# ANACONDA: An Improved Dynamic Regret Algorithm for Adaptive Non-Stationary Dueling Bandits

**Thomas Kleine Buening**
University of Oslo

**Aadirupa Saha***
TTI Chicago

## Abstract

We study the problem of non-stationary dueling bandits and provide the first adaptive dynamic regret algorithm for this problem. The only two existing attempts in this line of work fall short across multiple dimensions, including pessimistic measures of non-stationary complexity and non-adaptive parameter tuning that requires knowledge of the number of preference changes. We develop an elimination-based rescheduling algorithm to overcome these shortcomings and show a near-optimal $\tilde{O}(\sqrt{S^{\texttt{CW}}T})$ dynamic regret bound, where $S^{\texttt{CW}}$ is the number of times the Condorcet winner changes in $T$ rounds. This yields the first near-optimal dynamic regret bound for unknown $S^{\texttt{CW}}$. We further study other related notions of non-stationarity for which we also prove near-optimal dynamic regret guarantees under additional assumptions on the preference model.

## 1 Introduction

Multi-Armed Bandits (MABs) (Thompson, 1933; Robbins, 1952; Lattimore and Szepesvári, 2018) are a well-studied online learning framework, which can be used to model online decision-making under uncertainty. Due to its exploration-exploitation tradeoff, the MAB framework is able to model situations such as clinical trials or job scheduling, where the goal is to keep selecting the 'best item' in hindsight by sequentially querying one item at a time and subsequently observing a noisy reward feedback for the queried arm (Auer et al., 2002; Agrawal and Goyal, 2012; Bubeck et al., 2012).

The MAB framework has been studied and generalized to different settings, among which a popular variant known as Dueling Bandits (DB) has gained much attention in the machine learning community over the last two decades (Yue et al., 2012; Zoghi et al., 2014b, 2015; Wu and Liu, 2016). Dueling bandits are a preference-based variant of multi-armed bandits in which every round the learner selects a pair of items (or arms) whereupon a noisy preference between the two items is observed. Such a model is particularly useful in applications where direct numerical feedback is unavailable, but observed feedback or behavior implies a preference of one item over the other. For instance, the dueling bandit framework can be used for search engine optimization through interleaved comparisons (Radlinski and Craswell, 2013; Hofmann et al., 2011).

In the classical stochastic dueling bandit problem, it is assumed that the underlying preferences between items remain fixed over time. However, in many applications, this assumed stationarity of preferences is likely to be violated. For example, user preferences over movies may change depending on the season or other external influences. Despite its strong practical motivation, regret minimization in non-stationary dueling bandits has only recently been studied for the fist time (Saha and Gupta, 2022; Kolpaczki et al., 2022). In contrast to the classical stochastic (Yue et al., 2012; Zoghi et al., 2014a; Bengs et al., 2021) and adversarial (Gajane et al., 2015; Saha et al., 2021; Saha and Gaillard, 2022) dueling bandit problem, which measures performance in terms of static regret w.r.t. a fixed benchmark (or best item in hindsight), in non-stationary dueling bandits we consider the stronger *dynamic regret*, which compares the algorithm's selection against a dynamic benchmark every round.

In general, the achievable dynamic regret depends on the amount of non-stationarity in the environment. Here, prior work (Saha and Gupta, 2022; Kolpaczki et al., 2022) studied the number of changes in the preference matrix as a measure of non-stationary complexity. While the number of such preference switches indeed relates to the hardness of the problem, it is, however, a pessimistic measure of non-stationarity. For example, a change in the preference between two widely suboptimal arms or a minor change in the preference matrix under which the optimal arm remains optimal should not significantly impact our ability to achieve low dynamic regret. To this end, one question that we aim to address in this paper for the non-stationary dueling bandit problem is:

Proceedings of the 26$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. 
*Author is currently with Apple ML Research, USA.

**Q.1:** *Can we guarantee low dynamic regret for stronger and more meaningful notions of non-stationarity?*

Moreover, prior work on non-stationary dueling bandits (Saha and Gupta, 2022; Kolpaczki et al., 2022) assumes knowledge of the non-stationary complexity, i.e. prior knowledge of the total number of preference switches (or total variation), which is a very impractical assumption to impose. The second question we then address is:

**Q.2:** *Can we achieve near-optimal dynamic regret in non-stationary dueling bandits adaptively, without knowledge of the underlying non-stationary complexity?*

### 1.1 Our Contributions

We answer these two questions affirmatively. Our main contribution is an algorithm, termed ANACONDA, that adaptively achieves near-optimal regret with respect to the number of 'best arm' switches—a measure that is sensitive only to the variations of the best arms in the preference sequence and indifferent to any other 'background noise' due to suboptimal arms. More precisely, our contributions can be listed as follows:

• **Comparing Different Concepts of Non-Stationarity in Dueling Bandits.** We first give an overview over different notions of non-stationarity measures for dueling bandits and analyze their inter-dependencies towards a better understanding of the implications of one to another (Section 2.2).

• **Proposing Stronger and More Meaningful Notions of Non-Stationarity (towards Q.1).** We propose three new measures of non-stationarity for dueling bandits: (i) $S^{\texttt{CW}}$, which counts the number of times the Condorcet winner changes; (ii) $\tilde{V}$, which measures the variation in the winning probabilities of the Condorcet winner; and (iii) $\tilde{S}^{\texttt{CW}}$, which counts only the 'significant variations' in the Condorcet winner (Section 2.2). The novelty of our proposed non-stationarity measures lies in capturing only the non-stationarity observed for the 'best arms' of the preference sequence. They remain unaffected by any changes in the suboptimal arms and are thus stronger notions of non-stationarity than the previously studied preference shifts, $S^{\texttt{P}}$, or total variation, $V$. In particular, we show that $\tilde{S}^{\texttt{CW}} \leq S^{\texttt{CW}} \leq S^{\texttt{P}}$ and $\tilde{V} \leq V$ justifying the strength of our proposed non-stationarity measures.

• **Adaptive Algorithm (towards Q.2).** Besides using weaker notions of non-stationary complexity, another drawback of the existing non-stationary dueling bandit works are, in order to achieve near-optimal dynamic regret, they require exact knowledge of the non-stationarity present in the environment (e.g., $S^{\texttt{P}}$ or $V$), which are of course expected to be unknown to the system/algorithm designed ahead of time. Our next main contribution lies in designing an adaptive algorithm (ANACONDA, Algorithm 1), that does not require knowledge of any underlying non-stationary complexity—it can adapt to any unknown number of best arm switches $S^{\texttt{CW}}$ and yields a near-optimal regret bound of $\tilde{O}\big(\sqrt{S^{\texttt{CW}}T}\big)$ (Theorem 3.1, Section 3).[1]

• **Improved and (Near-)Optimal Dynamic Regret Bounds.** Owing to the fact that $S^{\texttt{CW}} \leq S^{\texttt{P}}$, our dynamic regret bounds can be much tighter compared to the previous results by Saha and Gupta (2022); Kolpaczki et al. (2022), who show a regret guarantee of $\tilde{O}\big(\sqrt{S^{\texttt{P}}T}\big)$ (Remark 2.1). Furthermore, our regret bound is also provably order optimal in $T$ and $S^{\texttt{CW}}$ as justified in Remark 3.1.

• **Better Guarantees for Structured Preferences.** In Section 5, we analyze a special class of preference matrices, which respect a type of transitivity, for which we can prove even stronger dynamic regret guarantees of $\tilde{O}\big(\sqrt{\tilde{S}^{\texttt{CW}}T}\big)$ in terms of Significant CW Switches $\tilde{S}^{\texttt{CW}}$ and $\tilde{O}\big(\tilde{V}^{1/3}T^{2/3}\big)$ in terms of Condorcet Winner Variation $\tilde{V}$. The optimality of these bounds is justified in Remark 5.2 and Remark 5.3.

### 1.2 Related Work

The non-stationary MAB problem has been extensively studied for various non-stationarity measures, e.g. total variation (Besbes et al., 2014, 2015), distribution switches (Garivier and Moulines, 2011; Allesiardo et al., 2017; Auer et al., 2019), or best arm switches (Abbasi-Yadkori et al., 2022; Suk and Kpotufe, 2022). Moreover, its study has been extended to more complex setups such as linear bandits (Russac et al., 2019, 2020) and contextual MAB (Luo et al., 2018; Chen et al., 2019; Wu et al., 2018). We will particularly take inspiration from the recent advances of Auer et al. (2019); Abbasi-Yadkori et al. (2022); Suk and Kpotufe (2022) that were able to achieve near-optimal dynamic regret rates without knowledge of the number of distribution (or best arm) changes.

While the non-stationary MAB problem has seen much attention in recent years, its DB counterpart remains widely unexplored. The only two earlier works that address the non-stationary DB problem are by Saha and Gupta (2022) and Kolpaczki et al. (2022). However, these works are limited in a) the weakness of the analyzed non-stationarity measures, namely, general preference switches or total variation (see Section 2.2), and b) in the fact that their algorithms require knowledge of the total amount of non-stationarity in advance to achieve near-optimal dynamic regret. Here, we improve upon prior work by designing an adaptive algorithm ANACONDA that does not require knowledge of the amount of non-stationarity in the environment and achieves near-optimal dynamic regret w.r.t. the number of Condorcet winner switches, a stronger notion of non-stationarity than preference switches. A more comprehensive review of previous work that is related to the non-stationary MAB and DB problem is provided in Appendix C.

[1] Here, the $\tilde{O}$ notation hides logarithmic factors.

## 2 Problem Setting

We consider preference matrices $\mathbf{P} \in [0,1]^{K \times K}$ such that $P(a,b)$ indicates the probability of arm $a$ being preferred over arm $b$. Here, $\mathbf{P}$ satisfies $P(a,b) = 1 - P(b,a)$ and $P(a,a) = 0.5$ for all $a, b \in [K]$. We say that $a$ dominates $b$ and write $a \succ b$ if $P(a,b) > 0.5$, i.e. arm $a$ has a higher chance of winning over arm $b$ in a duel $(a,b)$. A well-studied notion of a good benchmark arm in dueling bandits is the *Condorcet Winner* (CW): Given any preference matrix $\mathbf{P} \in [0,1]^{K \times K}$, an arm $a^* \in [K]$ is called a Condorcet winner of $\mathbf{P}$ if $P(a^*, b) > 0.5$ for all $b \in [K] \setminus \{a^*\}$ (Bengs et al., 2021).

Note that any preference matrix with a total ordering over the arms invariably has a Condorcet winner. For example, assuming a total ordering $1 \succ 2 \succ \ldots \succ K$ implies that the Condorcet winner is arm 1. Any RUM-based preference matrix (Saha and Gopalan, 2019a, 2020; Soufiani et al., 2013), or more generally any $\mathbf{P}$ with stochastic transitivity (Yue and Joachims, 2009), always respects a total ordering. However, note that CW-based preference matrices consider a much bigger class of pairwise relations than total ordering. In general, a preference matrix might not have a Condorcet winner, which led to more general notions of benchmark arms in DB, such as the Borda winner (Saha et al., 2021), the Copeland winner Zoghi et al. (2015) or the von Neumann winner (Dudík et al., 2015; Saha and Krishnamurthy, 2022).

### 2.1 Non-Stationary Dueling Bandits (`NSt-DB`)

We consider a decision space of $K$ arms denoted by $[K]$. At each round $t \in [T]$, the task of the learner is to select a pair of actions $(a_t, b_t) \in [K] \times [K]$, upon which a preference feedback $o_t(a_t, b_t) \sim \text{Ber}(P_t(a_t, b_t))$ is revealed to the learner according to the underlying preference matrix $\mathbf{P}_t \in [0,1]^{K \times K}$, where the sequence of preferences $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_T$ is generated adversarially. For any such preference matrix $\mathbf{P}_t$, we denote by

$$\delta_t(a,b) \coloneqq P_t(a,b) - 1/2$$

the gap or preference-strength of arm $a$ over arm $b$ in round $t$. We here assume that every preference matrix $\mathbf{P}_t$ has a Condorcet winner, which we refer to by $a_t^*$.

**Static Regret in Dueling Bandits.** In classical (stochastic) dueling bandits, where it is assumed that $\mathbf{P}_1 = \ldots = \mathbf{P}_T = \mathbf{P}$ (for some fixed preference matrix $\mathbf{P}$), the performance of the learner is often measured w.r.t. the CW of $\mathbf{P}$, defined as the *static regret*:

$$\mathrm{R}(T) \coloneqq \sum_{t=1}^{T} \frac{\delta_t(a^*, a_t) + \delta_t(a^*, b_t)}{2},$$

where $a^*$ is the CW of $\mathbf{P}$. Note that here $\delta_t(a^*, a) = P_t(a^*, a) - 1/2$ essentially quantifies the net loss of arm $a$ against the fixed benchmark arm $a^*$.

However, regret with respect to any fixed benchmark (comparator arm) soon becomes meaningless when the underlying preference matrices are changing over time, since no single fixed arm may represent a reasonably 'good benchmark' over $T$ rounds. Consider the following simple motivating example:

**Example 2.1.** Let $K = 2$ and define

$$\mathbf{P}_1 = \begin{bmatrix} 0.5 & 1 \\ 0 & 0.5 \end{bmatrix}, \quad \mathbf{P}_2 = \begin{bmatrix} 0.5 & 0 \\ 1 & 0.5 \end{bmatrix}.$$

Now, assume a preference sequence such that $\mathbf{P}_t = \mathbf{P}_1$ for the first $\lfloor T/2 \rfloor$ rounds and $\mathbf{P}_t = \mathbf{P}_2$ for the last $\lceil T/2 \rceil$ rounds. We see that a policy that plays any of the two arms all $T$ rounds, e.g. $\pi_t = 1$ for all $t \in [T]$, has regret $O(1)$ against any fixed benchmark arm, as $\delta_t(1,2) = 1/2$ for the first $T/2$ rounds and $\delta_t(1,2) = -1/2$ for last $T/2$ rounds. However, against a *dynamic benchmark*, e.g. arm 1 for $t < T/2$ and arm 2 for $t \geq T/2$, any policy that plays a fixed arm all $T$ rounds suffers $O(T/2)$ regret (while suffering only constant regret against any fixed benchmark).

**Dynamic Regret in Dueling Bandits.** Drawing motivation from the above, we seek to formulate a more meaningful notion of dueling bandit regret, where the benchmark at every round is chosen dynamically based on $\mathbf{P}_t$. More precisely, letting $a_t^*$ be the CW of $\mathbf{P}_t$, we define the *dynamic regret*:

$$\mathrm{DR}(T) \coloneqq \sum_{t=1}^{T} \frac{\delta_t(a_t^*, a_t) + \delta_t(a_t^*, b_t)}{2}.$$

### 2.2 Measures of Non-Stationarity

Clearly, without any control over the amount of non-stationarity in the sequence $\{\mathbf{P}_t\}_{t \in [T]}$, it is impossible for any learner to achieve a sublinear $o(T)$ dynamic regret in the worst case. To see this, consider the matrices from Example 2.1 and note that for any choice of arms $(a_t, b_t)$, the adversary can choose a matrix so as to guarantee instantaneous regret of at least $1/2$. This consequently leads to linear regret for the learner, implying that to achieve sublinear dynamic regret, we need to restrict the adversary in terms of the total amount of non-stationarity it can induce in the sequence $\{\mathbf{P}_t\}_{t \in [T]}$. But what could be a good measure of non-stationarity? In this paper, we study several of these measures, which we will now formally introduce and put in relation to one another.

① **Preference Switches.** A non-stationarity measure that has been studied in the previous `NSt-DB` literature is the number of times $\mathbf{P}_t$ changes (Kolpaczki et al., 2022; Saha and Gupta, 2022):

$$S^{\mathbb{P}} \coloneqq \sum_{t=2}^{T} \mathbf{1}\{\mathbf{P}_t \neq \mathbf{P}_{t-1}\}.$$

However, $S^{\mathrm{P}}$ can be a quite pessimistic measure of non-stationarity, as changes in the preference between two sub-optimal arms or minor preference shifts that do not change the CW are counted toward $S^{\mathrm{P}}$, whereas they should not significantly affect the performance of a good learning algorithm.

② **Condorcet Winner Switches.** A stronger measure of non-stationarity is then the total number of Condorcet Winner Switches, i.e. the number of times the identity of $a_t^*$ changes:

$$S^{\mathrm{CW}} := \sum_{t=2}^{T} \mathbf{1}\{a_t^* \neq a_{t-1}^*\}.$$

**Remark 2.1** ($S^{\mathrm{P}}$ vs $S^{\mathrm{CW}}$)**.** *From the definition, we always have $S^{\mathrm{CW}} \leq S^{\mathrm{P}}$. In fact, it is easy to construct a scenario where $S^{\mathrm{CW}} \ll S^{\mathrm{P}}$: Assume $K = 3$ and consider the following two preference matrices*

$$\mathbf{P}_1 = \begin{bmatrix} 0.5 & 0.55 & 0.55 \\ 0.45 & 0.5 & 1 \\ 0.45 & 0 & 0.5 \end{bmatrix}, \ \mathbf{P}_2 = \begin{bmatrix} 0.5 & 0.55 & 0.55 \\ 0.45 & 0.5 & 0 \\ 0.45 & 1 & 0.5 \end{bmatrix},$$

*and a preference sequence $\{\mathbf{P}_t\}_{t\in[T]}$ such that $\mathbf{P}_t = \mathbf{P}_1$ when $t$ is odd and $\mathbf{P}_t = \mathbf{P}_2$ otherwise. Then, $S^{\mathrm{CW}} = 0$ (since $1$ is the CW for all $t$), whereas $S^{\mathrm{P}} = T$.*

③ **Significant Condorcet Winner Switches.** Recently, Suk and Kpotufe (2022) proposed a new (and strong) notion of non-stationarity in multi-armed bandits, called *Significant Shifts*, that aims to account only for severe distribution shifts and comprises previous complexity measures. We can define a similar concept for dueling bandits: Let $\nu_0 := 1$ and define $\nu_{i+1}$ recursively as the first round in $[\nu_i, T)$ such that for all arms $a \in [K]$ there exist rounds $\nu_i \leq s_1 < s_2 < \nu_{i+1}$ such that $\sum_{t=s_1}^{s_2} \delta_t(a_t^*, a) \geq \sqrt{K(s_2 - s_1)}$. Let $\tilde{S}^{\mathrm{CW}}$ denote the number of such *Significant CW Switches* $\nu_1, \ldots, \nu_{\tilde{S}^{\mathrm{CW}}}$. We immediately see that $\tilde{S}^{\mathrm{CW}} \leq S^{\mathrm{CW}}$, as not all CW Switches are also Significant CW Switches. For example, a 'non-severe' and quickly reverted change of the Condorcet winner may not be counted towards $\tilde{S}^{\mathrm{CW}}$.

④ **Total Variation.** Another common notion of non-stationarity studied in the multi-armed bandit literature is the total variation present in the rewards (Besbes et al., 2014; Luo et al., 2018). Its analogue for dueling bandits can be defined as:

$$V := \sum_{t=2}^{T} \max_{a,b\in[K]} |P_t(a,b) - P_{t-1}(a,b)|,$$

which has been previously studied by Saha and Gupta (2022). However, $V$ can also be a pessimistic measure of complexity, as it can be of order $T$ even though the Condorcet winner remains fixed throughout all rounds.

⑤ **Condorcet Winner Variation.** We can then formulate a more refined version of total variation by accounting only for the maximal drift in the winning probabilities of the current Condorcet winner:

$$\tilde{V} := \sum_{t=2}^{T} \max_{a\in[K]} |P_t(a_t^*, a) - P_{t-1}(a_t^*, a)|.$$

**Remark 2.2** ($V$ vs $\tilde{V}$)**.** *It is clear from the definition that $\tilde{V} \leq V$. Moreover, we again see that the Condorcet Winner Variation can be much smaller than the Total Variation, i.e. $\tilde{V} \ll V$. For example, in the problem instance of Remark 2.1, we find that $\tilde{V} = 0$, whereas $V = T$. Thus, a regret bound in terms of the Condorcet Winner Variation $\tilde{V}$ can potentially be much stronger.*

# 3 Algorithm: ANACONDA

Following recent advances in non-stationary multi-armed bandits by Auer et al. (2019); Chen et al. (2019); Abbasi-Yadkori et al. (2022), and especially Suk and Kpotufe (2022), we construct an episode-based algorithm with a carefully chosen replay schedule.

Recall that we aim to minimize the dynamic regret w.r.t. a changing benchmark $a_t^*$. However, we quickly notice that we cannot reliably track the dynamic regret of some arm $a$, i.e. $\sum_t \delta_t(a_t^*, a)$, as the identity of the benchmark, $a_t^*$, changes at unknown times. As a resolution to this, the algorithmic idea that we deploy aims to detect relevant changes in the preference matrix by tracking the *static regret* $\max_{a'\in[K]} \sum_{t=s_1}^{s_2} \delta_t(a', a)$ instead. It will be the main challenge of our analysis to ensure that properly timed replays occur (and not too many of these) so that it is sufficient to track the static regret to guarantee low dynamic regret.

In the following, we explain our algorithmic approach in more detail. The algorithm is organized in episodes, denoted $\ell$. Similar to recent approaches to non-stationary multi-armed bandits (Auer et al., 2019; Abbasi-Yadkori et al., 2022; Suk and Kpotufe, 2022), the algorithm maintains a set of good arms, $\mathcal{A}_{\text{good}}$, and a replay schedule, $\{B_{s,m}\}_{s,m}$, within each episode. When no good arms are left in $\mathcal{A}_{\text{good}}$, a new episode begins and the set of good arms and the replay schedule are being reset.

Here, `ANACONDA` (Algorithm 1) is the meta procedure that initializes each episode by resetting the set of good arms to $[K]$, sampling a new replay schedule, and triggering the root call of $\texttt{CondaLet}(t_\ell, T+1-t_\ell)$.

When active in round $t$, a run of $\texttt{CondaLet}(t_0, m_0)$ (Algorithm 2) samples two arms uniformly at random from the active set of arms at round $t$, denoted $\mathcal{A}_t$. The set $\mathcal{A}_t$ is globally maintained by all calls of `CondaLet` and reset to $[K]$ at the beginning of each replay, i.e. call of `CondaLet`. When a child replay $\texttt{CondaLet}(t, m)$ is

**Algorithm 1** ANACONDA: Adaptive Non-stationAry CONdorcet Dueling Algorithm

**input:** horizon $T$
$t \leftarrow 1$
**while** $t \leq T$ **do**
  $t_\ell \leftarrow t$ // start of the $\ell$-th episode
  $\mathcal{A}_{\text{good}} \leftarrow [K]$
  **for** $m \in \{2, \ldots, 2^{\lceil \log(T) \rceil}\}$ and $s \in \{t_\ell + 1, \ldots, T\}$ **do** // set replay schedule
    Sample $B_{s,m} \sim \text{Bern}\left(\frac{1}{\sqrt{m(s-t_\ell)}}\right)$
  Run CondaLet$(t_\ell, T+1-t_\ell)$ // root replay in $\ell$-th episode

**Algorithm 2** CondaLet$(t_0, m_0)$

1: **input:** scheduled time $t_0$, duration $m_0$, replay schedule $\{B_{s,m}\}_{s,m}$
2: **initialize:** $t \leftarrow t_0$, $\mathcal{A}_t \leftarrow [K]$
3: **while** $t \leq T$ and $t \leq t_0 + m_0$ and $\mathcal{A}_{\text{good}} \neq \emptyset$ **do** // return if no good arms are left
4:   Play arm-pair $(a_t, b_t) \in \mathcal{A}_t$ with each arm being selected with probability $1/|\mathcal{A}_t|$
5:   $\mathcal{A}_{\text{good}} \leftarrow \mathcal{A}_{\text{good}} \setminus \{a \in [K] \colon \exists [s_1, s_2] \subseteq [t_\ell, t) \text{ s.t. (2) holds}\}$ // eliminate bad arms from $\mathcal{A}_{\text{good}}$
6:   $\mathcal{A}_{\text{local}} \leftarrow \mathcal{A}_t$ // save active set of arms locally
7:   $t \leftarrow t+1$
8:   **if** $\exists m$ such that $B_{t,m} = 1$ **then** // check for scheduled child replays
9:     Run CondaLet$(t, m)$ with $m = \max\{m \in \{2, \ldots, 2^{\lceil \log(T) \rceil}\} \colon B_{t,m} = 1\}$
10:   $\mathcal{A}_t \leftarrow \mathcal{A}_{\text{local}} \setminus \{a \in [K] \colon \exists [s_1, s_2] \subseteq [t_0, t) \text{ s.t. (2) holds}\}$ // eliminate bad arms from $\mathcal{A}_t$

scheduled in round $t$, i.e. $B_{t,m} = 1$ for some $m$, the parent algorithm, say CondaLet$(t_0, m_0)$, is interrupted (before eventually resuming if $t \leq t_0 + m_0$). To not overwrite arm eliminations of a parent by resetting $\mathcal{A}_t$ to $[K]$ in children, each version of CondaLet saves a local set of arms, $\mathcal{A}_{\text{local}}$, before checking for children.

**Gap Estimates.** Recall the definition of the gap between two arms as $\delta_t(a,b) = P_t(a,b) - 1/2$. Based on observed outcomes of duels, the ANACONDA maintains the following importance weighted estimates of $\delta_t(a,b)$:

$$\hat{\delta}_t(a,b) = |\mathcal{A}_t|^2 \, \mathbf{1}_{\{a_t = a, b_t = b\}} o_t(a,b) - 1/2. \qquad (1)$$

Wee see that whenever $a, b \in \mathcal{A}_t$, i.e. both arms are in the active set in round $t$, the estimator $\hat{\delta}_t(a,b)$ is an unbiased estimate of $\delta_t(a,b)$, as we select a pair of arms uniformly at random from $\mathcal{A}_t$ every round.

**Elimination Rule.** In Line 5 and Line 10 of Algorithm 2, we eliminate an arm $a \in [K]$ in round $t$ if there exist rounds $0 \leq s_1 < s_2 \leq t$ such that

$$\max_{a' \in [K]} \sum_{t=s_1}^{s_2} \hat{\delta}_t(a', a) > C \log(T) K \sqrt{(s_2 - s_1) \vee K^2} \quad (2)$$

where $C > 0$ is some universal constant that does not depend on $T$, $K$, or $S^{\text{CW}}$, and can be derived from the regret analysis.

### 3.1 Main Result

The main result of this paper is a $\tilde{O}(\sqrt{S^{\text{CW}} T})$ dynamic regret bound of ANACONDA without knowledge of the number of CW Switches $S^{\text{CW}}$. When $S^{\text{CW}} \ll S^{\text{P}}$, this bound substantially improves upon the *non-adaptive* $\tilde{O}(\sqrt{S^{\text{P}} T})$ rates in Saha and Gupta (2022) and Kolpaczki et al. (2022). In particular, as previously mentioned, the number of preference switches $S^{\text{P}}$ can be a very pessimistic measure of complexity. For example, a change in the preference between two suboptimal arms, or a minor change of the winning probabilities of the Condorcet winner under which it remains optimal, should not substantially affect the performance of a good algorithm (see Remark 2.1).

**Theorem 3.1** (Dynamic Regret of ANACONDA)**.** *Let $S^{\text{CW}}$ denote the unknown number of Condorcet Winner Switches in the sequence $\{\mathbf{P}_t\}_{t \in [T]}$. Let $\tau_1, \ldots, \tau_{S^{\text{CW}}}$ be the unknown times of these switches and let $\tau_0 := 1$ and $\tau_{S^{\text{CW}}+1} := T$. For some constant $c > 0$, the dynamic regret of ANACONDA is bounded as*

$$\text{DR}(T) \leq c \log^3(T) K \sum_{i=0}^{S^{\text{CW}}} \sqrt{\tau_{i+1} - \tau_i}.$$

An application of Jensen's inequality shows that this implies a dynamic regret bound of order $\tilde{O}(K\sqrt{S^{\text{CW}} T})$.

**Corollary 3.1** (Dynamic Regret w.r.t. $S^{\text{CW}}$)**.** *For some constant $c > 0$, the dynamic regret of ANACONDA is bounded as*

$$\text{DR}(T) \leq c \log^3(T) K \sqrt{(S^{\text{CW}} + 1) T}.$$

**Remark 3.1** (Regret Lower Bound and Justification of Optimality of Theorem 3.1)**.** *Note that a lower bound of $\Omega(\sqrt{KS^{\mathrm{P}}T})$ has recently been shown by Saha and Gupta (2022), which can also be seen to give a lower bound $\Omega(\sqrt{KS^{\mathrm{CW}}T})$ in terms of CW Switches $S^{\mathrm{CW}}$ as $S^{\mathrm{CW}} \leq S^{\mathrm{P}}$ (in particular, the lower bound problem instance used in Saha and Gupta (2022) is precisely such that $S^{\mathrm{CW}} = S^{\mathrm{P}}$). As a result, we find that the above bound is optimal up to logarithmic factors in its dependence on $S^{\mathrm{CW}}$ and $T$, while its dependence on $K$ may not be tight.*

## 4 Regret Analysis of ANACONDA

We build on recent advances in non-stationary multi-armed bandits, which are able to achieve near-optimal dynamic guarantees (Auer et al., 2019; Abbasi-Yadkori et al., 2022; Suk and Kpotufe, 2022) without knowledge of the non-stationary complexity. A common basis of the regret analysis in these works is a decomposition of the dynamic regret using the notion of good arms.

**Challenges in the Dueling Setting.** More precisely, within each episode $\ell$, prior work in multi-armed bandits (Auer et al., 2019; Abbasi-Yadkori et al., 2022; Suk and Kpotufe, 2022) decomposes the regret of their algorithm's selection $a_t$ into its relative regret against the last good arm $a_\ell^g \in \mathcal{A}_{\text{good}}$, and the relative regret of $a_\ell^g$ against the best arm $a_t^*$. A key advantage of this decomposition is that tracking the relative regret of some arm $a$ w.r.t. $a_\ell^g$ instead of $a_t^*$ is much easier. In particular, since $a_\ell^g$ is by definition considered good throughout the episode, it is always actively played, which guarantees unbiased estimates of the gap between any played arm $a$ and the last good arm $a_\ell^g$.

However, pairwise preferences are generally not transitive, let alone linear, so that a triangle inequality does not hold, i.e. $\delta_t(a_t^*, a) \not\leq \delta_t(a_t^*, a_\ell^g) + \delta_t(a_\ell^g, a)$. In NSt-DB, we can thus generally not rely on $a_\ell^g$ as a benchmark. Instead, in contrast to prior work in multi-armed bandits, we face the difficulty of having to argue directly that we can track the dynamic regret $\sum_t \delta_t(a_t^*, a)$ sufficiently well without a proxy benchmark such as $a_\ell^g$.

**Key Ideas to Overcome the Challenges.** To overcome these challenges, we consider every fixed arm $a \in [K]$ in isolation and split the horizon into the rounds before arm $a$ gets eliminated from $\mathcal{A}_{\text{good}}$ and the rounds after it gets eliminated from $\mathcal{A}_{\text{good}}$ in episode $\ell$. Letting $t_\ell^a$ be the elimination round of arm $a$, we will then argue that $t_\ell^a$ will occur sufficiently early to guarantee low regret (in episode $\ell$) before round $t_\ell^a$. For the rounds after elimination from $\mathcal{A}_{\text{good}}$, it will be key to dissect each possible replay of the eliminated arm and obtain replay-specific regret bounds, where we distinguish between 'confined' and 'unconfined' replays of arms. We now give an outline of our regret analysis.

### 4.1 Proof Sketch of Theorem 3.1

In the following, we let $\tilde{c} > 0$ denote a positive constant that does not dependent on $T$, $K$, or $S^{\mathrm{CW}}$, but may change from line to line. To begin our analysis, we state a concentration bound on the martingale difference sequence $\hat{\delta}_t(a,b) - \mathbb{E}[\hat{\delta}_t(a,b) \mid \mathcal{F}_{t-1}]$ as it can be found in similar form in (Beygelzimer et al., 2011) and (Suk and Kpotufe, 2022).

**Lemma 4.1.** *Let $\mathcal{E}$ be the event that for all rounds $1 \leq s_1 < s_2 \leq T$ and all arms $a, b \in [K]$:*

$$\left| \sum_{t=s_1}^{s_2} \hat{\delta}_t(a,b) - \sum_{t=s_1}^{s_2} \mathbb{E}\left[\hat{\delta}_t(a,b) \mid \mathcal{F}_{t-1}\right] \right| \leq c_1 \log(T) \left( K\sqrt{(s_2 - s_1)} + K^2 \right) \tag{3}$$

*for a sufficiently large constant $c_1 > 0$ and where $\mathcal{F} = \{\mathcal{F}_t\}_{t \in \mathbb{N}_0}$ denotes the canonical filtration. Then, event $\mathcal{E}$ occurs with probability at least $1 - 1/T^2$.*

Note that our elimination rule (2) has been chosen in accordance with the above concentration bound.

**Bounding Regret Within Episodes.** We proceed by bounding regret within each episode separately. Recall that we let $\tau_1 < \ldots < \tau_{S^{\mathrm{CW}}}$ denote the (unknown) rounds in which the Condorcet winner changes. We then refer to the interval $[\tau_i, \tau_{i+1})$ as the $i$-th phase, i.e. the interval for which $a_t^* = a_{\tau_i}^*$ for all $t \in [\tau_i, \tau_{i+1})$.

Let $\text{Phases}(t_1, t_2) = \{i : [\tau_i, \tau_{i+1}) \cap [t_1, t_2) \neq \emptyset\}$ be the set of phases $i$ such that $[\tau_i, \tau_{i+1})$ intersects with the interval $[t_1, t_2)$. Our main claim is the following upper bound on the dynamic regret within each episode:

$$\mathbb{E}\left[ \sum_{t=t_\ell}^{t_{\ell+1}-1} \frac{\delta_t(a_t^*, a_t) + \delta_t(a_t^*, b_t)}{2} \right] \leq \tilde{c} K \log^3(T)\, \mathbb{E}\left[ \sum_{i \in \text{Phases}(t_\ell, t_{\ell+1})} \sqrt{\tau_{i+1} - \tau_i} \right]. \tag{4}$$

By conditioning on $t_\ell$ and carefully applying the tower property, we can rewrite the expected dynamic regret within an episode in terms of fixed arms $a \in [K]$:

**Lemma 4.2.** *We have*

$$\mathbb{E}\left[ \sum_{t=t_\ell}^{t_{\ell+1}-1} \frac{\delta_t(a_t^*, a_t) + \delta_t(a_t^*, b_t)}{2} \right] = \mathbb{E}\left[ \sum_{a=1}^{K} \sum_{t=t_\ell}^{t_{\ell+1}-1} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|} \mathbf{1}_{\{a \in \mathcal{A}_t\}} \right].$$

In a next step, we split the RHS into the rounds before a fixed arm $a \in [K]$ has been eliminated from the good set, and the rounds after $a$ has been eliminated. Recall $t_\ell^a$ to be the round in episode $\ell$ in which arm $a$ is eliminated from $\mathcal{A}_{\text{good}}$ and consider

$$\underbrace{\mathbb{E}\left[\sum_{a=1}^{K}\sum_{t=t_\ell}^{t_\ell^a-1}\frac{\delta_t(a_t^*,a)}{|\mathcal{A}_t|}\right]}_{R_1(\ell)}+\underbrace{\mathbb{E}\left[\sum_{a=1}^{K}\sum_{t=t_\ell^a}^{t_{\ell+1}-1}\frac{\delta_t(a_t^*,a)}{|\mathcal{A}_t|}\mathbf{1}_{\{a\in\mathcal{A}_t\}}\right]}_{R_2(\ell)}$$

where we could drop the indicator in $R_1(\ell)$, since $\mathcal{A}_{\text{good}} \subseteq \mathcal{A}_t$ always. The remainder of our analysis is mostly concerned with showing that both, $R_1(\ell)$ and $R_2(\ell)$, are upper bounded by the RHS in (4).

**Regret Before Elimination.** The main difficulty in bounding $R_1(\ell)$ lies in the fact that some arm could have been eliminated due to being suboptimal, only to become the Condorcet winner soon after. As a result, large regret could go undetected, as the current Condorcet winner is not being actively played anymore. To this end, we have to argue that with high probability there will always be a replay scheduled that eliminates any bad arm from $\mathcal{A}_{\text{good}}$ in a timely manner, thereby eventually triggering a restart.

Here, we specifically consider calls of $\texttt{CondaLet}(s,m)$ that provably eliminate bad arms from $\mathcal{A}_{\text{good}}$. Importantly, by construction of our elimination rule (2), we can guarantee on the concentration event $\mathcal{E}$ that any replay that is scheduled within some phase $i$ will actively play the Condorcet winner of said phase.

**Lemma 4.3.** *On event $\mathcal{E}$, no call of $\texttt{CondaLet}(s,m)$ with $\tau_i \le s < \tau_{i+1}$ eliminates arm $a_i^*$ before round $\tau_{i+1}$.*

Roughly speaking, we can then argue that a replay that eliminates arm $a$ will be scheduled with high probability before the smallest round $s(a) > t_\ell$ such that

$$\sum_{t=t_\ell}^{s(a)}\delta_t(a_t^*,a) \gtrsim \sqrt{s(a)-t_\ell}.$$

In other words, arm $a$ is going to be eliminated from $\mathcal{A}_{\text{good}}$ before it suffers too much regret. Since $t_\ell^a$ is defined as the round in episode $\ell$ in which $a$ is eliminated from $\mathcal{A}_{\text{good}}$, we must have $t_\ell^a < s(a)$, which implies that the inner sum in $R_1(\ell)$ is at most of order $\sqrt{t_\ell^a - t_\ell}$ for every fixed arm $a \in [K]$. Finally, using that

$$\sqrt{t_\ell^a - t_\ell} \le \sum_{i\in\text{Phases}(t_\ell,t_\ell^a)}\sqrt{\tau_{i+1}-\tau_i}$$

and summing over all arms, we obtain the desired bound (4). Note that here summing over arms can be seen to account for a $\log(K)$ factor which we coarsely upper bound by $\log(T)$.

**Regret After Elimination.** $R_2(\ell)$ can be viewed as the regret due to replaying arms after they have been eliminated from the good set $\mathcal{A}_{\text{good}}$. We here distinguish between two types of replays:

**Definition 4.1.** We call $\texttt{CondaLet}(s,m)$ *confined* if there exists $i \in \text{Phases}(t_\ell,T)$ s.t. $[s,s+m] \subseteq [\tau_i,\tau_{i+1})$. In turn, we say that $\texttt{CondaLet}(s,m)$ is *unconfined* if for all $i \in \text{Phases}(t_\ell,T)$, we have $[s,s+m] \subseteq [\tau_i,\tau_{i+1})$.

To bound the regret within a confined replay, we recall that according to Lemma 4.3, on the concentration event $\mathcal{E}$, no replay will eliminate the Condorcet winner within the phase it is scheduled in. Thus, whenever some arm $a$ is being played by a confined replay, we obtain unbiased estimates of $\delta_t(a_t^*,a)$. It is then straightforward to show that for any confined $\texttt{CondaLet}(s,m)$, we have that $\sum_{t=s}^{s+m}\delta_t(a_t^*,a)$ is at most of order $\sqrt{m}$.

A similar line of argument does not work for unconfined replays, as they intersect with several phases. We then face a similar difficulty as when bounding $R_1(\ell)$, where the Condorcet winner of the current phase could have been eliminated (from the replay) in an earlier phase. Using similar arguments than for bounding $R_1(\ell)$, we show that for any unconfined $\texttt{CondaLet}(s,m)$, we have that $\sum_{t=s}^{s+m}\delta_t(a_t^*,a)$ is at most of order $\sqrt{s-t_\ell}+\sqrt{m}$.

Lastly, recall that in episode $\ell$ a replay $\texttt{CondaLet}(s,m)$ is scheduled with probability $1/\sqrt{m(s-t_\ell)}$. Crucially, any unconfined replay scheduled in $[\tau_i,\tau_{i+1})$ must have duration at least $m \ge \tau_{i+1}-s$ (otherwise it is not unconfined). Careful summation over confined and unconfined replays then yields the desired upper bound (4).

**Counting Episodes.** Lastly, we show that $\texttt{ANACONDA}$ will only restart if there has been a CW switch.

**Lemma 4.4.** *On event $\mathcal{E}$, for all episodes $\ell$ but the last there exists a change of the CW $t_\ell \le \tau_i < t_{\ell+1}$.*

This follows directly from the fact that on the concentration event $\mathcal{E}$ the current CW will never be eliminated from $\mathcal{A}_{\text{good}}$. Thus, if there is a restart, i.e. every arm has been eliminated from $\mathcal{A}_{\text{good}}$, there must have been a change of CW. Lemma 4.4 thus tells us that any phase intersects with at most two episodes. Summing the RHS of (4) over episodes then gives the claimed upper bound of

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{\delta_t(a_t^*,a_t)+\delta_t(a_t^*,b_t)}{2}\right] \le 2\tilde{c}K\log^3(T)\mathbb{E}\left[\sum_{i=1}^{S^{\text{CW}}}\sqrt{\tau_{i+1}-\tau_i}\right].$$

A detailed proof of Theorem 3.1 is given in Appendix A.

## 5 Tighter Bounds under SST and STI

We show that $\texttt{ANACONDA}$ can in fact yield a stronger regret guarantee in terms of a more refined notion of non-stationarity, Significant Condorcet Winner Switches (see

Section 2.2), under additional assumptions on the preference sequence $\mathbf{P}_1, \dots, \mathbf{P}_T$, namely, *Strong Stochastic Transitivity* (SST) and *Stochastic Triangle Inequality* (STI) (Yue and Joachims, 2009; Yue et al., 2012; Yue and Joachims, 2011). Let $a, b, c \in [K]$ and let $a \succ_t b$ denote that $a$ is preferred over $b$ in round $t$.

**Assumption 1** (Strong Stochastic Transitivity)**.** Every preference matrix $\mathbf{P}_t$ satisfies that if $a \succ_t b \succ_t c$, we have $\delta_t(a,c) \geq \delta_t(a,b) \vee \delta_t(b,c)$.

**Assumption 2** (Stochastic Triangle Inequality)**.** Every preference matrix $\mathbf{P}_t$ satisfies that if $a \succ_t b \succ_t c$, we have $\delta_t(a,c) \leq \delta_t(a,b) + \delta_t(b,c)$.

**Remark 5.1** (Example of SST & STI)**.** *Among the preference models that satisfy Assumption 1 and Assumption 2, are utility-based models with a symmetric and increasing link function $\sigma$. In these models, every arm $a$ has an associated (time-dependent) utility $u_t(a)$ and the probability of arm $a$ winning a duel against arm $b$ is given by $P_t(a \succ b) = \sigma(u_t(a) - u_t(b))$, where $\sigma$ is an increasing function with $\sigma(x) = 1 - \sigma(-x)$ and $\sigma(0) = 1/2$ that maps utility differences to probabilities (Yue et al., 2012; Bengs et al., 2021).*

### 5.1 Improved Dynamic Regret Analysis

Compared to Theorem 3.1, we now show how ANACONDA can achieve tighter regret guarantees in terms of both, Significant CW Switches and CW Variation.

① **Significant CW Switches.** Under Assumption 1 and Assumption 2, we are able to obtain the following adaptive dynamic regret bound w.r.t. $\tilde{S}^{\text{CW}}$.

**Theorem 5.1.** *Let $\tilde{S}^{\text{CW}}$ be the unknown number of Significant Condorcet Winner Switches. Under Assumption 1 and Assumption 2, ANACONDA has dynamic regret $\tilde{O}\big(K\sqrt{\tilde{S}^{\text{CW}}T}\big)$.*

**Remark 5.2.** *Recall from Section 2.2, since $\tilde{S}^{\text{CW}} \leq S^{\text{CW}}$ (as not all CW Switches are also Significant CW Switches), Theorem 5.1 gives an even tighter dynamic regret guarantee for the class of non-stationary preference sequences with SST and STI. Also note, this bound does not violate the $\Omega(\sqrt{KS^{\text{P}}T})$ lower bound as claimed in 3.1, as the lower bound is shown for a worst-case preference sequence $\{\mathbf{P}_t\}_{t\in[T]}$ where $\tilde{S}^{\text{CW}} = S^{\text{CW}} = S^{\text{P}}$.*

*Proof Overview.* With some additional effort, Assumption 1 and Assumption 2 allow us to utilize a dynamic regret decomposition similar to prior work on non-stationary multi-armed bandits, which yields a regret expression with episode-wise fixed benchmark arms. We can then reuse part of the regret analysis from Theorem 3.1 to show the claimed regret bound. Details can be found in Appendix B. □

We want to give a brief intuition about why additional assumptions are necessary when bounding dynamic regret w.r.t. Significant CW Switches $\tilde{S}^{\text{CW}}$ opposed to general CW Switches $S^{\text{CW}}$.[2] Consider a phase $[\nu_i, \nu_{i+1})$ in the sense of Significant CW Switches as defined in Section 2.2. As previously mentioned, the definition of a Significant CW Switch allows for several (non-severe) CW changes within each phase $[\nu_i, \nu_{i+1})$. As a result, we cannot guarantee that there will be any intervals during which the CW remains fixed, which would enable us to reliably estimate the relative regret $\sum_t \delta_t(a_t^*, a)$ so as to eliminate bad arms. Roughly speaking, assuming SST and STI enables us to identify bad arms based on knowledge of $\sum_t \delta_t(a', a)$ for some temporarily fixed benchmark $a'$. More details and a complete proof can be found in Appendix B.

② **Condorcet Winner Variation.** Recall the definition of the CW Variation $\tilde{V}$ from Section 2.2. As a consequence of Theorem 5.1, we can now show that ANACONDA also achieves near-optimal dynamic regret w.r.t. $\tilde{V}$.

**Corollary 5.1.** *Let $\tilde{V}$ be the unknown Condorcet Winner Variation in the sequence $\{\mathbf{P}_t\}_{t\in[T]}$. Under Assumption 1 and Assumption 2, ANACONDA has dynamic regret $\tilde{O}\big(K\sqrt{T} + \tilde{V}^{1/3}(KT)^{2/3}\big)$.*

**Remark 5.3.** *By definition, we have $\tilde{V} \leq V$, which means that Corollary 5.1 may yield a tighter dynamic regret bound than the (non-adaptive) $\tilde{O}\big(KV)^{1/3}T^{2/3}\big)$ guarantee by Saha and Gupta (2022). In view of the lower bound of $\Omega\big((KV)^{1/3}T^{2/3}\big)$ shown in (Saha and Gupta, 2022), the presented regret guarantee is also tight up to logarithmic factors and a factor of $K^{1/3}$. Note again that the lower bound by Saha and Gupta (2022) is not violated as their lower bound uses a worst-case preference sequence $\{\mathbf{P}_t\}_{t\in[T]}$ where $\tilde{V} = V$.*

## 6 Discussion and Future Work

We studied the problem of dynamic regret minimization in non-stationary dueling bandits and proposed an adaptive algorithm that yields provably optimal regret guarantees in terms of strong notions of non-stationary complexity. Our proposed algorithm is the first to achieve optimal dynamic dueling bandit regret without prior knowledge of the underlying non-stationary complexity.

While our results certainly close some of the practical open problems in preference elicitation in time varying preference models, it also leads many new questions along the line. In particular, as an extension to this work, one obvious question would be to understand non-stationary dueling bandits for more general preference matrices: What happens if the

[2] Note that this is a limitation of our regret analysis. It is an open question whether it is possible to achieve $O(\sqrt{\tilde{S}^{\text{CW}}T})$ dynamic regret in NSt-DB with general preference models.

preference sequence does not have a Condorcet winner in each round? What could be a good dynamic benchmark in that case? Hereto related, another open question is whether it is possible to obtain dynamic regret bounds in terms of Significant CW Switches ($\tilde{S}^{\texttt{CW}}$) for general preference sequences (without additional transitivity assumptions)? Extending the considered pairwise preference setting to more general subset-wise feedback (Saha and Gopalan, 2018, 2019b; Ghoshal and Saha, 2022) would be another interesting problem from a practical point of view.

## References


Yasin Abbasi-Yadkori, Andras Gyorgy, and Nevena Lazic. A new look at dynamic regret for non-stationary stochastic bandits. *arXiv preprint arXiv:2201.06532*, 2022.

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

Nir Ailon, Zohar Shay Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *ICML*, volume 32, pages 856–864, 2014.

Robin Allesiardo, Raphaël Féraud, and Odalric-Ambrym Maillard. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 3(4):267–283, 2017.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Peter Auer, Prateek Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. *In Proceedings of the 32nd International Conference on Learning Theory*, 99:138–158, 2019.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 2021.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in Neural Information Processing Systems*, 27: 199–207, 2014.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

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 Proceedings of the 32nd Conference on Learning Theory*, 99: 1–30, 2019.

Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pages 563–587, 2015.

Miroslav Dudik, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. *Conference on Learning Theory*, pages 563–587, 2015.

Pratik Gajane, Tanguy Urvoy, and Fabrice Clérot. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 218–227, 2015.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

Suprovat Ghoshal and Aadirupa Saha. Exploiting correlation to achieve faster learning rates in low-rank preference bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 456–482. PMLR, 2022.

Katja Hofmann, Shimon Whiteson, and Maarten De Rijke. A probabilistic method for inferring preferences from clicks. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 249–258, 2011.

Patrick Kolpaczki, Viktor Bengs, and Eyke Hüllermeier. Non-stationary dueling bandits. *arXiv preprint arXiv:2202.00935*, 2022.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2018.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. *In Proceedings of the 31st Conference On Learning Theory*, 75:1739–1776, 2018.

Filip Radlinski and Nick Craswell. Optimized interleaving for online retrieval evaluation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 245–254, 2013.

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. *Advances in Neural Information Processing Systems*, 32, 2019.

Yoan Russac, Olivier Cappé, and Aurélien Garivier. Algorithms for non-stationary generalized linear bandits. *arXiv preprint arXiv:2003.10113*, 2020.

Aadirupa Saha and Pierre Gaillard. Dueling bandits with adversarial sleeping. *Advances in Neural Information Processing Systems*, 34:27761–27771, 2021.

Aadirupa Saha and Pierre Gaillard. Versatile dueling bandits: Best-of-both-world analyses for online learning from preferences. 2022.

Aadirupa Saha and Aditya Gopalan. Battle of bandits. In *Uncertainty in Artificial Intelligence*, 2018.

Aadirupa Saha and Aditya Gopalan. Combinatorial bandits with relative feedback. In *Advances in Neural Information Processing Systems*, 2019a.

Aadirupa Saha and Aditya Gopalan. PAC Battling Bandits in the Plackett-Luce Model. In *Algorithmic Learning Theory*, pages 700–737, 2019b.

Aadirupa Saha and Aditya Gopalan. Best-item learning in random utility models with subset choices. In *International Conference on Artificial Intelligence and Statistics*, pages 4281–4291. PMLR, 2020.

Aadirupa Saha and Shubham Gupta. Optimal and efficient dynamic regret algorithms for non-stationary dueling bandits. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 19027–19049. PMLR, 2022.

Aadirupa Saha and Akshay Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *International Conference on Algorithmic Learning Theory*, pages 968–994. PMLR, 2022.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *International Conference on Machine Learning*, pages 9235–9244. PMLR, 2021.

Hossein Azari Soufiani, David C Parkes, and Lirong Xia. Preference elicitation for general random utility models. In *Uncertainty in Artificial Intelligence*, page 596. Citeseer, 2013.

Yanan Sui, Vincent Zhuang, Joel Burdick, and Yisong Yue. Multi-dueling bandits with dependent arms. In *Conference on Uncertainty in Artificial Intelligence*, UAI'17, 2017.

Joe Suk and Samory Kpotufe. Tracking most significant arm switches in bandits. In *Conference on Learning Theory*, pages 2160–2182. PMLR, 2022.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Huasen Wu and Xin Liu. Double Thompson sampling for dueling bandits. In *Advances in Neural Information Processing Systems*, pages 649–657, 2016.

Qingyun Wu, Naveen Iyer, and Hongning Wang. Learning contextual bandits in a non-stationary environment. *In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 495–504, 2018.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208. ACM, 2009.

Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 241–248, 2011.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The $k$-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Masrour Zoghi, Shimon Whiteson, Remi Munos, Maarten de Rijke, et al. Relative upper confidence bound for the $k$-armed dueling bandit problem. In *JMLR Workshop and Conference Proceedings*, number 32, pages 10–18. JMLR, 2014a.

Masrour Zoghi, Shimon A Whiteson, Maarten De Rijke, and Remi Munos. Relative confidence sampling for efficient on-line ranker evaluation. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 73–82. ACM, 2014b.

Masrour Zoghi, Zohar S Karnin, Shimon Whiteson, and Maarten De Rijke. Copeland dueling bandits. In *Advances in Neural Information Processing Systems*, pages 307–315, 2015.

# Appendix

## Table of Contents

**A Proof of Theorem 3.1** **12**
- A.1 Preliminaries . . . 12
- A.2 Bounding $R_1(\ell)$: Regret Before Elimination . . . 14
- A.3 Bounding $R_2(\ell)$: Regret After Elimination . . . 17
- A.4 Summing Over Episodes . . . 21

**B Missing Details from Section 5** **21**
- B.1 Significant CW Switches . . . 21
- B.2 Proof of Theorem 5.1 . . . 21
- B.3 Proof of Corollary 5.1 . . . 23

**C More Related Work** **24**

## Notation

| | |
|---|---|
| $a_t, b_t$ | Arms selected by the algorithm in round $t$ |
| $a, a', b$ | Generic fixed arms in $[K]$ |
| $\delta_t(a, b)$ | Gap between arm $a$ and arm $b$ |
| $\hat{\delta}_t(a, b)$ | Importance weighted gap estimate |
| $a_t^*$ | Condorcet winner in round $t$ |
| $t_\ell$ | First round in the $\ell$-th episode |
| $t_\ell^a$ | Round in the $\ell$-th episode in which $a$ is eliminated from $\mathcal{A}_{\text{good}}$ |
| $S^{\texttt{CW}}$ | Number of Condorcet Winner Switches |
| $\tau_1, \dots, \tau_{S^{\texttt{CW}}}$ | Rounds in which the Condorcet winner changes |
| $a_i^*$ | Condorcet winner in phase $i \in [S^{\texttt{CW}}]$, i.e. $a_t^* = a_i^*$ for $t \in [\tau_i, \tau_{i+1})$ |
| $\tilde{S}^{\texttt{CW}}$ | Number of Significant Condorcet Winner Switches |
| $\nu_1, \dots, \nu_{\tilde{S}^{\texttt{CW}}}$ | Rounds of Significant CW Switches |
| $a_i^s$ | Last safe arm in phase $[\nu_i, \nu_{i+1})$, i.e. last arm to satisfy (30) |
| $\tilde{V}$ | Condorcet Winner Variation |

# A Proof of Theorem 3.1

We organize the proof of Theorem 3.1 as follows. Section A.1 contains basic preliminary facts that will be the foundation of the upcoming proof. Section A.2 then bounds the regret any fixed arm suffers within each episode *before* being eliminated from the good set. Complementary to this, Section A.3 then deals with the regret an arm suffers *after* being eliminated.

## A.1 Preliminaries

In this preliminary section, we introduce a concentration bound on the sum of our estimates $\hat{\delta}_t$ in Section A.1.1. We then show in Section A.1.2 that the beginning of a new episode implies that the Condorcet winner has changed (on the concentration event), which will be useful later. Finally, Section A.1.3 decomposes the regret in terms episodes, arms, and rounds, which will form the basis of our analysis.

### A.1.1 Martingale Concentration Bound

We will rely on a similar martingale tail bound as Beygelzimer et al. (2011) and Suk and Kpotufe (2022), which is based on a version of Freedman's inequality given below.

**Lemma A.1** (Theorem 1 in Beygelzimer et al. (2011))**.** *Let $(X_t)_{t\in\mathbb{N}}$ be a martingale difference sequence w.r.t. some filtration $(\mathcal{F}_t)_{t\in\mathbb{N}_0}$. Assume that is $X_t$ is almost surely uniformly bounded, i.e. $X_t \leq R$ a.s. for some constant $R$. Moreover, suppose that $\sum_{s=1}^{t}\mathbb{E}[X_s^2 \mid \mathcal{F}_{s-1}] \leq V_t$ a.s. for some sequence of constants $(V_t)_{t\in\mathbb{N}}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\sum_{s=1}^{t} X_s \leq (e-1)\left(\sqrt{V_t \log(1/\delta)} + R\log(1/\delta)\right). \tag{5}$$

*Proof.* See Theorem 1 in Beygelzimer et al. (2011) and Lemma 1 in Suk and Kpotufe (2022). □

We now apply the above concentration bound to the martingale difference sequence $\hat{\delta}_t(a,b) - \mathbb{E}[\hat{\delta}_t(a,b) \mid \mathcal{F}_{t-1}]$.

**Lemma A.2.** *Let $\mathcal{E}$ be the event that for all rounds $s_1 < s_2$ and all arms $a, b \in [K]$:*

$$\left|\sum_{t=s_1}^{s_2} \hat{\delta}_t(a,b) - \sum_{t=s_1}^{s_2} \mathbb{E}\left[\hat{\delta}_t(a,b) \mid \mathcal{F}_{t-1}\right]\right| \leq c_1 \log(T)\left(K\sqrt{(s_2-s_1)} + K^2\right) \tag{6}$$

*for an appropriately large constant $c_1 > 0$ and where $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}_0}$ is the canonical filtration generated by observations in past rounds. Then, event $\mathcal{E}$ occurs with probability at least $1 - 1/T^2$.*

*Proof.* Note that $\hat{\delta}_t(a,b) - \mathbb{E}[\hat{\delta}_t(a,b) \mid \mathcal{F}_{t-1}]$ is naturally a martingale difference, since $\mathbb{E}\big[\hat{\delta}_t(a,b) - \mathbb{E}[\hat{\delta}_t(a,b) \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}\big] = 0$ a.s. Using that $|\mathcal{A}_t| \leq K$, we have that $X_t \leq 2K^2$ a.s. for all rounds $t$. Moreover, we get that

$$\sum_{t=s_1}^{s_2} \mathbb{E}\left[\hat{\delta}_t^2(a,b) \mid \mathcal{F}_{t-1}\right] \leq \sum_{t=s_1}^{s_2} |\mathcal{A}_t|^4\, \mathbb{E}\left[\mathbf{1}_{\{a_t=a, b_t=b\}} \mid \mathcal{F}_{t-1}\right] = \sum_{t=s_1}^{s_2} |\mathcal{A}_t|^2 \leq K^2(s_2-s_1).$$

We can thus apply Lemma A.1 with $R = K^2$ and $V_t = 2K^2 t$. Using $|x-y| \leq |x| + |y|$ and taking union bounds over $a, b$ and $s_1, s_2$, we then obtain Lemma A.2. □

### A.1.2 Episodes and Condorcet Winner Switches

**Lemma A.3.** *On event $\mathcal{E}$, for each episode $[t_\ell, t_{\ell+1})$ with $t_{\ell+1} \leq T$, there exists a change of the CW $\tau_i \in [t_\ell, t_{\ell+1})$.*

This implies that any phase $[\tau_i, \tau_{i+1})$ will intersect with at most two episodes.

*Proof.* The start of a new episode means that every arm $a \in [K]$ has been eliminated from $\mathcal{A}_{\text{good}}$ at some round in $t_\ell^a \in [t_\ell, t_{\ell+1})$. As a result, there must exist an interval $[s_1, s_2] \subseteq [t_\ell, t_\ell^a)$ and some arm $a' \in [K]$ so that the elimination rule (2) holds. Using Lemma A.2, we then find that for some constant $c_2 > 0$:

$$\sum_{t=s_1}^{s_2} \mathbb{E}\left[\hat{\delta}_t(a',a) \mid \mathcal{F}_{t-1}\right] > c_2 \log(T) K \sqrt{(s_2-s_1) \vee K^2}. \tag{7}$$

Note that by construction of $\hat{\delta}_t(a', a)$, we always have $\delta_t(a', a) \geq \mathbb{E}[\hat{\delta}_t(a', a) \mid \mathcal{F}_{t-1}]$ since

$$\mathbb{E}[\hat{\delta}_t(a', a) \mid \mathcal{F}_{t-1}] = \begin{cases} \delta_t(a', a) & a', a \in \mathcal{A}_t \\ -1/2 & \text{otherwise.} \end{cases} \tag{8}$$

Thus, in view of inequality (7), there exists no arm $a \in [K]$ such that $\max_{a'} \delta_t(a', a) = 0$ for all $t \in [t_\ell, t_{\ell+1})$, i.e. no fixed arm is optimal throughout the episode and there must have been a change of Condorcet winner. □

### A.1.3 Decomposing Regret across Episodes and Arms

We will bound regret of the algorithm withing each episode separately, i.e. we consider

$$\mathbb{E}\left[\sum_{t=t_\ell}^{t_{\ell+1}-1} \frac{\delta_t(a_t^*, a_t) + \delta_t(a_t^*, b_t)}{2}\right], \tag{9}$$

where $t_\ell$ is the first round in episode $\ell$ and $a_t^*$ is the Condorcet winner in round $t \in [T]$.

Recall that, every round $t \in [T]$, the algorithm selects an arm $a$ uniformly at random from the active set $\mathcal{A}_t$. It will then be useful to rewrite (11) in terms of fixed arms $a \in [K]$.

**Lemma A.4.** *We can write* (11) *in terms of the regret suffered by fixed arms:*

$$\mathbb{E}\left[\sum_{t=t_\ell}^{t_{\ell+1}-1} \frac{\delta_t(a_t^*, a_t) + \delta_t(a_t^*, b_t)}{2}\right] = \mathbb{E}\left[\sum_{a=1}^{K}\sum_{t=t_\ell}^{t_{\ell+1}} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|}\mathbf{1}_{\{a \in \mathcal{A}_t\}}\right] \tag{10}$$

*Proof.* As the algorithm independently and symmetrically selects two arms $(a_t, b_t)$ in each round (Line 4 in Algorithm 2), we can focus on bounding regret for one of the two arms, say $a_t$, by writing

$$\mathbb{E}\left[\sum_{t=t_\ell}^{t_{\ell+1}-1} \frac{\delta_t(a_t^*, a_t) + \delta_t(a_t^*, b_t)}{2}\right] = \mathbb{E}\left[\sum_{t=t_\ell}^{t_{\ell+1}-1} \delta_t(a_t^*, a_t)\right]. \tag{11}$$

Conditioning on $t_\ell$ and using the tower property, we then further find that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=t_\ell}^{t_{\ell+1}} \delta_t(a_t^*, a_t)\right] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=t_\ell}^{t_{\ell+1}} \delta_t(a_t^*, a_t) \mid t_\ell\right]\right] \\
&= \mathbb{E}\left[\sum_{t=t_\ell}^{T} \mathbb{E}\left[\mathbf{1}_{\{t < t_{\ell+1}\}} \mathbb{E}\left[\delta_t(a_t^*, a_t) \mid \mathcal{F}_{t-1}\right] \mid t_\ell\right]\right] \\
&= \mathbb{E}\left[\sum_{t=t_\ell}^{T}\sum_{a \in \mathcal{A}_t} \mathbb{E}\left[\mathbf{1}_{\{t < t_{\ell+1}\}} \mid t_\ell\right] \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|}\right] = \mathbb{E}\left[\sum_{t=t_\ell}^{t_{\ell+1}}\sum_{a \in \mathcal{A}_t} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|}\right],
\end{aligned}$$

where we used that $\mathbf{1}_{\{t < t_{\ell+1}\}}$ is $\mathcal{F}_{t-1}$-measurable and

$$\mathbb{E}\left[\delta_t(a_t^*, a_t) \mid \mathcal{F}_{t-1}\right] = \sum_{a \in \mathcal{A}_t} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|}.$$

Lastly, Lemma A.4 then follows from rewriting the sum over $a \in \mathcal{A}_t$ using the indicator $\mathbf{1}_{\{a \in \mathcal{A}_t\}}$ and swapping the order of the sums. □

In an important next step, we split the dynamic regret for *each fixed arm* $a \in [K]$ into:

(i) the regret we suffer from playing arm $a$ in the $\ell$-th episode before its elimination from $\mathcal{A}_{\text{good}}$,

(ii) the regret we suffer from (re)playing arm $a$ in the $\ell$-th episode after its elimination from $\mathcal{A}_{\text{good}}$.

Recall that $t_\ell^a \in [t_\ell, t_{\ell+1})$ denotes the time that arm $a$ is eliminated from $\mathcal{A}_{\text{good}}$ in episode $\ell$. Using Lemma A.4, we then decompose the dynamic regret in episode $\ell$ as

$$E\left[\sum_{t=t_\ell}^{t_{\ell+1}-1} \frac{\delta_t(a_t^*, a_t) + \delta_t(a_t^*, b_t)}{2}\right] = \underbrace{\mathbb{E}\left[\sum_{a=1}^{K} \sum_{t=t_\ell}^{t_\ell^a - 1} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|}\right]}_{R_1(\ell)} + \underbrace{\mathbb{E}\left[\sum_{a=1}^{K} \sum_{t=t_\ell^a}^{t_{\ell+1}-1} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|} \mathbf{1}_{\{a \in \mathcal{A}_t\}}\right]}_{R_2(\ell)}, \tag{12}$$

where for $R_1(\ell)$ we used that $a \in \mathcal{A}_{\text{good}}$ implies $a \in \mathcal{A}_t$ by construction of these sets. For every fixed arm, $R_1(\ell)$ corresponds to the regret suffered before said arm is eliminated from the master set. Accordingly, $R_2(\ell)$ is the regret due to replaying an arm after its elimination from the master set. The remainder of the proof is mainly concerned with bounding $R_1(\ell)$ and $R_2(\ell)$ appropriately.

### A.2 Bounding $R_1(\ell)$: Regret Before Elimination

We begin by assuming w.l.o.g. that $t_\ell^1 \le \cdots \le t_\ell^K$ so that for each round $t < t_\ell^a$ all arms $a' \ge a$ are element in $\mathcal{A}_{\text{good}} \subseteq \mathcal{A}_t$. As a result, we have $|\mathcal{A}_t| \ge K + 1 - a$ for all $t \le t_\ell^a$, and thus

$$\mathbb{E}\left[\sum_{a=1}^{K} \sum_{t=t_\ell}^{t_\ell^a - 1} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|}\right] \le \mathbb{E}\left[\sum_{a=1}^{K} \sum_{t=t_\ell}^{t_\ell^a - 1} \frac{\delta_t(a_t^*, a)}{K + 1 - a}\right]. \tag{13}$$

As we can see, the denominator will eventually account for a factor of $\log(K) \approx \sum_{a=1}^{K} 1/a$. We now concentrate on bounding the inner sum in (13), i.e. the regret of any fixed arm before being eliminated in the $\ell$-th episode.

#### A.2.1 Bounding $\mathbb{E}[\sum_{t=t_\ell}^{t_\ell^a - 1} \delta_t(a_t^*, a)]$ for any fixed arm $a \in [K]$

This section is devoted to proving the following upper bound.

**Lemma A.5.** *For some constant $c > 0$:*

$$\mathbb{E}\left[\sum_{t=t_\ell}^{t_\ell^a - 1} \delta_t(a_t^*, a)\right] \le c \log^2(T) K \, \mathbb{E}\left[\sum_{i \in \text{Phases}(t_\ell, t_\ell^a)} \sqrt{\tau_{i+1} - \tau_i}\right] + \frac{K}{T^2} + \frac{1}{T}. \tag{14}$$

To prove Lemma A.5, we will divide the interval $[t_\ell, t_\ell^a)$ into segments over the course of which arm $a$ suffers large regret and show that not too many of such segments will occur in interval $[t_\ell, t_\ell^a)$, i.e. until arm $a$ is being eliminated from $\mathcal{A}_{\text{good}}$. The definition of such bad segments is analogous to their construction in Abbasi-Yadkori et al. (2022) and Suk and Kpotufe (2022). Whereas prior work utilizes such segments to bound the regret of the last arm considered good in an episode, i.e. the last arm in $\mathcal{A}_{\text{good}}$, we will instead derive a regret bound for *any fixed arm $a$*. While the according regret bound will be in some sense weaker, it will still be sufficiently tight for our purposes. We here follow the notation in Suk and Kpotufe (2022).

**Definition A.1** (Bad Segments)**.** Fix $t_\ell$ and let $[\tau_i, \tau_{i+1})$ be any phase intersecting $[t_\ell, T)$. For an arm $a$, define rounds $s_{i,j}(a) \in [t_\ell \vee \tau_i, \tau_{i+1})$ recursively as follows: let $s_{i,0}(a) = t_\ell \vee \tau_i$ and define $s_{i,j+1}(a)$ as the smallest round in $(s_{i,j}(a), \tau_{i+1})$ such that arm $a$ satisfies for some constant $c_3 > 0$:

$$\sum_{t=s_{i,j}(a)}^{s_{i,j+1}(a)} \delta_t(a_i^*, a) > c_3 \log(T) K \sqrt{s_{i,j+1}(a) - s_{i,j}(a)}, \tag{15}$$

if such round $s_{i,j+1}(a)$ exists. Otherwise, we let $s_{i,j+1}(a) = \tau_{i+1} - 1$. We refer to the intervals $[s_{i,j}, s_{i,j+1})$ as bad segments if (15) is satisfied. If a segment does not satisfy (15), we refer to them as non-bad segments.[3]

Note that the concept of bad segments will become useful later as, for a fixed $t_\ell$, by definition of the bad segments, we can always upper bound the dynamic regret on an interval $[s_{i,j}(a), s_{i,j+1}(a))$ by

$$\sum_{t=s_{i,j}(a)}^{s_{i,j+1}(a)-1} \delta_t(a_t^*, a) \le c_3 \log(T) K \sqrt{s_{i,j+1}(a) - s_{i,j}(a)}. \tag{16}$$

[3]Note that by definition every segment but the last segment in a given phase must always satisfy (15)

We now define the *bad round* for an arm $a$ as the smallest round when the aggregated regret of bad segments exceeds $\sqrt{\text{interval length}}$ regret.

**Definition A.2** (Bad Round)**.** Fix $t_\ell$ and some arm $a$. The bad round $s(a) > t_\ell$ is defined as the smallest round which satisfies for some universally fixed constant $c_4 > 0$:

$$\sum_{(i,j):\ s_{i,j+1}(a) < s(a)} \sqrt{s_{i,j+1}(a) - s_{i,j}(a)} > c_4 \log(T) \sqrt{s(a) - t_\ell}, \tag{17}$$

where the sum is over all bad segments with $s_{i,j+1}(a) < s(a)$.

For a given episode $\ell$, we will show that arm $a$ is eliminated with high probability by the time the bad round $s(a)$ occurs. To this end, we will introduce perfect replays, i.e. those runs of `CondaLet` which are properly timed and eliminate arm $a$ before it aggregates large regret.

### A.2.2 Perfect Replays

The following result will become very useful and makes the intuition precise that on the concentration event the Condorcet winner will not be eliminated. More precisely, any run of $\texttt{CondaLet}(s, m)$ scheduled in phase $i$ will never eliminate $a_i^*$ inside phase $i$ as long as our concentration bound holds.

**Lemma A.6.** *On event $\mathcal{E}$, no run of $\texttt{CondaLet}(s, m)$ with $s \in [\tau_i, \tau_{i+1})$ ever eliminates arm $a_i^*$ before round $\tau_{i+1}$.*

*Proof.* Suppose the contrary that some $\texttt{CondaLet}(s, m)$ with $s \in [\tau_i, \tau_{i+1})$ eliminates arm $a_i^*$ before round $\tau_{i+1}$. Then, we must have for some arm $a \in [K]$ and interval $[s_1, s_2] \subseteq [s, \tau_{i+1})$ that

$$C \log(T) K \sqrt{(s_2 - s_1) \vee K^2} < \sum_{t=s_1}^{s_2} \hat{\delta}_t(a, a_i^*), \tag{18}$$

which using the concentration bound (6) implies on event $\mathcal{E}$ that

$$c_2 \log(T) K \sqrt{(s_2 - s_1) \vee K^2} < \sum_{t=s_1}^{s_2} \mathbb{E}\left[\hat{\delta}_t(a, a_i^*) \mid \mathcal{F}_{t-1}\right] \leq \sum_{t=s_1}^{s_2} \delta_t(a, a_i^*), \tag{19}$$

where the last inequality holds by merit of (8). Now, by the definition of arm $a_i^*$ as the Condorcet winner in phase $i$, we must have $\delta_t(a, a_i^*) \leq 0$ for all $t \in [\tau_i, \tau_{i+1})$ and all $a \in [K]$. Lemma A.6 then follows from contradiction. ☐

This leads to the following important property of `CondaLet` that states that properly timed replays of sufficient length will eliminate arms from $\mathcal{A}_{\text{good}}$ in the course of their bad segments. We call such calls of `CondaLet` *perfect replays*.

**Proposition A.1** (Perfect Replay)**.** *Suppose that event $\mathcal{E}$ holds. Let $[s_{i,j}(a), s_{i,j+1}(a))$ be a bad segment w.r.t. arm $a$ and let $\tilde{s}_{i,j}(a) = \left\lceil \frac{s_{i,j}(a) + s_{i,j+1}(a)}{2} \right\rceil$ be the midpoint of the interval. It holds that any run of $\texttt{CondaLet}(s, m)$ with $s \in [s_{i,j}(a), \tilde{s}_{i,j}(a)]$ and $m \geq s_{i,j+1}(a) - s_{i,j}(a)$ will eliminate arm $a$ from $\mathcal{A}_{good}$. We refer to such calls of `CondaLet` as* perfect replays *w.r.t. arm $a$.*

*Proof.* Let $\texttt{CondaLet}(s, m)$ be a replay such that $s \in [s_{i,j}(a), \tilde{s}_{i,j}(a)]$ and $m \geq s_{i,j+1}(a) - s_{i,j}(a)$. As any bad segment is by definition contained inside a phase, Lemma A.6 tells us that $a_i^* \in \mathcal{A}_t$ for all $t \in [\tilde{s}_{i,j}(a), s_{i,j+1}(a)]$. Recall that the estimates $\hat{\delta}_t(a_i^*, a)$ are unbiased if $a, a_i^* \in \mathcal{A}_t$ and we are thus able to obtain unbiased estimates of $\sum_{t=\tilde{s}_{i,j}(a)}^{s_{i,j+1}(a)} \delta_t(a_i^*, a)$. What is left to show is that in fact arm $a$ suffers sufficiently large regret to cause its elimination on this interval. To this end, by definition of the bad segments and basic algebraic manipulation, we find that

$$\begin{aligned}
\sum_{t=\tilde{s}_{i,j}(a)}^{s_{i,j+1}(a)} \delta_t(a_i^*, a) &= \sum_{t=s_{i,j}(a)}^{s_{i,j+1}(a)} \delta_t(a_i^*, a) - \sum_{t=s_{i,j}(a)}^{\tilde{s}_{i,j}(a)-1} \delta_t(a_i^*, a) \\
&\overset{(15)}{\geq} c_3 \log(T) K \left( \sqrt{s_{i,j+1}(a) - s_{i,j}(a)} - \sqrt{\tilde{s}_{i,j}(a) - 1 - s_{i,j}(a)} \right) \\
&\geq \frac{c_3}{4} \log(T) K \sqrt{s_{i,j+1}(a) - \tilde{s}_{i,j}(a)}.
\end{aligned}$$

Using that $\sum_{t=\tilde{s}_{i,j}(a)}^{s_{i,j+1}(a)} \hat{\delta}_t(a_i^*, a)$ is an unbiased estimate of $\sum_{t=\tilde{s}_{i,j}(a)}^{s_{i,j+1}(a)} \delta_t(a_i^*, a)$ and applying the concentration bound (6), this shows that arm $a$ satisfies the elimination rule (2) over interval $[\tilde{s}_{i,j}(a), s_{i,j+1}(a)]$ and will thus be eliminated by `CondaLet`$(s, m)$. □

#### A.2.3 Perfect replays are scheduled w.h.p.

Following Suk and Kpotufe (2022), we will now show that a perfect replay that eliminates arm $a$ is scheduled before round $s(a)$ with high probability. A replay `CondaLet`$(s, m)$ is scheduled if $B_{s,m} = 1$ and the random variables $B_{s,m}$ with $s \geq t_\ell$ are conditionally independent on $t_\ell$ (see Line 7 in Algorithm 1). We are thus interested in perfect replays `CondaLet`$(s, m)$ such that for any bad segment $[s_{i,j}(a), s_{i,j+1}(a))$ with $s_{i,j+1}(a) < s(a)$, we have $s \in [s_{i,j}(a), \tilde{s}_{i,j}(a)]$ and $m \geq s_{i,j+1}(a) - s_{i,j}(a)$. Moreover, we define $m_{i,j}$ as the smallest element in $\{2, \ldots, 2^{\lceil \log(T) \rceil}\}$ such that $m_{i,j} \geq s_{i,j+1}(a) - s_{i,j}(a)$, which implies that $s_{i,j+1}(a) - s_{i,j}(a) \geq \frac{m_{i,j}}{2}$. We will obtain the high probability guarantee via concentration on the sum

$$X(t_\ell, s(a)) = \sum_{(i,j):\ s_{i,j+1}(a) < s(a)} \ \sum_{s=s_{i,j}(a)}^{\tilde{s}_{i,j}(a)} B_{s,m_{i,j}}. \tag{20}$$

**Lemma A.7.** *Let $\mathcal{E}'(t_\ell)$ denote the event that $X(t_\ell, s(a)) \geq 1$ for all arms $a$, i.e. a perfect replay is scheduled before round $s(a)$. We have $\mathbb{P}(\mathcal{E}'(t_\ell) \mid t_\ell) \geq 1 - K/T^3$.*

*Proof.* Recalling that $B_{s,m} \mid t_\ell \sim \text{Bernoulli}\left(\frac{1}{\sqrt{m(s-t_\ell)}}\right)$, we find that

$$\begin{aligned}
\mathbb{E}[X(t_\ell, s(a)) \mid t_\ell] &\geq \frac{1}{\sqrt{2}} \sum_{\substack{(i,j):\\ s_{i,j+1}(a) < s(a)}} \frac{\tilde{s}_{i,j}(a) - s_{i,j}(a)}{\sqrt{s_{i,j+1}(a) - s_{i,j}(a)}\sqrt{s(a) - t_\ell}} \\
&\geq \frac{1}{4} \sum_{\substack{(i,j):\\ s_{i,j+1}(a) < s(a)}} \sqrt{\frac{s_{i,j+1}(a) - s_{i,j}(a)}{s(a) - t_\ell}} \overset{(17)}{\geq} \frac{c_4}{4} \log(T)
\end{aligned}$$

For $c_4$ sufficiently large the standard Chernoff bound tells us that

$$\mathbb{P}\left(X(t_\ell, s(a)) \leq \frac{\mathbb{E}[X(t_\ell, s(a)) \mid t_\ell]}{2} \mid t_\ell\right) \leq \exp\left(-\frac{\mathbb{E}[X(t_\ell, s(a)) \mid t_\ell]}{8}\right) \leq \frac{1}{T^3}.$$

The desired bound then follows from taking a union bound over all arms in $[K]$. □

Now, on event $\mathcal{E} \cap \mathcal{E}'(t_\ell)$, it must hold that $t_\ell^a < s(a)$ for all arms $a \in [K]$, since otherwise $a$ would have been eliminated by some perfect replay before round $t_\ell^a$ (by definition of event $\mathcal{E}'(t_\ell)$). As the bad round $s(a)$ is defined as the *smallest* round satisfying (17), we then have

$$\sum_{(i,j):\ s_{i,j+1}(a) < t_\ell^a} \sqrt{s_{i,j+1}(a) - s_{i,j}(a)} \leq c_4 \log(T) K \sqrt{t_\ell^a - t_\ell}. \tag{21}$$

Hence, in view of equation (16), over the bad segments, the regret of arm $a$ is at most of order $\log^2(T)\sqrt{t_\ell^a - t_\ell}$. Moreover, for every last segment in some phase $i$, $[s_{i,j}, s_{i,j+1}(a))$, as well as the final segment $[s_{i,j}(a), t_\ell^a)$, we know that the regret suffered from playing $a$ is upper bounded by $c_3 \log(T)\sqrt{\tau_{i+1} - \tau_i}$ by definition of non-bad segments (Definition A.1). Therefore, on event $\mathcal{E} \cap \mathcal{E}'(t_\ell)$, it follows from equation (21) and the above that

$$\sum_{t=t_\ell}^{t_\ell^a - 1} \delta_t(a_t^*, a) \leq c_5 K \log^2(T) \sum_{i \in \text{Phases}(t_\ell, t_\ell^a)} \sqrt{\tau_{i+1} - \tau_i}, \tag{22}$$

where we used that $\sqrt{t_\ell^a - t_\ell} \leq \sum_{i \in \text{Phases}(t_\ell, t_\ell^a)} \sqrt{\tau_{i+1} - \tau_i}$. Finally, we obtain Lemma A.5 by taking expectation and using that $\mathcal{E} \cap \mathcal{E}'(t_\ell)$ holds with high probability,

$$\mathbb{E}\left[\sum_{t=t_\ell}^{t_\ell^a - 1} \delta_t(a_t^*, a)\right] \leq \mathbb{E}\left[\left[\mathbf{1}_{\{\mathcal{E} \cap \mathcal{E}'(t_\ell)\}} \sum_{t=t_\ell}^{t_\ell^a - 1} \delta_t(a_t^*, a) \mid t_\ell\right]\right] + T\big(\mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{E}'(t_\ell)^c \mid t_\ell)\big)$$
$$\leq c_5 K \log^2(T)\, \mathbb{E}\left[\sum_{i \in \text{Phases}(t_\ell, t_\ell^a)} \sqrt{\tau_{i+1} - \tau_i}\right] + \frac{1}{T} + \frac{K}{T^2}.$$

### A.2.4 Summing Over Arms

Note that $t_\ell^a \leq t_{\ell+1}$ for all $a \in [K]$ by definition of $t_\ell^a$. Then, summing over all arms, it follows from Lemma A.5 and (13) that for some constant $c_6 > 0$:

$$\mathbb{E}\left[\sum_{a=1}^{K} \sum_{t=t_\ell}^{t_\ell^a - 1} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|}\right] \leq c_6 K \log^3(T)\, \mathbb{E}\left[\sum_{i \in \text{Phases}(t_\ell, t_{\ell+1})} \sqrt{\tau_{i+1} - \tau_i}\right], \tag{23}$$

where we loosely upper bound $\log(K)$ by $\log(T)$.

## A.3 Bounding $R_2(\ell)$: Regret After Elimination

Before we can begin, we will have to lay some groundwork to simplify the analysis in later steps. Recall the definition of bad segments from Section A.2 and define for every phase $[\tau_i, \tau_{i+1})$ intersecting with $[t_\ell^a, t_{\ell+1})$, i.e. $i \in \text{Phases}(t_\ell^a, t_{\ell+1})$, the segments $[s_{i,j}(a), s_{i,j+1})$ as in Definition A.1.

We will split the regret due to bad segments, i.e. those that satisfy (15), from the regret due to non-bad segments, i.e. the last segments in a phase that do no satisfy (15). For a fixed arm $a \in [K]$, we let bad$(a)$ denote the rounds $t \in [t_\ell, t_{\ell+1})$ such that $t \in [s_{i,j}(a), s_{i,j+1}(a))$ for any *bad* segment $[s_{i,j}(a), s_{i,j+1}(a))$.

By the definition of a non-bad segment (w.r.t. arm $a$), we know that that there is at most one such segment in every phase and that the regret of arm $a$ in each segment is upper bounded by $c_3 \log(T)\sqrt{\tau_{i+1} - \tau_i}$, where $[\tau_i, \tau_{i+1})$ is the phase that contains the segment. To take care of the denominator $|\mathcal{A}_t|$, assume w.l.o.g. that there is a run of `CondaLet`$(t_\ell^a, m)$ that remains active and uninterrupted until the final round $T$.[4] We can then reorder arms $a \in [K]$ according to the round that they are being eliminated by `CondaLet`$(t_\ell^a, m)$, which gives $|\mathcal{A}_t| \geq K + 1 - a$ whenever $a \in \mathcal{A}_t$. As before, this yields a factor of $\log(K)$ when summing over all arms. We then bound $R_2(\ell)$ over non-bad segments as

$$\mathbb{E}\left[\sum_{a=1}^{K} \sum_{t=t_\ell^a}^{t_{\ell+1}-1} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|} \mathbf{1}_{\{a \in \mathcal{A}_t, t \notin \text{bad}(a)\}}\right] \leq c_3 K \log(K) \log(T) \mathbb{E}\left[\sum_{i \in \text{Phases}(t_\ell^a, t_{\ell+1})} \sqrt{\tau_{i+1} - \tau_i}\right]. \tag{24}$$

The more challenging task is now to bound $R_2(\ell)$ for rounds in bad segments. Recall that, for a fixed arm $a \in [K]$, the sum in question relates to the expected regret suffered within an episode from replaying arm $a$ after it has been eliminated from $\mathcal{A}_{\text{good}}$, i.e. after time $t_\ell^a$. We begin by a straightforward upper bound. To this end, for a given replay `CondaLet`$(s, m)$, let $M(s, m, a)$ be the last round in $[s, s+m]$, where arm $a$ is active in `CondaLet`$(s, m)$ and all of its children. Then,

$$\mathbb{E}\left[\sum_{a=1}^{K} \sum_{t=t_\ell^a}^{t_{\ell+1}-1} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|} \mathbf{1}_{\{a \in \mathcal{A}_t, t \in \text{bad}(a)\}}\right] \leq \mathbb{E}\left[\sum_{a=1}^{K} \sum_{s=t_\ell+1}^{t_{\ell+1}-1} \sum_{m} \mathbf{1}_{\{B_{s,m}=1\}} \sum_{t=s \vee t_\ell^a}^{M(s,m,a)} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|} \mathbf{1}_{\{t \in \text{bad}(a)\}}\right], \tag{25}$$

where the most inner sum on the right hand side is for $m \in \{2, \ldots, 2^{\lceil \log(T) \rceil}\}$. We will keep the convention that whenever a sum over $m$ is not further specified, it will be over the above set. Note that (25) is a loose upper bound. While of course only a single call of `CondaLet` can be active at any point in time, we here sum over every possible replay and ignore the potential nesting and interleaving of replays. In particular, this upper bound is justified as each $\delta_t(a_t^*, a)$ is non-negative

[4] Note that this is w.l.o.g. when bounding $1/|\mathcal{A}_t|$ as any interrupting call of `CondaLet` would only increase $|\mathcal{A}_t|$ by resetting it to $[K]$.

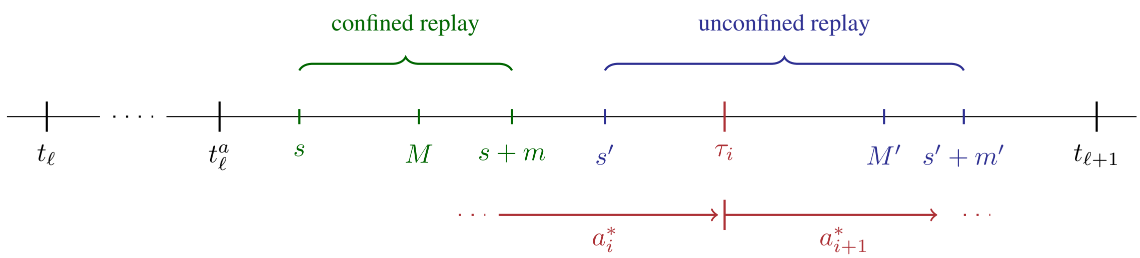


Figure 1: For some episode $[t_\ell, t_{\ell+1})$ and arm $a \in [K]$, an example of a confined replay and a unconfined replay, where $M = M(s, m, a)$ and $M' = M(s', m', a)$. When a replay $\texttt{CondaLet}(s', m')$ intersects with more than one phase, the CW in the next phase $[\tau_i, \tau_{i+1})$, denoted $a^*_{i+1}$, could be evicted before the beginning of that phase, i.e. in the interval $[s', \tau_i)$.

by definition of the CW $a_t^*$. The looseness of (25) will pose no obstacle, as the remainder of our upper bounds will be sufficiently tight as we will see.

Again, we first take care of the dependence on $K$ due to the denominator on the right hand side of (25). Note that for a fixed $\texttt{CondaLet}(s, m)$ if $a_k$ is the $k$-th arm to be eliminated by $\texttt{CondaLet}(s, m)$, then $\min_{t\in[s,M(s,m,a_k)]} |\mathcal{A}_t| \geq K + 1 - k$. Similarly to before, this will result in a multiplicative $\log(K)$ term when eventually switching the order of the sums and summing over all arms. For now, we therefore focus on the expression

$$\mathbb{E}\left[\sum_{s=t_\ell+1}^{t_{\ell+1}-1} \sum_m \mathbf{1}_{\{B_{s,m}=1\}} \sum_{t=s\vee t_\ell^a}^{M(s,m,a)} \delta_t(a_t^*, a) \mathbf{1}_{\{t\in \text{bad}(a)\}}\right] \tag{26}$$

for any fixed arm $a \in [K]$. To deal with this quantity, it will be helpful to distinguish between two types of replays, i.e. calls of $\texttt{CondaLet}$, which we refer to as confined and unconfined replays.

**Definition A.3** (Confined and Unconfined Replays)**.** For a fixed $t_\ell$, we call $\texttt{CondaLet}(s, m)$ *confined* if there exists $i \in \text{Phases}(t_\ell, T)$ such that $[s, s+m] \subseteq [\tau_i, \tau_{i+1})$, i.e. the replay intersects with a single phase only. In turn, we say that $\texttt{CondaLet}(s, m)$ is *unconfined* if for all $i \in \text{Phases}(t_\ell, T)$, we have $[s, s+m] \not\subseteq [\tau_i, \tau_{i+1})$.

An illustration of confined and unconfined replays is given in Figure 1.

We proceed by upper bounding the inner sum $\sum_{t=s\vee t_\ell^a}^{M(s,m,a)} \delta_t(a_t^*, a)\mathbf{1}_{\{t\in\text{bad}(a)\}}$ for confined and unconfined replays separately. The bound for confined replays comes with no major intricacies, whereas bounding the regret due to unconfined replays is slightly more involved.

### A.3.1 Bounding Regret for Confined Replays

We begin by bounding, the inner sum $\sum_{t=s\vee t_\ell^a}^{M(s,m,a)} \delta_t(a_t^*, a)$ for any confined replay in terms of the replay duration $m$.

**Lemma A.8.** *On event $\mathcal{E}$, for any fixed arm $a$ and confined replay $(s, m)$, it holds that*

$$\sum_{t=s\vee t_\ell^a}^{M(s,m,a)} \delta_t(a_t^*, a) \leq c_2 \log(T) K \sqrt{m}.$$

*Proof of Lemma A.8.* Consider any confined replay $\texttt{CondaLet}(s, m)$ with $[s, s+m] \subseteq [\tau_i, \tau_{i+1})$ for some phase $i$. This implies that on interval $[s, s+m]$ the Condorcet winner remains the same, i.e. $a_t^* = a_i^*$ for all $t \in [s, s+m]$. Now, recall from Lemma A.6 that, on event $\mathcal{E}$, arm $a_i^*$ will not be eliminated inside of $[s, s+m]$ as it is a subset of phase $[\tau_i, \tau_{i+1})$. As a result, we must have $a, a_i^* \in \mathcal{A}_t$ for all $t \in [s \vee t_\ell^a, M(s, m, a)]$ and our estimate $\hat{\delta}_t(a_i^*, a)$ is thus unbiased. Since $M(s, m, a)$ is the last round that arm $a$ is retained by $\texttt{CondaLet}(s, m)$ (and its children), it follows from the elimination rule (2) and the concentration bound (6) that

$$\sum_{t=s\vee t_\ell^a}^{M(s,m,a)} \delta_t(a_i^*, a) \leq c_2 \log(T) K \sqrt{M(s,m,a) - s \vee t_\ell^a} \leq c_2 \log(T) K \sqrt{m},$$

where the last inequality uses that $M(s, m, a) \leq s + m$. $\square$

### A.3.2 Bounding Regret for Unconfined Replays

**Lemma A.9.** *On event $\mathcal{E} \cap \mathcal{E}''(t_\ell)$, for any fixed arm $a$ and unconfined replay $(s, m)$, it holds that*

$$\sum_{t=s \vee t_\ell^a}^{M(s,m,a)} \delta_t(a_t^*, a) \mathbf{1}_{\{t \in bad(a)\}} \leq c_5 \log^2(T) K \big(\sqrt{s - t_\ell} + 2\sqrt{m}\big).$$

*Here, the event $\mathcal{E}''(t_\ell)$ is a concentration event similar to that in Lemma A.7 and will be defined in the following.*

*Proof of Lemma A.9.* Consider any unconfined replay $\texttt{CondaLet}(s, m)$ with $s \in [t_\ell, t_{\ell+1})$. Let $i$ be the phase so that $s \in [\tau_{i-1}, \tau_i)$. We can then split the sum over $t \in [s \vee t_\ell^a, M(s, m, a)]$ into the rounds before the Condorcet winner changes for the first time within $[s, s + m]$ and the remaining rounds, i.e.

$$\sum_{t=s \vee t_\ell^a}^{M(s,m,a)} \delta_t(a_t^*, a) = \sum_{t=s \vee t_\ell^a}^{\tau_i - 1} \delta_t(a_i^*, a) + \sum_{t=\tau_i}^{M(s,m,a)} \delta_t(a_t^*, a). \tag{27}$$

Note that the interval $[\tau_i, M(s, m, a)]$ can itself span over several phases. The first sum on the right hand side can be bounded as in Lemma A.8. Using Lemma A.6, the elimination rule, and the concentration bound, we get

$$\sum_{t=s \vee t_\ell^a}^{\tau_i - 1} \delta_t(a_i^*, a) \leq c_2 \log(T) K \sqrt{m}.$$

The second sum cannot be bounded in a similar way, as we cannot guarantee that the Condorcet winner in some round $t \in [\tau_i, M(s, m, a)]$ has not been eliminated in prior rounds $[s \vee t_\ell^a, \tau_i)$. For example in Figure 1, the unconfined replay $\texttt{CondaLet}(s', m')$ could have eliminated $a_{i+1}^*$ on interval $[s', \tau_i)$ before it became the Condorcet winner. We may therefore fail to detect that $a$ suffers large regret without additional replays.

To resolve this difficulty, we can reuse part of the arguments from Section A.2. Define the bad segments $[s_{k,j}(a), s_{k,j+1}(a))$ for $k \geq i$ as in Definition A.1. Similarly to before, we now define the bad round $s'(a)$ as the smallest round $s'(a) > \tau_i$ such that for the same constant $c_4 > 0$ as in (17)

$$\sum_{(k,j):\ s_{k,j+1}(a) < s'(a)} \sqrt{s_{k,j+1}(a) - s_{k,j}(a)} > c_4 \log(T) \sqrt{s'(a) - t_\ell}, \tag{28}$$

where the sum is over all bad segments with $k \geq i$ and $s_{k,j+1}(a) < s'(a)$.

Importantly, for this definition of $s'(a)$ and with the sum $X(t_\ell, s'(a))$ defined accordingly, the high probability guarantee of Lemma A.7 still holds. This implies that a perfect replay (see Proposition A.1) that eliminates arm $a$ (from the unconfined replay $\texttt{CondaLet}(s, m)$) is scheduled w.h.p. before the bad round $s'(a)$ occurs. Let the corresponding event denote $\mathcal{E}''(t_\ell)$ as in Lemma A.7.

The round $M(s, m, a)$ was defined as the last round for which $a$ is retained in $\texttt{CondaLet}(s, m)$ and all of its children. Hence, on event $\mathcal{E} \cap \mathcal{E}''(t_\ell)$, we must have $M(s, m, a) < s'(a)$ as otherwise $a$ would have been eliminated from $\texttt{CondaLet}(s, m)$ (or one of its children) before round $M(s, m, a)$, a contradiction. By merit of (16), this yields

$$\sum_{(k,j):\ s_{k,j+1}(a) < M(s,m,a)} \sqrt{s_{k,j+1}(a) - s_{k,j}(a)} \leq c_4 \log(T) K \sqrt{M(s, m, a) - t_\ell}$$

The regret on the final segment $[s_{k,j}(a), M(s, m, a)]$ can trivially be bounded by $c_3 \log(T) K \sqrt{m}$, as it must be a non-bad segment and $M(s, m, a) - s_{k,j}(a) \leq m$. Finally, in view of (16), it follows that

$$\begin{aligned}\sum_{t=s \vee t_\ell^a}^{M(s,m,a)} \delta_t(a_t^*, a) \mathbf{1}_{\{t \in \text{bad}(a)\}} &\leq c_5 \log^2(T) K \big(\sqrt{M(s, m, a) - t_\ell} + \sqrt{m}\big) \\ &\leq c_5 \log^2(T) K \big(\sqrt{s - t_\ell} + 2\sqrt{m}\big),\end{aligned}$$

where the second inequality uses $\sqrt{M(s, m, a) - t_\ell} \leq \sqrt{s - t_\ell} + \sqrt{m}$, since $M(s, m, a) \leq s + m$ and $s \geq t_\ell$.

□

### A.3.3 Combining Confined and Unconfined Replays

We will now conclude the bound on $R_2(\ell)$. To this end, recall that the replay schedule is chosen according to $B_{s,m} \mid t_\ell \sim \text{Bern}\big(1/\sqrt{m(s-t_\ell)}\big)$. Then, conditioning on $t_\ell$, we have

$$\mathbb{E}\left[\sum_{s=t_\ell+1}^{t_{\ell+1}} \sum_m \mathbf{1}_{\{B_{s,m}\}}\right] = \mathbb{E}\left[\sum_{s=t_\ell+1}^{T} \sum_m \mathbb{E}\left[\mathbf{1}_{\{B_{s,m}\}} \mid t_\ell\right] \mathbb{E}\left[\mathbf{1}_{\{s<t_{\ell+1}\}} \mid t_\ell\right]\right] = \mathbb{E}\left[\sum_{s=t_{\ell+1}}^{t_{\ell+1}-1} \frac{1}{\sqrt{m(s-t_\ell)}}\right].$$

Moreover, note that we can rewrite a sum over $s \in [t_\ell+1, t_{\ell+1})$ as a double sum over $i \in \text{Phases}(t_\ell, t_{\ell+1})$ and $s \in [\tau_i \vee (t_\ell+1), \tau_{i+1} \wedge t_{\ell+1})$. For unconfined replays, we notice that when $\texttt{CondaLet}(s,m)$ is scheduled with $s \in [\tau_i, \tau_{i+1})$, it must hold that $m \geq \tau_{i+1} - s$, as $\texttt{CondaLet}(s,m)$ would otherwise not be unconfined.

Now, combining Lemma A.8 and Lemma A.9, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{\{\mathcal{E}\cap\mathcal{E}''(t_\ell)\}} \sum_{a=1}^{K} \sum_{t=t_\ell^a}^{t_{\ell+1}-1} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|} \mathbf{1}_{\{a\in\mathcal{A}_t, t\in\text{bad}(a)\}}\right] \\
&\leq \mathbb{E}\left[\mathbf{1}_{\{\mathcal{E}\cap\mathcal{E}''(t_\ell)\}} \sum_{s=t_\ell+1}^{t_{\ell+1}-1} \sum_m \mathbf{1}_{\{B_{s,m}=1\}} \sum_{t=s\vee t_\ell^a}^{M(s,m,a)} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|} \mathbf{1}_{\{t\in\text{bad}(a)\}}\right] \\
&\leq c_2 K \log(K) \log(T)\, \mathbb{E}\left[\sum_{s=t_\ell}^{t_{\ell+1}-1} \sum_m \frac{\sqrt{m}}{\sqrt{m(s-t_\ell)}}\right] \\
&\qquad\qquad + c_5 K \log(K) \log^2(T)\, \mathbb{E}\left[\sum_{i\in\text{Phases}(t_\ell,t_{\ell+1})} \sum_{s=\tau_i}^{\tau_{i+1}-1} \sum_{m\geq\tau_{i+1}-s} \frac{\sqrt{s-t_\ell}+2\sqrt{m}}{\sqrt{m(s-t_\ell)}}\right] \\
&\leq c_2 K \log^3(T)\, \mathbb{E}\left[\sqrt{t_{\ell+1}-t_\ell}\right] + c_5 K \log^3(T)\, \mathbb{E}\left[\sum_{i\in\text{Phases}(t_\ell,t_{\ell+1})} \sum_{s=\tau_i}^{\tau_{i+1}-1} \frac{1}{\sqrt{\tau_{i+1}-s}} + 2\sqrt{t_{\ell+1}-t_\ell}\right] \\
&\leq c_7 K \log^3(T)\, \mathbb{E}\left[2 \sum_{i\in\text{Phases}(t_\ell,t_{\ell+1})} \sqrt{\tau_{i+1}-\tau_i} + \sum_{i\in\text{Phases}(t_\ell,t_{\ell+1})} \sqrt{\tau_{i+1}-\tau_i} + 4 \sum_{i\in\text{Phases}(t_\ell,t_{\ell+1})} \sqrt{\tau_{i+1}-\tau_i}\right] \\
&\leq 7c_7 K \log^3(T)\, \mathbb{E}\left[\sum_{i\in\text{Phases}(t_\ell,t_{\ell+1})} \sqrt{\tau_{i+1}-\tau_i}\right].
\end{aligned}$$

We here repeatedly used that $\sum_{k=1}^n 1/\sqrt{k} \leq 2\sqrt{n}$ in the third and fourth inequality. In particular, the fourth inequality holds as $\sqrt{t_{\ell+1}-t_\ell} \leq \sum_{i\in\text{Phases}(t_\ell,t_{\ell+1})} \sqrt{\tau_{i+1}-\tau_i}$ and

$$\sum_{s=\tau_i}^{\tau_{i+1}-1} \frac{1}{\sqrt{\tau_{i+1}-s}} = \sum_{s=1}^{\tau_{i+1}-\tau_i-1} \frac{1}{\sqrt{s}} \leq \sqrt{\tau_{i+1}-\tau_i}.$$

Further note that, as explained before, the denominator $|\mathcal{A}_t|$ can be seen to account for a factor of $\log(K)$, which we loosely upper bounded by $\log(T)$. Together with (24), we then obtain for some constant $c_8 > 0$ the desired bound of

$$E\left[\sum_{a=1}^{K} \sum_{t=t_\ell^a}^{t_{\ell+1}-1} \frac{\delta_t(a_t^*, a)}{|\mathcal{A}_t|} \mathbf{1}_{\{a\in\mathcal{A}_t\}}\right] \leq c_8 K \log^3(T)\, \mathbb{E}\left[\sum_{i\in\text{Phases}(t_\ell,t_{\ell+1})} \sqrt{\tau_{i+1}-\tau_i}\right]. \tag{29}$$

### A.4 Summing Over Episodes

In Section A.2 and Section A.3, we bounded the regret of arms within an episode before and after their elimination, respectively. Combining (23) and (29), and summing over episodes, we then obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} \frac{\delta_t(a_t^*, a_t) + \delta_t(a_t^*, b_t)}{2}\right] \leq c_9 K \log^3(T)\, \mathbb{E}\left[\mathbf{1}_{\{\mathcal{E}\}} \sum_{\ell=1}^{L} \sum_{i \in \text{Phases}(t_\ell, t_{\ell+1})} \sqrt{\tau_{i+1} - \tau_i}\right] + \frac{1}{T}.$$

Now, on the concentration event $\mathcal{E}$, Lemma A.3 tells us that any phase $[\tau_i, \tau_{i+1})$ intersects with at most two episodes. Recall that $\tau_0 := 1$ and $\tau_{S^{\texttt{CW}}+1} := T$. It then follows from the above that

$$\mathbb{E}\left[\sum_{t=1}^{T} \frac{\delta_t(a_t^*, a_t) + \delta_t(a_t^*, b_t)}{2}\right] \leq 2c_9 K \log^3(T) \sum_{i=0}^{S^{\texttt{CW}}} \sqrt{\tau_{i+1} - \tau_i} + \frac{1}{T}.$$

## B Missing Details from Section 5

### B.1 Significant CW Switches

Let us first recall the definition of Significant Condorcet Winner Switches from Section 2.2. Let $\nu_0 := 1$ and define $\nu_{i+1}$ recursively as the first round in $[\nu_i, T)$ such that for all arms $a \in [K]$ there exist rounds $\nu_i \leq s_1 < s_2 < \nu_{i+1}$ such that

$$\sum_{t=s_1}^{s_2} \delta_t(a_t^*, a) \geq \sqrt{K(s_2 - s_1)}. \tag{30}$$

Let $\tilde{S}^{\texttt{CW}}$ denote the number of such Significant CW Switches $\nu_1, \ldots, \nu_{\tilde{S}^{\texttt{CW}}}$. The key idea of Suk and Kpotufe (2022) when developing this notion of non-stationarity (for multi-armed bandits) is that a restart in exploration is only warranted if there are no *safe* arms left to play, i.e. there is no arm left that does not suffer regret (30) on some interval $[s_1, s_2]$. For every phase $[\nu_i, \nu_{i+1})$, we denote by $a_i^s$ the last safe arm in phase $i$, i.e. the last arm to satisfy (30) in phase $i$. Moreover. we define the sequence of safe arms as $a_t^s = a_i^s$ for $t \in [\nu_i, \nu_{i+1})$.

Significant CW Switches are able to reconcile switch-based non-stationarity measures such as CW Switches $S^{\texttt{CW}}$ and variation-based non-stationarity measures such as the CW Variation $\tilde{V}$. More specifically, it naturally holds that $\tilde{S}^{\texttt{CW}} \leq S^{\texttt{CW}}$ and Corollary 5.1 shows that near-optimal dynamic regret w.r.t. $\tilde{S}^{\texttt{CW}}$ also implies near-optimal dynamic regret w.r.t. $\tilde{V}$.

### B.2 Proof of Theorem 5.1

For convenience, we recall the assumptions of Theorem 5.1.

**Assumption 1** (Strong Stochastic Transitivity)**.** Every preference matrix $\mathbf{P}_t$ satisfies that if $a \succ_t b \succ_t c$, we have $\delta_t(a,c) \geq \delta_t(a,b) \vee \delta_t(b,c)$.

**Assumption 2** (Stochastic Triangle Inequality)**.** Every preference matrix $\mathbf{P}_t$ satisfies that if $a \succ_t b \succ_t c$, we have $\delta_t(a,c) \leq \delta_t(a,b) + \delta_t(b,c)$.

We see that together Assumption 1 and Assumption 2 imply a more general type of triangle inequality for any triplet $a, b, c \in [K]$ with $a \succ b$ and $a \succ c$.

**Lemma B.1.** *Under Assumption 1 and Assumption 2, for any triplet $a, b, c \in [K]$ with $a \succ_t b$ and $a \succ_t c$, it holds that*

$$\delta_t(a,c) \leq 2\delta_t(a,b) + \delta_t(b,c).$$

*Proof.* Suppose that $b \succ_t c$. Then, the claim follows directly from the stochastic triangle inequality, since $\delta_t(a,c) \leq \delta_t(a,b) + \delta_t(b,c)$. Suppose that $c \succ_t b$. Leveraging strong stochastic transitivity of the triplet $a \succ_t c \succ_t b$, we have

$$\delta_t(a,b) \geq \delta_t(a,c) \vee \delta_t(c,b).$$

This implies that $\delta_t(a,c) \leq \delta_t(a,b)$ as well as $\delta_t(c,b) \leq \delta_t(a,b)$. By definition of the gaps, this also yields $|\delta_t(b,c)| \leq \delta_t(a,b)$, since $c \succ_t b$. Consequently, it holds that $\delta_t(a,c) \leq 2\delta_t(a,b) + \delta_t(b,c)$. ∎

As briefly discussed in Section 5, these assumptions on the preference sequence $\mathbf{P}_1, \dots, \mathbf{P}_T$ allow us to decompose the dynamic regret so that we can compare against a temporarily fixed benchmark.

We can w.l.o.g. assume that $a_t^* \succ_t a_t$ and $a_t^* \succ_t a_t^s$. To see that this assumption is valid, note that $a_t^*$ is the Condorcet winner in round $t$ and it is then easy to verify that Lemma B.1 also holds if $a_t^*$ equals one (or both) of $a_t$ and $a_t^s$. Applying Lemma B.1 to $a_t^*$, $a_t^s$ and $a_t$, we have

$$\delta_t(a_t^*, a_t) \le 2\delta_t(a_t^*, a_t^s) + \delta_t(a_t^s, a_t).$$

Recalling equation (11) from Section A, we then get the following decomposition of the dynamic regret within each episode as

$$\mathbb{E}\left[\sum_{t=t_\ell}^{t_{\ell+1}-1} \frac{\delta_t(a_t^*, a_t) + \delta_t(a_t^*, b_t)}{2}\right] \le 2\,\underbrace{\mathbb{E}\left[\sum_{t=t_\ell}^{t_{\ell+1}-1} \delta_t(a_t^*, a_t^s)\right]}_{\tilde{R}_1(\ell)} + \underbrace{\mathbb{E}\left[\sum_{t=t_\ell}^{t_{\ell+1}-1} \delta_t(a_t^s, a_t)\right]}_{\tilde{R}_2(\ell)}.$$

### B.2.1 Bounding $\tilde{R}_1(\ell)$

We can bound $\tilde{R}_1(\ell)$ directly using the definition of Significant CW Switches. By definition of $a_i^s$ as the last safe arm in phase $[\nu_i, \nu_{i+1})$, i.e. the last arm to satisfy (30) for some interval $[s_1, s_2] \subseteq [\nu_i, \nu_{i+1})$, it holds that

$$\sum_{t=\nu_i}^{\nu_{i+1}} \delta_t(a_t^*, a_i^s) \le \sqrt{K(\nu_{i+1} - \nu_i)}.$$

We can then sum over all phases $i \in [\tilde{S}^{\texttt{CW}}]$ to obtain

$$\sum_{t=1}^{T} \delta_t(a_t^*, a_t^s) \le \sum_{i=1}^{\tilde{S}^{\texttt{CW}}} \sqrt{K(\nu_{i+1} - \nu_i)}.$$

### B.2.2 Bounding $\tilde{R}_2(\ell)$

As briefly mentioned in the main text, the difficulty in bounding $\sum_{t=t_\ell}^{t_{\ell+1}-1} \delta_t(a_t^*, a_t)$ for Significant CW Switches is that the identity of the Condorcet winner, i.e. $a_t^*$, may change several times within each significant phase $i \in [\tilde{S}^{\texttt{CW}}]$. This makes accurately tracking $\delta_t(a_t^*, a)$ (nearly) impossible even across small intervals and the arguments that we used to prove Theorem 3.1 fail.

In contrast, when we consider the relative regret of $a_t$ against the last safe arm $a_t^s$ (or sequence thereof), this difficulty can be resolved. Considering $a_t^s$ (instead of $a_t^*$) as a benchmark ensures that within each phase $i \in [\tilde{S}^{\texttt{CW}}]$ the comparator arm is fixed, since $a_t^s = a_i^s$ for all $t \in [\nu_i, \nu_{i+1})$. Hence, the relative regret w.r.t. $a_t^s$ can still be dealt with. In particular, the proof of Theorem 3.1 from Section A can be seen to hold with minor changes when swapping $a_t^*$ for $a_t^s$ and considering significant phases $\nu_1, \dots, \nu_{\tilde{S}^{\texttt{CW}}}$. For completeness, we reformulate and prove two important lemmas from Section A that relied on properties of $a_t^*$ and $\tau_1, \dots, \tau_{S^{\texttt{CW}}}$. We want to emphasise that we here again rely on Assumption 1 and Assumption 2.

The following lemma shows that the beginning of a new episode implies a Significant CW Switch, i.e. every arm suffers at least (30) much regret over some interval within the episode.

**Lemma B.2** (Lemma A.3 for $\tilde{S}^{\texttt{CW}}$)**.** *On event $\mathcal{E}$, for each episode $[t_\ell, t_{\ell+1})$ with $t_{\ell+1} \le T$, there exists a Significant CW Switch $\nu_i \in [t_\ell, t_{\ell+1})$.*

*Proof.* The start of a new episode means that every arm $a \in [K]$ has been eliminated from $\mathcal{A}_{\text{good}}$ at some round in $t_\ell^a \in [t_\ell, t_{\ell+1})$. As a result, there must exist an interval $[s_1, s_2] \subseteq [t_\ell, t_\ell^a)$ and some arm $a' \in [K]$ so that the elimination rule (2) holds. Using Lemma A.2, we then find that for some constant $c_2 > 0$:

$$\sum_{t=s_1}^{s_2} \mathbb{E}\left[\hat{\delta}_t(a', a) \mid \mathcal{F}_{t-1}\right] > c_2 \log(T) K \sqrt{(s_2 - s_1) \vee K^2}. \tag{31}$$

Note that by construction of $\hat{\delta}_t(a', a)$, we always have $\delta_t(a', a) \geq \mathbb{E}[\hat{\delta}_t(a', a) \mid \mathcal{F}_{t-1}]$ since

$$\mathbb{E}[\hat{\delta}_t(a', a) \mid \mathcal{F}_{t-1}] = \begin{cases} \delta_t(a', a) & a', a \in \mathcal{A}_t \\ -1/2 & \text{otherwise.} \end{cases} \tag{32}$$

Applying Lemma B.1 to the triplet $(a_t^*, a', a)$, we get that $\delta_t(a_t^*, a) \geq 2\delta_t(a_t^*, a') + \delta_t(a', a) \geq \delta_t(a', a)$. Thus, (31) tells us that there exists no arm $a \in [K]$ such that for all $[s_1, s_2] \subseteq [t_\ell, t_{\ell+1})$

$$\sum_{t=s_1}^{s_2} \delta_t(a_t^*, a) < \sqrt{K(s_2 - s_1)}.$$

In other words, there is no arm that remains safe to play throughout the episode and there must have been a Significant CW Switch $\nu_i \in [t_\ell, t_{\ell+1})$. □

The following lemma ensures that the last safe arm $a_i^s$ within phase $i$ is not being eliminated before round $\nu_{i+1}$ by any replay `CondaLet`$(s, m)$ that is scheduled in said phase.

**Lemma B.3** (Lemma A.6 for $a_t^s$). *On event $\mathcal{E}$, no run of `CondaLet`$(s, m)$ with $s \in [\nu_i, \nu_{i+1})$ ever eliminates arm $a_i^s$ before round $\nu_{i+1}$.*

*Proof.* Suppose on the contrary that some `CondaLet`$(s, m)$ with $s \in [\nu_i, \nu_{i+1})$ eliminates arm $a_i^s$ before round $\nu_{i+1}$. Then, we must have for some arm $a \in [K]$ and interval $[s_1, s_2] \subseteq [s, \nu_{i+1})$ that

$$C \log(T) K \sqrt{(s_2 - s_1) \vee K^2} < \sum_{t=s_1}^{s_2} \hat{\delta}_t(a, a_i^s), \tag{33}$$

In view of the concentration bound (6), this implies on event $\mathcal{E}$ that

$$c_2 \log(T) K \sqrt{(s_2 - s_1) \vee K^2} < \sum_{t=s_1}^{s_2} \mathbb{E}\left[\hat{\delta}_t(a, a_i^s) \mid \mathcal{F}_{t-1}\right] \leq \sum_{t=s_1}^{s_2} \delta_t(a, a_i^s), \tag{34}$$

where the last inequality holds by merit of (32). Now, by the definition of $a_i^s$ as the last safe arm in phase $i$, it must hold that $\delta_t(a, a_i^s) < \sqrt{K(s_2 - s_1)}$ for all $t \in [\nu_i, \nu_{i+1})$ and all $a \in [K]$. This stands in contradiction to the above which proves Lemma B.3. □

Now, following the same steps as in the proof of Theorem 3.1 in Section A, we obtain for some constant $\tilde{c} > 0$

$$\tilde{R}_2(\ell) \leq \tilde{c} K \log^3(T) \mathbb{E}\left[\sum_{i \in \text{Phases}_{\tilde{S}^{\text{CW}}}(t_\ell, t_{\ell+1})} \sqrt{\nu_{i+1} - \nu_i}\right],$$

where $\nu_{\tilde{S}^{\text{CW}}+1} \coloneqq T$ and $\text{Phases}_{\tilde{S}^{\text{CW}}}(t_1, t_2) \coloneqq \{i \in [\tilde{S}^{\text{CW}}] \colon [\nu_i, \nu_{i+1}) \cap [t_1, t_2) \neq \emptyset\}$. Lastly, in view of the modified Lemma B.2, it follows that (cf. Section A.4)

$$\text{DR}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \frac{\delta_t(a_t^*, a_t) + \delta_t(a_t^*, b_t)}{2}\right] \leq \tilde{2}cK \log^3(T) \sum_{i=0}^{\tilde{S}^{\text{CW}}} \sqrt{\nu_{i+1} - \nu_i}. \tag{35}$$

An application of Jensen's inequality shows that $\text{DR}(T) \leq \tilde{O}(K\sqrt{\tilde{S}^{\text{CW}} T})$.

### B.3 Proof of Corollary 5.1

Recall the definition of the Condorcet Winner Variation from Section 2.2:

$$\tilde{V} \coloneqq \sum_{t=2}^{T} \max_{a \in [K]} \left| P_t(a_t^*, a) - P_{t-1}(a_t^*, a) \right|.$$

We define the CW Variation over phase $[\nu_i, \nu_{i+1})$ as $\tilde{V}_{[\nu_i,\nu_{i+1})} := \sum_{t=\nu_i+1}^{\nu_{i+1}} \max_{a\in[K]} |P_t(a_t^*, a) - P_{t-1}(a_t^*, a)|$. Note that in view of the bound in (35), it suffices to show that $\sum_{i=0}^{\tilde{S}^{\texttt{CW}}} K\sqrt{\nu_{i+1}-\nu_i} \le K\sqrt{T} + \tilde{V}^{1/3}(KT)^{2/3}$.

Consider a phase $[\nu_i, \nu_{i+1})$ with $0 \le i < \tilde{S}^{\texttt{CW}}$. By definition of Significant CW Switches, every arm $a \in [K]$ must satisfy on some interval $[s_1, s_2] \subseteq [\nu_i, \nu_{i+1})$ that

$$\sum_{t=s_1}^{s_2} \delta_t(a_t^*, a) \ge \sqrt{K(s_2 - s_1)}.$$

In particular, this is also the case for the Condorcet winner $a^*_{\nu_{i+1}}$ in round $\nu_{i+1}$. Then, since $\sqrt{s_2 - s_1} > \sum_{t=s_1}^{s_2} \frac{1}{\nu_{i+1}-\nu_i}$, there exists a round $t \in [s_1, s_2]$ such that $\delta_t(a_t^*, a^*_{\nu_{i+1}}) \ge \sqrt{\frac{K}{\nu_{i+1}-\nu_i}}$. We then have

$$\begin{aligned}
\sqrt{\frac{K}{\nu_{i+1}-\nu_i}} &\le \delta_t(a_t^*, a^*_{\nu_{i+1}}) \\
&\le \delta_t(a_t^*, a^*_{\nu_{i+1}}) + \delta_{\nu_{i+1}}(a^*_{\nu_{i+1}}, a_t^*) \\
&\le \delta_t(a_t^*, a^*_{\nu_{i+1}}) - \delta_{\nu_{i+1}}(a_t^*, a^*_{\nu_{i+1}}) \\
&\le \left|\delta_t(a_t^*, a^*_{\nu_{i+1}}) - \delta_{\nu_{i+1}}(a_t^*, a^*_{\nu_{i+1}})\right| \\
&= \left|P_t(a_t^*, a^*_{\nu_{i+1}}) - P_{\nu_{i+1}}(a_t^*, a^*_{\nu_{i+1}})\right| \\
&\le \sum_{s=t+1}^{\nu_{i+1}} \max_{a\in[K]} |P_t(a_t^*, a) - P_{t-1}(a_t^*, a)| \le \tilde{V}_{[\nu_i,\nu_{i+1}]},
\end{aligned}$$

where we used that $\delta_{\nu_{i+1}}(a^*_{\nu_{i+1}}, a_t^*) \ge 0$ and $\delta_{\nu_{i+1}}(a^*_{\nu_{i+1}}, a_t^*) = -\delta_t(a_t^*, a^*_{\nu_{i+1}})$ in the second and third inequality, respectively. We can now apply Hölder's inequality to obtain

$$\begin{aligned}
\sum_{i=0}^{\tilde{S}^{\texttt{CW}}} K\sqrt{\nu_{i+1}-\nu_i} &\le K\sqrt{T} + \sum_{i=0}^{\tilde{S}^{\texttt{CW}}-1} K\sqrt{\nu_{i+1}-\nu_i} \\
&\le K\sqrt{T} + \left(\sum_{i=0}^{\tilde{S}^{\texttt{CW}}} \sqrt{\frac{K}{\nu_{i+1}-\nu_i}}\right)^{1/3} \left(\sum_{i=0}^{\tilde{S}^{\texttt{CW}}} K^{5/4}(\nu_{i+1}-\nu_i)\right)^{2/3} \\
&\le K\sqrt{T} + \left(\sum_{i=0}^{\tilde{S}^{\texttt{CW}}} \tilde{V}_{[\nu_i,\nu_{i+1})}\right)^{1/3} K^{5/6}\, T^{2/3} \\
&= K\sqrt{T} + \tilde{V}^{1/3} K^{5/6}\, T^{2/3}.
\end{aligned}$$

The above dependence on $K$ can be improved to $K^{4/9}$ (which is even smaller than the $K^{2/3}$ dependence in Corollary 5.1) by modifying the definition of Significant CW Switches so that $\nu_{i+1}$ is the first round in $[\nu_i, T)$ such that for all arms $a \in [K]$ there exist rounds $\nu_i \le s_1 < s_2 < \nu_{i+1}$ with

$$\sum_{t=s_1}^{s_2} \delta_t(a_t^*, a) \ge K\sqrt{s_2 - s_1}.$$

It is straightforward to check that Theorem 5.1 holds true also for this definition of Significant CW Switches.

## C More Related Work

Related to the non-stationary dueling bandit problem studied in this paper are adversarial dueling bandits Ailon et al. (2014); Gajane et al. (2015); Saha et al. (2021); Sui et al. (2017). Here, Ailon et al. (2014) was the first to study the dueling bandit problem in an adversarial setup and introduced a popular sparring idea, which has been picked up by many follow-up works Gajane et al. (2015); Dudik et al. (2015); Saha et al. (2021); Saha and Gupta (2022). The settings in Ailon et al. (2014) and Gajane et al. (2015) are restricted to utility-based preference models, where each arm has an associated utility in each round. This entails a complete ordering over the arms in each round, which only covers a small subclass of $[K] \times [K]$ preference matrices. Moreover, Gajane et al. (2015) assume that the feedback includes not only the winner but also the difference in the utilities between the winning and losing arm, which is more similar to MAB feedback and than the $0/1$ one bit preference feedback considered by us. Saha et al. (2021) consider the dueling bandit setup for general adversarial preferences, but they measure performance in terms of (static) regret w.r.t. *Borda-scores*. This measure of regret is very different from our preference-based regret objective. In general, the adversarial dueling bandit problem aims to minimize *static regret* w.r.t. some fixed benchmark $a^*$, whereas we study *dynamic regret* w.r.t. a time-varying benchmark $a_t^*$. As discussed in Section 2, static regret can be an undesirable measure of performance when no single fixed arm represents a reasonably good benchmark over all rounds (see Example 2.1).

Another somewhat related line of work considers the sleeping dueling bandit problem, where the action space is non-stationary (as opposed to the preference sequence). The objective here is to be competitive w.r.t. the best active arm at each round. Saha and Gaillard (2021) studies the setup for adversarial sleeping but assumes a fixed preference matrix across all rounds.