mathcal{O}}(d^p \gamma_T^q (\log |\Pi|)^r \sqrt{N_T T})$.

In PS-KBs, $q \geq 1/2$, $p \geq 0$ and $r = 0$. We can upper bound N_{e} using the fact that $|\mathcal{V}_T| \leq \lceil \sqrt{d} R / 2\delta_T \rceil^d$. Thus, $N_e \leq \lceil C \gamma_T^{2q/d} \rceil^d$ with $\delta_T = \frac{R d^{1/2-2p/d}}{2(C\gamma_T^{2q})^{1/d}}$ and $R_T = \tilde{\mathcal{O}}((d^{2p} \gamma_T^{2q} N_T T)^{1/2} + d^p \gamma_T^q \sqrt{N_T T}) = \tilde{\mathcal{O}}(d^p \gamma_T^q (\log |\Pi|)^r \sqrt{N_T T})$.

We emphasize that when the number of action is smaller than the covering number, i.e., $|\mathcal{A}| < \lceil C \gamma_T^{2q/d} \rceil^d \leq \gamma_T$, then we can set $\mathcal{A}_{\mathrm{e}}$ to be the entire action set $\mathcal{A}$. In this case, $N_{\mathrm{e}} < \gamma_T$, guaranteeing order-optimal regret.

In PS-CBs, $N_{\mathrm{e}} \leq |\mathcal{A}|$, $r \geq 1/2$, $p = q = 0$, and $|\Pi| = |\mathcal{A}|^{|\mathcal{C}|}$. Thus, $R_T = \tilde{\mathcal{O}}((|\mathcal{A}| \log |\Pi|)^r \sqrt{N_T T} + \sqrt{|\mathcal{C}||\mathcal{A}| N_T T}) = \tilde{\mathcal{O}}(d^p \gamma_T^q (|\mathcal{A}| \log |\Pi|)^r \sqrt{N_T T})$.