Tree-Guided Identify-Then-Exploit: A Unified Framework of Best Arm Identification and Regret Minimization for Dueling Bandits

Pu Wang
State Key Lab of CAD&CG
Zhejiang University
puwang0508@gmail.com

Yao-Xiang Ding
State Key Lab of CAD&CG
Zhejiang University
dingyx.gm@gmail.com

Abstract

We study N-armed stochastic dueling bandits under the Condorcet-winner assumption with a positive global gap, where three widely adopted objectives are considered: best-arm identification (BAI), weak regret, and strong regret. We propose *Tree-Guided Identify-Then-Exploit* (TG-ITE), the first unified framework to tackle all these objectives to our knowledge. Without requiring transitivity assumptions, we propose a shared tree-guided identification approach to find a high-confidence incumbent within $O(N)$ comparisons. We further propose varied exploitation strategies to utilize this warm-start stage to optimize the specific objectives at hand. For a fixed global gap Δ, this methodology enables our approach to (1) build the first winner-stays-style algorithm to achieve $O(N)$ weak regret; (2) realize the joint optimization of BAI and weak regret with $O(N)$ guarantees for both, eliminating the sub-optimal $O(\log N)$ factors of the existing approach; (3) attain the same $O(N \log T)$ order as specialized strong-regret approaches; and (4) achieve $O(N)$ sample complexity in BAI, matching a worst-case lower bound up to an additional $\log(1/\Delta)$ term. If a lower bound on the minimal winner gap is known, this term can be removed. Our results suggest that a single identification primitive can serve BAI and both regret objectives with linear dependence on N.

1 Introduction

Dueling bandits [Yue and Joachims, 2009; Bengs *et al.*, 2021] model sequential learning from noisy relative feedback: instead of observing a numerical reward, the learner compares two arms and observes a binary preference outcome. This feedback model is natural in ranking [Zoghi *et al.*, 2015b], recommendation [Ma *et al.*, 2025], and human-preference learning [Christiano *et al.*, 2017; Ouyang *et al.*, 2022; Rafailov *et al.*, 2023], where relative judgments are easier to obtain than calibrated absolute scores. In this paper, we study N-armed dueling bandits under the Condorcet winner assumption. In this setting, there exists a single best arm—the Condorcet winner $a^\star$—that beats every other arm, while preferences among the remaining arms may be cyclic. We consider three common objectives in this setting: (1) *best arm identification* (BAI), which requires returning $a^\star$ at a data-dependent stopping time, as well as two regret minimization (RM) objectives, i.e. (2) *online weak regret*, where comparisons involving $a^\star$ are free; (3) *online strong regret*, where only self-comparison $(a^\star, a^\star)$ is free.[1] Our target is to explore the following question: *Can we build a unified algorithmic framework that leads to optimally performed algorithms in all three objectives?*

[1] We use the binary conventions of Chen and Frazier [2017], formalized in Section 2. They differ from terminal simple regret and from gap-weighted cumulative regret (Remark 1).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Gap-oblivious guarantees under Assumption 1. The joint row bounds both the identification time and R_∞^{weak}. Section 5 treats the case of a known lower bound on the winner gap.

Objective	Policy	Guarantee (in expectation)
Exact δ-BAI	TREEASCENT, then stop	$O\big(N\Delta^{-2}[\log(1/\delta)+\log(1/\Delta)]\big)$
Weak regret	Constant-confidence warm start, then SCREENANDREPLACE	$O\big(N\Delta^{-2}\log(1/\Delta)\big)$ for R_∞^{weak}
Joint BAI and weak regret	Report the δ-confidence output, then continue screening	$O\big(N\Delta^{-2}[\log(1/\delta)+\log(1/\Delta)]\big)$
Strong regret, known H	Identify, then self-play	$O\big(N\Delta^{-2}[\log H+\log(1/\Delta)]\big)$
Strong regret, anytime	Repeated fresh identifications and self-play blocks	$O\big(N\Delta^{-2}[\log T+\log(1/\Delta)\log\log T]\big)$

Intuitively, these objectives agree on what must eventually be learned: the Condorcet winner $a^\star$. But they disagree on how the learned winner should be exploited. This observation suggests a natural *identify-then-exploit* design template: first, identify the Condorcet winner with high confidence, then exploit it according to the objective at hand. However, the Explore-Then-Commit (ETC) strategy for stochastic multi-armed bandit (MAB), which is built exactly on this clean idea, has been shown to be necessarily suboptimal on cumulative regret in the seminal work of Garivier *et al.* [2016]. This result is not incidental: The pioneering work of Audibert *et al.* [2010] has already shown that in MAB, BAI requires highly exploratory strategies. These results reveal a fundamental principle in bandits: BAI and RM involve different trade-offs in exploration and exploitation [Bubeck *et al.*, 2009]. Generally, if one algorithm is optimal on one side, then it may not be optimal on another.

On the other hand, allowing to choose two arms at one time seems to make dueling bandits more benign between BAI and RM. This benefit allows an algorithm to utilize one arm for exploration and another for exploitation. This property especially fits for weak regret optimization: One could keep a current incumbent in one coordinate (the weak regret can be eliminated if the choice is correct) while testing challengers in another (to fix the possible wrong choice of incumbent). This is exactly the methodology behind the *winner-stays* (WS) strategy for weak regret [Chen and Frazier, 2017]. Furthermore, this naturally inspires us to utilize identify-then-exploit strategy to jointly optimize BAI and weak regret: Introducing a BAI warm start phase to identify the high-confidence incumbent, and then conducting WS-style exploitation for controlling weak regret. Recently, Ma *et al.* [2025] proposes a WS-style algorithm to achieve this target. Denote by N the number of arms. Under the assumptions that Condorcet winner exists and the arms have a total order, their approach achieves $O(N\log N)$ weak regret, which matches the WS-W algorithm from [Chen and Frazier, 2017], and $O(N\log N)$ BAI sample complexity, $O(\log N)$ factor worse than the existing optimal BAI algorithm. Though promising, important limitations still remain: (1) The $O(\log N)$ gap in BAI remains unclosed; (2) In comparison to the current best $O(N)$ weak regret algorithm from [Saad *et al.*, 2024], which builds upon a different paradigm of online mirror descent with Tsallis regularizer, the result is still sub-optimal and depends on additional total order assumption; (3) As shown in [Chen and Frazier, 2017], WS can be tuned to optimize strong regret, which is not considered in Ma *et al.* [2025].

In this work, we propose *Tree-Guided Identify-Then-Exploit* (TG-ITE), the first unified framework for BAI, weak regret, and strong regret in dueling bandits to our knowledge. The main results are illustrated in Table 1. For each objective, we introduce a shared identification subroutine TREEASCENT, followed by objective-specific exploitation strategies. TREEASCENT serves as the common warm start and identification routine of the following exploitation stages. The key challenge lies in identifying an incumbent with high confidence within $O(N)$ comparisons without requiring assumptions beyond the existence of a Condorcet winner and a positive global gap. We address this by introducing the technique of *tree-based tournament decomposition*, as detailed in Section 3.2. For BAI, gap-oblivious TREEASCENT matches the $O(N)$ bound of Falahatgar *et al.* [2017b] up to an additional $\log(1/\Delta)$ term, without transitivity assumptions. Given a lower bound on the winner gap, a budgeted variant (Section 5) removes this term and matches a worst-case lower bound (Proposition 2). For strong regret, we integrate TREEASCENT with the self-play exploitation strategy, achieving $O(N\log T)$ strong regret for fixed gaps, which matches the leading order of the WS-S algorithm in Chen and Frazier [2017], as well as other classical algorithms specialized for strong

regret minimization [Yue and Joachims, 2011; Yue *et al.*, 2012; Zoghi *et al.*, 2014]. For weak regret, we integrate TREEASCENT with a novel WS-style exploitation strategy SCREENANDREPLACE. The resulting TG-ITE-WEAK algorithm improves the WS-W guarantee from $O(N^2)$, under the Condorcet assumption only, to $O(N)$, matching the linear dependence on N of the current best result in Saad *et al.* [2024]. Notably, TG-ITE-WEAK can be used to simultaneously conduct BAI with $O(N)$ sample complexity and achieve $O(N)$ weak regret, closing both gaps of $O(\log N)$ in [Ma *et al.*, 2025]. Overall, our unified framework suggests that, in dueling bandits, one identification primitive can serve both identification and regret minimization with linear dependence on N.[2]

2 Problem Formulation

We study stochastic dueling bandits with a finite arm set $[N] \coloneqq \{1, \dots, N\}$, $N \geq 2$. At each round t, the learner selects an ordered pair $(i_t, j_t) \in [N]^2$ based on the past and observes an outcome $Y_t \sim \text{Bernoulli}(p_{i_t,j_t})$, where $Y_t = 1$ indicates that i_t wins the duel. Given the selected pair, Y_t is independent of the past. The fixed, unknown preference matrix $P = (p_{i,j})_{i,j\in[N]}$ satisfies $p_{i,j} \in (0,1)$ and $p_{i,j} = 1 - p_{j,i}$. Self-comparisons are allowed and uninformative, since $p_{i,i} = 1/2$.

The Condorcet winner, if it exists, is the unique arm $a^\star$ with $p_{a^\star,j} > 1/2$ for all $j \neq a^\star$. Define the winner gaps $\Delta_i^\star \coloneqq p_{a^\star,i} - 1/2$ and the pairwise gaps $\Delta_{i,j} \coloneqq |p_{i,j} - 1/2|$. We call

$$\Delta^\star \coloneqq \min_{i \neq a^\star} \Delta_i^\star \qquad \text{and} \qquad \Delta \coloneqq \min_{i \neq j} \Delta_{i,j}$$

the *minimal winner gap* and the *global gap*. Then $\Delta^\star > 0$ and $\Delta \leq \Delta^\star$, but $\Delta = 0$ if nonwinners tie.

Assumption 1 (Condorcet winner with positive global gap) *A Condorcet winner exists, and* $\Delta > 0$.

Assumption 1 allows cycles among the nonwinner arms. Section 5 removes the condition $\Delta > 0$ when a lower bound $0 < g \leq \Delta^\star$ on the minimal winner gap is known.

Best-arm identification (δ-BAI). In the fixed-confidence BAI problem, the learner receives a confidence parameter $\delta \in (0,1)$, stops at a data-dependent time τ, and outputs an arm $\widehat{a} \in [N]$. The algorithm is δ-*correct* if $\mathbb{P}(\widehat{a} = a^\star) \geq 1 - \delta$, and its performance is measured by the expected stopping time $\mathbb{E}[\tau]$. A BAI algorithm terminates and produces no further comparisons after stopping. In some settings, the lower bound g is available as side information, which removes the $\log(1/\Delta)$ term from the sample complexity (Section 5).

Regret minimization. At each round t, the learner incurs the (binary) *weak regret* or *strong regret*

$$r_t^{\text{weak}} \coloneqq \mathbf{1}\{i_t \neq a^\star,\ j_t \neq a^\star\}, \qquad r_t^{\text{str}} \coloneqq 1 - \mathbf{1}\{i_t = a^\star,\ j_t = a^\star\}, \qquad R_T^x \coloneqq \sum_{t=1}^{T} r_t^x$$

for $x \in \{\text{weak}, \text{str}\}$. Thus a round incurs zero weak regret whenever at least one of the two compared arms is the Condorcet winner, but zero strong regret only at $(a^\star, a^\star)$. Weak-regret guarantees are stated for the horizonless total $R_\infty^{\text{weak}} \coloneqq \lim_{T\to\infty} R_T^{\text{weak}}$ by default. For strong regret, we treat both *known-horizon* algorithms, which take T as input, and *horizonless* algorithms, which are evaluated at every finite T without knowing it in advance.

Remark 1 (Binary, gap-weighted, and terminal criteria) The binary losses above follow Chen and Frazier [2017]. A common alternative uses the *gap-weighted losses* $r_t^{\text{weak,gw}} = \min\{\Delta_{i_t}^\star, \Delta_{j_t}^\star\}$ and $r_t^{\text{str,gw}} = (\Delta_{i_t}^\star + \Delta_{j_t}^\star)/2$, which have the same zero-loss actions and are related to the binary losses in Appendix D.3. A separate criterion, terminal *simple regret*, evaluates only the final recommendation, for example through $\Delta_{\widehat{a}}^\star$ or $\mathbf{1}\{\widehat{a} \neq a^\star\}$. The losses above instead account for the entire interaction.

Objective-specific instantiations. Our goal is not a single policy that is simultaneously optimal for all three criteria. Instead, the shared tree-guided identification primitive TREEASCENT is instantiated with objective-specific confidence schedules and exploitation tails: termination after a high-confidence identification (BAI), a champion-centric online tail (weak regret), and interleaved identifications with self-play (strong regret).

[2]We delay more discussions of the comparison between our results and existing works to the remarks in Section 4. Due to space limitation, we leave full related work discussions in Appendix A and all formal proofs in the appendix.

Algorithm 1 DUELTEST(u, v, δ): Pairwise testing over two arms

Require: Arms $u \neq v$ and confidence level $\delta \in (0, 1)$.

$n \leftarrow 0$, $S \leftarrow 0$.
repeat
  $n \leftarrow n + 1$.
  Observe $Y_n \sim \text{Bernoulli}(p_{u,v})$, where $Y_n = 1$ means that u wins.
  $S \leftarrow S + Y_n$ and $\widehat{\mu}_n \leftarrow S/n$.
until $|\widehat{\mu}_n - 1/2| > c(n, \delta)$, where $c(n, \delta) = \sqrt{\frac{\log\left(2n(n+1)/\delta\right)}{2n}}$.
if $\widehat{\mu}_n > 1/2$ **then**
  return u.
else
  return v.
end if

Algorithm 2 BKT(S, δ): Balanced knockout tournament on a subset of arms

Require: A nonempty arm set $S \subseteq [N]$ and confidence level $\delta \in (0, 1)$.

$X \leftarrow S$, $r \leftarrow 1$.
while $|X| > 1$ **do**
  Set $\delta_r \leftarrow 6\delta/(\pi^2 r^2)$.
  Pair the arms in X according to a fixed deterministic order; one arm receives a bye if $|X|$ is odd.
  For each pair (u, v), keep the winner returned by DUELTEST(u, v, δ_r).
  Let X be the set of winners together with the possible bye arm; set $r \leftarrow r + 1$.
end while
return the unique arm in X.

3 Tree-Guided Identification

This section develops the identification phase used by all later policies. The primitive TREEASCENT takes an arbitrary starting arm b and returns a high-confidence candidate for the Condorcet winner using $O(N)$ expected comparisons for a fixed global gap and confidence level, uniformly over b. Section 4 then attaches objective-specific exploitation policies for the three objectives.

The design of TREEASCENT starts from a graph-level picture. Under Assumption 1, the arms and their pairwise preferences form a (possibly cyclic) directed graph in which the Condorcet winner $a^\star$ uniquely dominates every other arm. We overlay this graph with a deterministic balanced binary tree $\mathcal{T}$ on the N arms. $\mathcal{T}$ is not learned from data and serves only as a geometric scaffold that prescribes which comparisons to run: it groups the arms into sibling blocks of geometrically growing size around b, so that a size-proportional confidence allocation keeps the total identification cost linear in N. The knockout comparisons induced by $\mathcal{T}$ then carve out a directed beating tree, which contains no cycles even when the preference graph does, and the surviving candidate is the arm at the top of this tree. As the dominating arm, $a^\star$ reaches the top with high probability: it either starts as b or is selected by the tournament on the block containing it, and every later merge preserves it. This capture-and-preservation argument needs only the comparisons involving $a^\star$ to be correct, so the nonwinners never need to be ordered correctly and no transitivity assumption is required. For readability the main text assumes $N = 2^L$, and the general case uses formal byes (Appendix C).

3.1 Statistical Primitives

We first introduce two algorithmic subroutines. The first (Algorithm 1) is a sequential test for a single pair. It repeatedly compares u and v until the empirical mean separates from $1/2$ by an anytime confidence radius, and then returns the empirically better arm. The second (Algorithm 2) is a balanced knockout tournament that returns a representative of an arbitrary subset. Lemma 5 shows that its comparison cost is linear in the subset size, and that if the subset contains $a^\star$, then the tournament returns that arm with high probability.

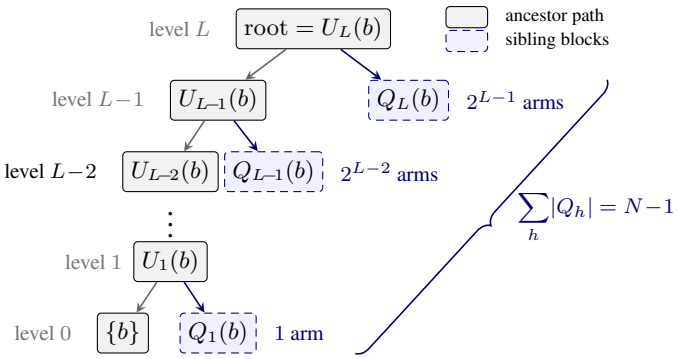


Figure 1: The ancestor path of a starting arm b decomposes the arms into sibling blocks $Q_1(b), \ldots, Q_L(b)$. TREEASCENT walks up the ancestor path, runs a knockout tournament on each sibling block to obtain a representative, and merges that representative with the current champion. If a block contains $a^\star$, the representative is $a^\star$ with high probability. Block sizes grow geometrically ($|Q_h| = 2^{h-1}$), enabling a size-proportional confidence allocation that keeps the total identification cost linear in N.

3.2 Tree-Based Tournament Decomposition

We now make the geometric scaffold concrete. The key mechanism is to decompose global identification of the Condorcet winner into a series of local tournaments with geometrically growing sizes, organized along the ancestor path of the starting arm b. Figure 1 illustrates the construction.

Fix a deterministic balanced binary tree $\mathcal{T}$ on the $N = 2^L$ arms, and write $S(v)$ for the set of arms in the subtree below an internal node v. Now fix a starting arm $b \in [N]$ and walk from the leaf $\{b\}$ up to the root. At each level $h \in \{0, 1, \ldots, L\}$, this walk reaches an ancestor whose subtree covers 2^h arms; we call this set the *ancestor block*

$$U_h(b), \qquad U_0(b) = \{b\}, \qquad U_L(b) = [N], \qquad |U_h(b)| = 2^h.$$

By the binary structure, $U_h(b)$ splits into two halves of equal size: the previous ancestor block $U_{h-1}(b)$, and the arms freshly attached at level h, which we call the *sibling block*

$$Q_h(b) \;:=\; U_h(b) \setminus U_{h-1}(b), \qquad |Q_h(b)| = 2^{h-1}.$$

Thus $Q_1(b)$ is a single arm and $Q_L(b)$ already contains half of all arms. Iterating the split yields the disjoint partition

$$[N] = \{b\} \sqcup Q_1(b) \sqcup \cdots \sqcup Q_L(b), \qquad m_h := |Q_h(b)| = 2^{h-1}, \qquad \sum_{h=1}^{L} m_h = N - 1.$$

The Condorcet winner $a^\star$ therefore either coincides with b or lies in exactly one sibling block. TREEASCENT will identify which by walking the ancestor path and running a knockout tournament on each block.

3.3 Global Winner Identification by TREEASCENT

Based on the tree-based tournament decomposition, we establish the tree-guided identification algorithm as illustrated in Algorithm 3. At each level, the pass runs a knockout tournament on the newly attached sibling block and merges the returned representative with the current champion. The proof only tracks the first level at which $a^\star$ enters the path, and then shows that correct merge tests preserve it.

The subtree budgets β_h are proportional to the block sizes m_h: this proportionality is what keeps $\sum_h m_h \log(1/\beta_h)$ linear in N, rather than $N \log\log N$ as a uniform allocation would give. The merge budgets η_h are uniform and contribute at most $\epsilon/2$ total failure probability.

Theorem 1 (Tree-guided identification) *Under Assumption 1, for every fixed balanced tree $\mathcal{T}$, every starting arm $b \in [N]$, and every $\epsilon \in (0, 1)$, the output of* TREEASCENT$(b, \mathcal{T}, \epsilon)$ *satisfies*

$$\mathbb{P}(\widetilde{b} = a^\star) \geq 1 - \epsilon.$$

Algorithm 3 TREEASCENT$(b, \mathcal{T}, \epsilon)$: Tree-guided winner identification

Require: Starting arm $b \in [N]$; balanced tree $\mathcal{T}$ on $N = 2^L$ leaves; confidence $\epsilon \in (0, 1)$.
Ensure: An identified candidate $\widetilde{b}$.
1: Compute $v_0(b) \subset \cdots \subset v_L(b) = \text{root}$ and sibling blocks $Q_1(b), \ldots, Q_L(b)$.
2: Set $c_0 \leftarrow b$.
3: **for** $h = 1, 2, \ldots, L$ **do**
4: $\quad m_h \leftarrow |Q_h(b)|$.
5: $\quad \beta_h \leftarrow \frac{\epsilon}{2}\frac{m_h}{N-1}, \qquad \eta_h \leftarrow \frac{\epsilon}{2L}$.
6: $\quad g_h \leftarrow \text{BKT}(Q_h(b), \beta_h)$.
7: $\quad c_h \leftarrow \text{DUELTEST}(c_{h-1}, g_h, \eta_h)$.
8: **end for**
9: **return** $\widetilde{b} \leftarrow c_L$.

Moreover, if $\tau_{\text{TA}}(b, \epsilon)$ denotes the total number of comparisons used by the pass, then

$$\mathbb{E}[\tau_{\text{TA}}(b, \epsilon)] \le \frac{CN}{\Delta^2}\left(\log\frac{1}{\epsilon} + \log\frac{1}{\Delta} + 1\right)$$

for a universal constant $C > 0$. The same statement holds for arbitrary $N \ge 2$ under the formal-bye implementation in Appendix C.

We say that a subroutine call uses *fresh samples* when every comparison outcome drawn during the call has not been used by the algorithm before. The pair choices inside the call may still be adaptive. Since Theorem 1 holds for every starting arm, it also applies when b is chosen based on earlier comparisons, provided the pass uses fresh samples.

4 Objective-Specific Exploitation

4.1 BAI: Identify Then Stop

The BAI policy TG-ITE-BAI runs TREEASCENT$(b_0, \mathcal{T}, \delta)$, outputs the returned arm $\widehat{a}$, and stops.

Corollary 1 (Fixed-confidence δ-BAI) *Under Assumption 1, for every $\delta \in (0, 1)$,* TG-ITE-BAI *satisfies*

$$\mathbb{P}(\widehat{a} = a^\star) \ge 1 - \delta$$

and its expected stopping time is bounded by

$$\mathbb{E}[\tau_{\text{BAI}}] \le \frac{CN}{\Delta^2}\left(\log\frac{1}{\delta} + \log\frac{1}{\Delta} + 1\right).$$

Remark. For fixed gap and confidence parameters, the result in Corollary 1 guarantees $O(N)$ sample complexity without requiring commonly adopted assumptions such as strong stochastic transitivity and stochastic triangle inequality. Proposition 2 in Appendix C.4 gives a worst-case lower bound $\Omega(N\Delta^{-2}\log(1/\delta))$ on instances in which every pairwise gap equals Δ, even for learners that know Δ. The upper bound matches this rate up to an additional $\log(1/\Delta)$ term, which comes from the anytime confidence radius of DUELTEST. The detailed discussions are provided in Section A.2.

4.2 Weak Regret: Winner-Centric Exploitation

The weak-regret policy uses TREEASCENT for a constant-confidence warm start to identify an initial incumbent, then runs the iterative winner-centric exploitation subroutine SCREENANDREPLACE (Algorithm 4). In each epoch, the algorithm screens all challengers against the current incumbent and, if any appears to beat it, runs a replacement tournament on the incumbent together with those apparent beaters. The resulting weak-regret algorithm TG-ITE-WEAK is illustrated in Algorithm 5. Each screening comparison includes the incumbent, so when the incumbent is $a^\star$, weak regret arises only if a challenger wrongly appears to beat it, which happens with probability at most ρ per challenger. When the incumbent is not $a^\star$, the screen also compares it with $a^\star$, and the epoch returns $a^\star$ with probability at least $1 - \rho - \kappa$. Because Algorithm 5 shrinks ρ_s and κ_s geometrically, such errors quickly become rare and the total weak regret stays finite.

Algorithm 4 SCREENANDREPLACE(c, ρ, κ)

Require: Incumbent arm c; screening confidence ρ; replacement confidence κ.

$B \leftarrow \emptyset$.
for each $j \in [N] \setminus \{c\}$ in a fixed order **do**
  $w \leftarrow$ DUELTEST(c, j, ρ) using fresh samples.
  if $w = j$ **then**
    $B \leftarrow B \cup \{j\}$.
  end if
end for
if $B = \emptyset$ **then**
  return c.
else
  return BKT$(B \cup \{c\}, \kappa)$ using fresh samples.
end if

Algorithm 5 TG-ITE-WEAK$(b_0, \mathcal{T})$

Require: Starting arm $b_0 \in [N]$; fixed balanced tree $\mathcal{T}$.

Run a constant-confidence warm start $c_0 \leftarrow$ TREEASCENT$(b_0, \mathcal{T}, 1/4)$.
for $s = 1, 2, 3, \ldots$ **do**
  Set $\rho_s \leftarrow 4^{-(s+2)}$ and $\kappa_s \leftarrow 4^{-(s+2)}$.
  $c_s \leftarrow$ SCREENANDREPLACE$(c_{s-1}, \rho_s, \kappa_s)$.
end for

Theorem 2 (Online weak regret with linear dependence on N) *Under Assumption 1, the weak regret of Algorithm 5 satisfies*

$$\mathbb{E}[R_\infty^{\mathrm{weak}}] \le \frac{CN}{\Delta^2}\left(\log\frac{1}{\Delta} + 1\right)$$

for a universal constant $C > 0$. In particular, $R_\infty^{\mathrm{weak}} < \infty$ almost surely.

Furthermore, by adjusting the warm start confidence in Algorithm 5 to δ, we can obtain an algorithm, TG-ITE-DUAL, jointly optimizing BAI and weak regret. It reports c_0 as its BAI answer when the warm start ends and then continues screening.

Corollary 2 (Joint BAI and weak regret) *Under Assumption 1, for every $\delta \in (0, 1)$,* TG-ITE-DUAL *satisfies $\Pr(c_0 = a^\star) \ge 1 - \delta$, and its answer time τ_{Dual} and weak regret satisfy*

$$\max\left\{\mathbb{E}[\tau_{\mathrm{Dual}}],\ \mathbb{E}[R_\infty^{\mathrm{weak}}]\right\} \le \frac{CN}{\Delta^2}\left(\log\frac{1}{\delta} + \log\frac{1}{\Delta} + 1\right).$$

Its identification stage is the same TREEASCENT call as TG-ITE-BAI, so with the same inputs and feedback, both report the same arm at the same time. The $\log(1/\delta)$ term in the weak regret pays for this high-confidence warm start.

Remark. Under the positive global gap assumption, our approach is the first winner-stays (WS) style algorithm that achieves $O(N)$ weak regret, improving the original WS-W algorithm [Chen and Frazier, 2017], whose guarantee is $O(N^2)$ under the Condorcet assumption and $O(N \log N)$ under a total order. Corollary 2 improves the sub-optimal $O(\log N)$ factors in Ma *et al.* [2025], yielding bounds that are linear in N for both BAI and weak regret, although their sample-complexity guarantee is a high-probability bound, whereas ours bounds the expected answer time.

Unlike these binary-loss analyses, Saad *et al.* [2024] bound gap-weighted weak regret. Converting the losses round by round (Appendix D.3) shows that their algorithm attains binary weak regret $O(N/(\Delta^\star)^2)$ and that this order is unavoidable in the worst case. Theorem 2 matches its linear dependence on N but depends on Δ instead of $\Delta^\star$. Proposition 1 shows that Algorithm 5 actually incurs this dependence, because its pairwise tests can spend arbitrarily long on two nearly tied nonwinners even when the winner gap stays fixed.

Proposition 1 (Global-gap obstruction for the weak-regret policy) *Fix $d \in (0, 1/4)$. For every sufficiently small $\gamma \in (0, d)$, there is a four-arm instance, fixed tree, and starting arm with $\Delta^\star = d$*

Algorithm 6 TG-ITE-STRONG-ANYTIME$(b_0, \mathcal{T})$

Require: Starting arm b_0; fixed balanced tree $\mathcal{T}$.

$c_0 \leftarrow b_0$.
for $s = 1, 2, 3, \ldots$ **do**
  Set $\epsilon_s \leftarrow 2^{-2^s}$ and $B_s \leftarrow 2^{2^s}$.
  $c_s \leftarrow$ TREEASCENT$(c_{s-1}, \mathcal{T}, \epsilon_s)$ using fresh samples.
  Play (c_s, c_s) for the next B_s rounds.
end for

and $\Delta = \gamma$ such that Algorithm 5 satisfies

$$\mathbb{E}[R_\infty^{\mathrm{weak}}] \geq c_d \, \gamma^{-2} \log(1/\gamma)$$

for some $c_d > 0$ independent of γ. Consequently, the unchanged weak policy has no finite uniform bound depending only on N and $\Delta^\star$.

4.3 Strong Regret: Self-Play Exploitation

Binary strong regret has the opposite geometry: any comparison of distinct arms incurs one unit of regret, even if one of them is $a^\star$. The policy therefore identifies a reliable incumbent and self-plays.

Known horizon. If the horizon H is known, the policy (Algorithm 8 in Appendix E.1) runs TREEASCENT at confidence $\epsilon_H = 1/(H+1)$ and then self-plays.

Theorem 3 (Known-horizon strong regret) *Under Assumption 1, Algorithm 8 satisfies, for every horizon $H \geq 1$,*

$$\mathbb{E}[R_H^{\mathrm{str}}] \leq \frac{CN}{\Delta^2} \left(\log(H+1) + \log \frac{1}{\Delta} + 1 \right) + 1.$$

Unknown horizon. When the horizon is unknown, the policy cannot choose a single confidence level $1/(T+1)$. An anytime alternative is to interleave increasingly confident identified incumbents with increasingly long self-play blocks.

Theorem 4 (Horizonless strong regret) *Under Assumption 1, Algorithm 6 satisfies, for every $T \geq 2$,*

$$\mathbb{E}[R_T^{\mathrm{str}}] \leq \frac{CN}{\Delta^2} \left(\log(T+1) + (\log\log(T+2) + 1) \left(\log \frac{1}{\Delta} + 1 \right) \right) + C'(\log\log(T+2) + 1)$$

for universal constants $C, C' > 0$. For any fixed $\Delta > 0$, the leading term as $T \to \infty$ is $O(N\Delta^{-2} \log T)$.

Remark. Existing strong regret specialized algorithms [Yue and Joachims, 2011; Yue *et al.*, 2012; Saha and Gaillard, 2022] usually achieve $O(N \log T/\Delta^\star)$ under gap-weighted loss. Converting the losses as in Section 4.2 turns this rate into binary strong regret $O(N \log T/(\Delta^\star)^2)$, and Proposition 3 in Appendix E.3 shows that this order is asymptotically unavoidable for strongly consistent policies on total orders in which every pairwise gap equals $\Delta^\star$. In comparison, for fixed gaps our results are optimal in N, T, but depend on Δ instead of $\Delta^\star$. Under gap-weighted loss, our general bounds retain this dependence on Δ, because TREEASCENT also tests nonwinner pairs at high confidence. Improving this dependence while keeping an explicit BAI guarantee remains open.

5 Known-Gap Budgeted Extension

This section assumes only that a Condorcet winner exists and that the learner knows a lower bound $0 < g \leq \Delta^\star$ on the minimal winner gap. As in Section 3, whenever every test involving $a^\star$ returns $a^\star$, BKT returns $a^\star$ if it is a candidate, and TREEASCENT and SCREENANDREPLACE return $a^\star$. In each such test, $a^\star$ beats its opponent by at least $\Delta^\star \geq g$. We can therefore replace DUELTEST with BUDGETEDDUELTEST (Algorithm 9 in Appendix F), a variant of the comparison primitive of Falahatgar *et al.* [2017b]. It caps each test at a budget fixed by g and δ, stops earlier once the empirical mean separates from $1/2$ by its confidence radius minus g, and returns the arm with more wins. A test involving $a^\star$ still returns $a^\star$ with the required confidence. A test between two nonwinners, even a tied one, can stop early or at the budget, and its answer does not affect correctness.

Lemma 1 (Validity of the budgeted pairwise test) *For every pair $u \neq v$, Algorithm 9 uses at most $\lceil (2g^2)^{-1} \log(4/\delta) \rceil$ comparisons deterministically. If $|p_{u,v} - 1/2| \geq g$, then it returns the better arm with probability at least $1 - \delta$.*

The tree, the confidence schedules, and the exploitation rules stay unchanged, so the pairwise guarantee of Lemma 1 propagates to all objectives as follows.

Theorem 5 (Known-gap budgeted TG-ITE) *Assume that a Condorcet winner exists and that the learner knows $0 < g \leq \Delta^\star$, and run the policies of Section 4 with* BUDGETEDDUELTEST *in place of* DUELTEST. *For universal constants $C, C' > 0$ and every $\delta \in (0, 1)$, the following bounds hold.*

$$\textit{Exact } \delta\textit{-BAI: } \Pr(\widehat{a} = a^\star) \geq 1 - \delta, \quad \tau_{\mathrm{BAI}} \leq \frac{CN}{g^2}\left(\log\frac{1}{\delta} + 1\right) \textit{ deterministically;}$$

$$\textit{Weak regret: } \mathbb{E}[R_\infty^{\mathrm{weak}}] \leq \frac{CN}{g^2};$$

$$\textit{Weak regret with a } \delta\textit{-BAI warm start: } \quad \mathbb{E}[R_\infty^{\mathrm{weak}}] \leq \frac{CN}{g^2}\left(\log\frac{1}{\delta} + 1\right);$$

$$\textit{Strong regret, known } H: \quad \mathbb{E}[R_H^{\mathrm{str}}] \leq \frac{CN}{g^2}\left(\log(H+1) + 1\right) + 1;$$

$$\textit{Strong regret, anytime: for every } T \geq 2,$$

$$\mathbb{E}[R_T^{\mathrm{str}}] \leq \frac{CN}{g^2}\left(\log(T+1) + \log\log(T+2) + 1\right) + C'(\log\log(T+2) + 1).$$

Compared with Table 1, the global gap Δ is replaced by g, which can be as large as $\Delta^\star$, the $\log(1/\Delta)$ terms disappear, and the BAI stopping time is bounded deterministically. Since only the pairwise test has changed, all of the dependence on Δ in Table 1 comes from gap-oblivious testing rather than from the tree. In the setting of Proposition 2, with g equal to the common gap, the BAI budget matches the lower bound up to constants.

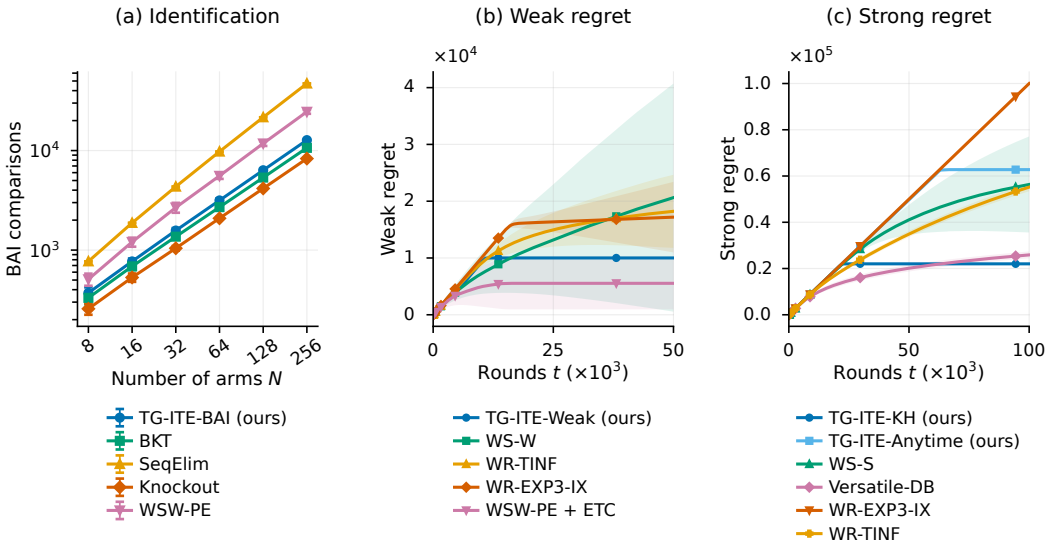


Figure 2: (a) Cyclic BAI. (b) Perturbed cyclic weak regret. (c) Cyclic strong regret. Error bars and bands show one sample SD over 30 seeds.

6 Experiments

On synthetic preference matrices, we evaluate TG-ITE-BAI, TG-ITE-WEAK, and the known-horizon and anytime strong-regret policies, abbreviated TG-ITE-KH and TG-ITE-ANYTIME. The

baselines are BKT (Algorithm 2 on all arms), SEQELIM [Falahatgar *et al.*, 2017a], KNOCKOUT [Falahatgar *et al.*, 2017b], WS-W and WS-S [Chen and Frazier, 2017], WR-TINF and WR-EXP3-IX [Saad *et al.*, 2024], WSW-PE [Ma *et al.*, 2025], and VERSATILE-DB [Saha and Gaillard, 2022]. Since WSW-PE stops after outputting an arm, its online runs add an explore-then-commit continuation (WSW-PE+ETC). In the cyclic instances, the winner beats every other arm with gap $\Delta^\star$ and the nonwinners beat one another cyclically with gap d, so no total order exists and $\Delta = \min\{\Delta^\star, d\}$. The main weak-regret comparison uses a randomly perturbed version. All online costs include identification. Appendix H specifies the full protocol and reports all results.

Comparison. On cyclic instances with $(\Delta^\star, d) = (.32, .38)$ and $\delta = .1$ (Figure 2(a)), the cost of TG-ITE-BAI doubles with N, as Corollary 1 predicts. TG-ITE-BAI uses 10% to 20% more comparisons than BKT. This overhead is expected, since TREEASCENT is designed as a shared stage for all three objectives rather than for BAI alone. Knockout needs even fewer comparisons, but it is given the tolerance $\epsilon = .95\Delta^\star$, computed from the true winner gap. On the perturbed instance with $N = 256$, $\Delta^\star = .26$, and $T = 50{,}000$ (Figure 2(b)), the weak regret of TG-ITE-Weak stays flat once its warm start identifies the winner, as Theorem 2 predicts, and ends below those of WR-TINF, WR-EXP3-IX, and WS-W with a much smaller spread across seeds. Because its incumbent takes part in every duel, WSW-PE+ETC accrues no weak regret once that incumbent is the winner and ends lower on average, but its spread across seeds is far larger. On the cyclic instance with $N = 256$, $(\Delta^\star, d) = (.30, .38)$, and $T = 100{,}000$ (Figure 2(c)), the strong regret of TG-ITE-KH stays flat once self-play begins, as Theorem 3 predicts, and ends below those of all baselines. Versatile-DB comes closest but is still growing at the horizon. Without the horizon, TG-ITE-Anytime repeats identification at increasing confidence levels and ends above WR-TINF and WS-S. It discards past outcomes at the start of every phase, as its analysis assumes. In a paired ablation on this configuration with $T = 10^6$ (Table 9), it uses 2.94 times as many new identification comparisons as a heuristic that caches pair statistics across phases, since the first phase already identifies the winner in every run and later phases only confirm it.

Dependence on the global gap. With $\Delta^\star$ held at .30, the mean cost of TG-ITE-BAI grows by a factor of about 4.6 each time Δ halves, consistent with the $\Delta^{-2}\log(1/\Delta)$ dependence in Corollary 1 (Table 5). TG-ITE-Weak and TG-ITE-KH begin with a TREEASCENT pass, so their regret grows at a similar rate. Once $\Delta \le .05$, this pass no longer finishes within $T = 10^6$ (Tables 6 and 7). WR-TINF's guarantee depends on $\Delta^\star$ instead, so the comparison with it hinges on which gap is small. TG-ITE-Weak falls behind it on this sweep once $\Delta \le .20$, but with $\Delta^\star = \Delta = .05$ and nonwinner gap .20, it incurs about a tenth as much weak regret (Table 8). Given $g = \Delta^\star$, the known-gap variants of Section 5 cap every test, so their cost and regret level off, as Theorem 5 predicts.

7 Limitations and Future Work

Nearly tied nonwinners can dominate the cost of the gap-oblivious policies even when $\Delta^\star$ is large (Proposition 1, Section 6). In the Gaussian formulation of Jamieson *et al.* [2014, Appendix A, Theorem 6], Farrell's theorem shows that a test with a fixed error level for a single pair with unknown gap γ requires $\Omega(\gamma^{-2}\log\log(1/\gamma))$ expected samples along some gaps tending to zero. Thus uniformly correct, gap-oblivious pairwise tests cannot have a uniform $O(\gamma^{-2})$ expected stopping-time bound. Section 5 removes this cost and the $\log(1/\Delta)$ term at the price of knowing a lower bound $0 < g \le \Delta^\star$ in advance, so its policies are no longer completely gap-oblivious. Our known-gap experiments supply $g = \Delta^\star$ as an oracle input, and learning such a bound online while keeping these guarantees remains open. Finally, the strong-regret bounds hold only in expectation, and a rare wrong output followed by a long self-play block can produce regret near the horizon (Appendix H.3). We list three directions for future work. *(i)* For a known horizon H, we conjecture that relaxing the BAI objective in TREEASCENT to a coarse capture pass followed by direct incumbent-versus-challenger verification attains gap-weighted strong regret $O(N\Delta^{-2}\log(1/\Delta)) + O(\sum_{i\neq a^\star}\log H/\Delta_i^\star)$. Proving this requires controlling the cost incurred while the incumbent is incorrect. *(ii)* Caching statistics across phases of the strong-regret policy saves substantial cost in Section 6 but has no guarantee yet. A martingale-based analysis could provide one. *(iii)* Without gap information, iterated-logarithm confidence sequences [Howard *et al.*, 2021] could reduce the $\log(1/\Delta)$ term to $\log\log(1/\Delta)$ if the error and expected stopping-time bounds of DUELTEST carry over.

Acknowledgement

This work was supported by National Natural Science Foundation of China (U23A20311, 62206245). The authors declare no competing interests. We would like to thank the anonymous reviewers for their constructive suggestions, as well as Yu-Hu Yan, Peng Zhao, and Fang Kong for insightful discussions.

References

Arpit Agarwal, Rohan Ghuge, and Viswanath Nagarajan. Batched dueling bandits. In *International Conference on Machine Learning*, pages 89–110. PMLR, 2022.

Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*, pages 856–864. PMLR, 2014.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *The 23rd Conference on Learning Theory*, pages 41–53, 2010.

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*, 22(7):1–108, 2021.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

Robert Busa-Fekete, Eyke Huellermeier, and Balázs Szörényi. Preference-based rank elicitation using statistical models: The case of mallows. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1071–1079. PMLR, 2014.

Bangrui Chen and Peter I Frazier. Dueling bandits with weak regret. In *International Conference on Machine Learning*, pages 731–739. PMLR, 2017.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Rémy Degenne, Thomas Nedelec, Clément Calauzènes, and Vianney Perchet. Bridging the gap between regret minimization and best arm identification, with application to a/b tests. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1988–1996. PMLR, 2019.

Yihan Du, Siwei Wang, and Longbo Huang. Dueling bandits: From two-dueling to multi-dueling. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 348–356. International Foundation for Autonomous Agents and Multiagent Systems, 2020.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 563–587. PMLR, 2015.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006.

Moein Falahatgar, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. Maxing and ranking with few assumptions. In *Advances in Neural Information Processing Systems*, 2017.

Moein Falahatgar, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Maximum selection and ranking under noisy comparisons. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1088–1096. PMLR, 2017.

Moein Falahatgar, Ayush Jain, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. The limits of maxing, ranking, and preference learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1427–1436. PMLR, 2018.

Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in neural information processing systems*, 25, 2012.

Aurelien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Björn Haddenhorst, Viktor Bengs, Jasmin Brandt, and Eyke Hüllermeier. Testification of condorcet winners in dueling bandits. In *Uncertainty in Artificial Intelligence*, pages 1195–1205. PMLR, 2021.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 423–439. PMLR, 2014.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1141–1154. PMLR, 2015.

Lanjihong Ma, Yao-Xiang Ding, Zhen-Yu Zhang, and Zhi-Hua Zhou. Achieving nearly-optimal regret and sample complexity in dueling bandits with applications in online recommendations. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1008–1019, 2025.

Arnab Maiti, Ross Boczar, Kevin Jamieson, and Lillian Ratliff. Near-optimal pure exploration in matrix games: A generalization of stochastic bandits & dueling bandits. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2602–2610. PMLR, 2024.

Soheil Mohajer, Changho Suh, and Adel Elmahdy. Active learning for top-k rank aggregation from noisy comparisons. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2488–2497. PMLR, 2017.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741, 2023.

Wenbo Ren, Jia Kevin Liu, and Ness Shroff. On sample complexity upper and lower bounds for exact ranking from noisy comparisons. *Advances in Neural Information Processing Systems*, 32, 2019.

Wenbo Ren, Jia Liu, and Ness Shroff. The sample complexity of best-k items selection from pairwise comparisons. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8051–8072. PMLR, 2020.

El Mehdi Saad, Alexandra Carpentier, Tomáš Kocák, and Nicolas Verzelen. On weak regret analysis for dueling bandits. *Advances in Neural Information Processing Systems*, 37:40752–40791, 2024.

El Mehdi Saad, Victor Thuot, and Nicolas Verzelen. The sampling complexity of condorcet winner identification in dueling bandits, 2026. arXiv:2603.15189.

Aadirupa Saha and Pierre Gaillard. Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19011–19026. PMLR, 2022.

Yanan Sui, Vincent Zhuang, Joel W. Burdick, and Yisong Yue. Multi-dueling bandits with dependent arms. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, volume R15 of *Proceedings of Machine Learning Research*, pages 551–560. PMLR, 2017.

Huasen Wu and Xin Liu. Double thompson sampling for dueling bandits. *Advances in neural information processing systems*, 29, 2016.

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, 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.

Zixin Zhong, Wang Chi Cheung, and Vincent Tan. Achieving the pareto frontier of regret minimization and best arm identification in multi-armed bandits. *Transactions on Machine Learning Research*, 2023.

Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten de Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 10–18. PMLR, 2014.

Masrour Zoghi, Zohar S Karnin, Shimon Whiteson, and Maarten De Rijke. Copeland dueling bandits. *Advances in neural information processing systems*, 28, 2015.

Masrour Zoghi, Shimon Whiteson, and Maarten De Rijke. MergeRUCB: A method for large-scale online ranker evaluation. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 17–26, 2015.

A Related Work

This section reviews the work most related to our setting. We group prior results by objective: regret minimization, pure exploration, and attempts to unify multiple objectives.

A.1 Regret minimization for dueling bandits

Dueling bandits were introduced to model online learning from noisy pairwise preferences, with early applications in interactive information retrieval and online ranker evaluation [Yue and Joachims, 2009; Yue *et al.*, 2012]. The regret-minimization literature has since developed along several assumptions and benchmarks. Early algorithms such as Interleaved Filter and Beat-the-Mean rely on latent-order or stochastic-transitivity-type assumptions, and achieve logarithmic-in-T regret under such structure [Yue and Joachims, 2011; Yue *et al.*, 2012]. Later work relaxed or changed the structural requirements: Ailon *et al.* [2014] reduce dueling bandits to conventional cardinal bandits; Zoghi *et al.* [2014] propose RUCB for the Condorcet setting and prove an $O(N \log T)$ finite-time regret bound; Zoghi *et al.* [2015b] introduce MergeRUCB for large-scale online ranker evaluation; and Komiyama *et al.* [2015] give an asymptotic lower bound based on information divergence together with RMED,

the first asymptotically matching algorithm under the Condorcet assumption. Other works study alternative winner notions or extensions, including Copeland winners, Thompson-sampling-style methods, contextual dueling bandits, multi-dueling bandits, dependent-arm models, and batched feedback [Zoghi *et al.*, 2015a; Wu and Liu, 2016; Dudík *et al.*, 2015; Sui *et al.*, 2017; Du *et al.*, 2020; Agarwal *et al.*, 2022]. These directions are mostly orthogonal to our finite-arm Condorcet setting.

Weak regret. Weak regret counts a round as regret-free whenever at least one selected arm is the Condorcet winner. This objective can be much easier than strong regret because one coordinate can exploit the current incumbent while the other coordinate explores challengers. Chen and Frazier [2017] made this distinction explicit and proposed Winner-Stays-Weak (WS-W), whose expected cumulative weak regret is $O(N^2)$ under the Condorcet-winner assumption and improves to $O(N \log N)$ when the arms follow a total order. Their Theorem 1 gives the explicit total-order bound $O(N\Delta^{-6} \log N)$, where Δ is a minimum pairwise gap. More recently, Saad *et al.* [2024] showed that weak regret under the sole Condorcet-winner assumption is governed by the full preference matrix: they prove an $\Omega(N/\Delta^\star)$ lower bound in the large-optimality-gap regime, where $\Delta^\star := \min_{i \neq a^\star}(p_{a^\star,i} - 1/2)$, and propose WR-TINF, which matches this order in that regime; they also give another algorithm for regimes where the Condorcet gap is not the right hardness parameter. The closest joint RM–BAI work is Ma *et al.* [2025], which builds a reduction from weak-regret minimization to fixed-confidence BAI; instantiated with WS-W, their WSW-PE algorithm preserves the weak-regret guarantee $O(N\Delta^{-6} \log N)$ and identifies the Condorcet winner with high probability using $O\Big(R_w + \frac{N}{(\Delta^\star)^2} \log \frac{N}{\Delta^\star \delta}\Big)$ comparisons. In contrast, our TG-ITE-Weak keeps the Winner-Stays-style incumbent/challenger structure but replaces the online discovery phase with a tree-guided warm start and screening tail. Under the positive global-gap assumption $\Delta := \min_{i \neq j} |p_{i,j} - 1/2|$, it satisfies

$$\mathbb{E}[R_\infty^{\mathrm{weak}}] \leq O\bigg(\frac{N}{\Delta^2}\left(\log \frac{1}{\Delta} + 1\right)\bigg).$$

Thus our weak-regret guarantee is time-independent and linear in N without the $\log N$ factor of WS-W/WSW-PE. Compared with Saad *et al.* [2024], our uniform bound depends on the global minimum gap Δ rather than the instance-dependent optimal gap $\Delta^\star$, which makes it coarser in its gap dependence. However, this is the cost for a simple, reusable identification primitive that supports BAI, weak regret, and strong regret simultaneously. This global-gap parameterization also follows the convention of the classical noisy-comparison literature [Feige *et al.*, 1994] and of the Winner-Stays line of work [Chen and Frazier, 2017], which state guarantees in terms of a gap governing all pairwise comparisons rather than only those involving the winner.

Strong regret. Strong regret requires both selected coordinates to be the Condorcet winner. This is the geometry captured by the standard Condorcet regret used in much of the dueling-bandit literature, where pulling a suboptimal arm in either coordinate incurs gap-weighted loss. Yue *et al.* [2012] prove an $\Omega(N \log T)$ worst-case expected-regret lower bound under the Condorcet-winner assumption, and Interleaved Filter achieves $O(N \log T)$ under strong stochastic transitivity. Beat-the-Mean further gives a high-probability regret bound of order $O\Big(\sum_{i \neq a^\star} \frac{\gamma^8}{\Delta_i} \log T\Big)$, $\Delta_i := p_{a^\star,i} - 1/2$ under relaxed stochastic transitivity and stochastic triangle inequality [Yue and Joachims, 2011]. RUCB obtains an $O(N \log T)$ finite-time bound in the Condorcet setting [Zoghi *et al.*, 2014], while RMED is asymptotically optimal without additional structural assumptions [Komiyama *et al.*, 2015]. Saha and Gaillard [2022] later give a reduction-based best-of-both-worlds algorithm, Versatile-DB, which is optimal for both stochastic and adversarial preferences and, in the stochastic Condorcet case, achieves the gap-dependent bound $O\Big(\sum_{i \neq a^\star} \frac{\log T}{\Delta_i}\Big)$ against the Condorcet-winner benchmark. In the binary strong-regret formulation of Chen and Frazier [2017], WS-S obtains $O(N^2 + N \log T)$ expected cumulative strong regret, improved to $O(N \log N + N \log T)$ under total order. Our strong-regret policy instead identifies a reliable incumbent and self-plays. With known horizon T, it guarantees

$$\mathbb{E}[R_T^{\mathrm{str}}] \leq O\bigg(\frac{N}{\Delta^2}\left(\log T + \log \frac{1}{\Delta} + 1\right)\bigg),$$

and the horizonless version has the same leading $O(N\Delta^{-2} \log T)$ term up to lower-order $\log \log T$ factors. Therefore, our strong-regret result removes the N^2 exploration term of WS-S and matches the linear-in-N, logarithmic-in-T behavior of specialized strong-regret algorithms, while using the same TREEASCENT primitive as the BAI and weak-regret policies.

A.2 Pure Exploration for Dueling Bandits

Pure exploration has a long history in ordinary stochastic bandits, where best-arm identification is separated from cumulative regret minimization. In the standard stochastic MAB setting, fixed-confidence and fixed-budget BAI have been extensively studied, with sample complexity governed by inverse squared reward gaps, such as $\sum_{i \neq i^\star} \Delta_i^{-2}$, up to logarithmic factors in the confidence parameter [Even-Dar *et al.*, 2006; Audibert *et al.*, 2010; Gabillon *et al.*, 2012; Kaufmann *et al.*, 2016]. Dueling bandits replace numerical rewards by noisy pairwise comparisons, so pure exploration first requires specifying what the "best" arm means. Common winner notions include Condorcet, Borda, and Copeland winners; in this work we focus on the Condorcet winner [Bengs *et al.*, 2021].

A large body of pure-exploration work for dueling bandits studies winner identification under structural assumptions on the preference matrix. Let $\Delta_{i,j} = p_{i,j} - 1/2$ denote the signed pairwise gap. Weak stochastic transitivity (WST) requires

$$\Delta_{i,j} \geq 0,\ \Delta_{j,k} \geq 0 \quad \Longrightarrow \quad \Delta_{i,k} \geq 0,$$

which induces a transitive preference relation and hence, up to ties or tie-breaking, a latent total order. Strong stochastic transitivity (SST) is stronger: for ordered arms $i \succ j \succ k$, it requires

$$\Delta_{i,k} \geq \max\{\Delta_{i,j}, \Delta_{j,k}\}.$$

Thus SST not only implies a latent total order but also imposes monotonicity on comparison margins. This distinction matters statistically.

Under total-order-type assumptions, Yue and Joachims [2011] proposed BTM-PAC, which identifies an ϵ-optimal arm with sample complexity $O\big(\frac{N}{\epsilon^2} \log \frac{N}{\epsilon\delta}\big)$. Subsequent noisy-comparison work sharpened this dependence under stronger structural assumptions. In particular, Falahatgar *et al.* [2017b] gave a knockout-style maximum-selection algorithm under SST and stochastic triangle inequality (STI) with $O\big(\frac{N}{\epsilon^2}\big(\log \frac{1}{\delta}\big)\big)$ comparisons, which is optimal up to constants for PAC maxing. Later work clarified how much structure is needed: Falahatgar *et al.* [2017a] showed that SST alone is enough for linear-complexity PAC maxing, while Falahatgar *et al.* [2018] showed that WST alone can require $\Omega(N^2)$ comparisons for general PAC maxing, even though WST still induces a latent order.

Another line studies exact identification rather than approximate maxing. Exact Condorcet identification can be viewed as the regime $0 < \epsilon < \Delta^\star$, where $\Delta^\star := \min_{j \neq a^\star} |p_{a^\star,j} - 1/2|$ is the Condorcet minimum gap. Classical noisy-comparison results analyze selection when each comparison errs with a fixed probability [Feige *et al.*, 1994]. On instances in which every pairwise gap equals Δ, any δ-correct algorithm needs $\Omega\big(\frac{N}{\Delta^2} \log \frac{1}{\delta}\big)$ comparisons, even when Δ is known (Proposition 2). Under parametric or ranking-style assumptions, Busa-Fekete *et al.* [2014] proposed MallowsMPI for exact BAI under a Mallows model, with sample complexity $O\big(\frac{N}{\Delta^2} \log \frac{N}{\delta\Delta}\big)$. Mohajer *et al.* [2017] studied active top-K rank aggregation from noisy comparisons using fixed-repetition tournaments with comparison budgets calibrated to the gaps. More recent exact-ranking and best-k selection results further refine this picture: Ren *et al.* [2019] characterize exact ranking from noisy comparisons, and Ren *et al.* [2020] study PAC and exact best-k selection under SST and STI, obtaining matching PAC best-k bounds of order $\Theta\big(\frac{N}{\epsilon^2} \log \frac{k}{\delta}\big)$ for $1 \leq k \leq N/2$ and exact top-1 guarantees that are optimal up to logarithmic or doubly-logarithmic factors.

The closest setting to ours is Condorcet winner identification without requiring a latent total order among all arms. Haddenhorst *et al.* [2021] study statistical procedures for testing and identifying Condorcet winners. Maiti *et al.* [2024] study pure exploration in matrix games, a framework that subsumes stochastic bandits and dueling bandits, give near-optimal guarantees for identifying pure-strategy Nash equilibria, and obtain a high-probability Condorcet-identification bound governed by winner-to-arm gaps. Most recently, Saad *et al.* [2026] introduce an identification procedure that exploits the full gap matrix $\Delta_{i,j} = p_{i,j} - 1/2$ rather than only the winner-to-arm gaps $\{\Delta_{a^\star,j}\}$, leveraging informative comparisons among non-winners to improve, up to logarithmic factors, the best known instance-dependent sample-complexity rates. They further establish the first lower bounds for Condorcet identification under the sole Condorcet-winner assumption, matching their upper bounds up to logarithmic factors and isolating the intrinsic cost of locating and estimating informative entries of the gap matrix; this reveals new regimes in which the optimal complexity is governed by the full preference matrix rather than by $\Delta^\star$ alone.

Our BAI guarantee uses a global-gap accounting with $\Delta := \min_{i \neq j} |p_{i,j} - 1/2|$ and obtain a δ-correct identification primitive with expected sample complexity

$$O\left(\frac{N}{\Delta^2}\left(\log\frac{1}{\delta} + \log\frac{1}{\Delta} + 1\right)\right).$$

Compared with early BTM-PAC or PE-style bounds containing $\log(N/(\epsilon\delta))$ or $\log(N/(\Delta\delta))$, the confidence dependence in our bound is $\log(1/\delta)$ rather than $\log(N/\delta)$; the remaining $\log(1/\Delta)$ term comes from the sequential pairwise tests used inside the tournament primitive. Compared with SST/STI-based maxing algorithms, our procedure does not rely on a total order or magnitude-transitivity assumptions among all arms. Compared with the instance-dependent CW-identification guarantees of Saad *et al.* [2026], our bound is coarser in the gap structure, but the resulting primitive is simple, starting-arm-uniform, and directly reusable in the weak- and strong-regret components of our framework.

A.3 Targeting Multi-Objective Unification

The relation between regret minimization and best-arm identification is well studied in ordinary bandits. Fixed-confidence identification requires enough exploration to rule out alternatives, while regret minimization rewards early exploitation. This difference is reflected in classical BAI algorithms and lower bounds [Audibert *et al.*, 2010; Kaufmann *et al.*, 2016]. Explore-then-commit strategies make the tension explicit and can be suboptimal for regret minimization [Garivier *et al.*, 2016]. Several works study the interface more directly, including algorithms for A/B testing and Pareto-frontier analyses of regret minimization versus best-arm identification [Degenne *et al.*, 2019; Zhong *et al.*, 2023].

In dueling bandits, the learner chooses two arms per round, which creates additional ways to combine exploration and exploitation. Winner-Stays-style algorithms use one coordinate as an incumbent and the other to test challengers, which is especially natural for weak regret [Chen and Frazier, 2017]. The work closest to our multi-objective motivation is Ma *et al.* [2025], which studies joint regret minimization and sample complexity for dueling bandits and proposes a black-box reduction transforming any regret-minimization algorithm into a best-arm identification algorithm. However, their proposed algorithm only achieves the WS-W style weak regret bound of order $O(N \log N)$, a $\log N$ factor above the $O(N)$ optimum of Saad *et al.* [2024]. And their algorithm is a further $\log N$ factor suboptimal for BAI.

B Primitive Tests and Tournament Subroutines

This appendix collects the statistical primitives used throughout the paper and gives their complete proofs. We keep the statements in a global-gap form because this is the accounting used by the main results. Recall that

$$\Delta_{i,j} := \left|p_{i,j} - \frac{1}{2}\right|, \qquad \Delta := \min_{i \neq j} \Delta_{i,j}.$$

All logarithms are natural. Universal numerical constants may change from line to line.

All guarantees below hold conditionally on the history before a fresh call, which fixes the arms, candidate set, tree, and confidence argument. Given this history, the new observations of each selected pair are i.i.d. Bernoulli with the prescribed mean. Applying the pairwise bounds test by test and averaging over their entry histories therefore extends them to the whole subroutine. This allows the online policies to choose their inputs adaptively.

B.1 DUELTEST

Lemma 2 (Anytime confidence sequence) *Let $Y_1, Y_2, \ldots$ be i.i.d. Bernoulli random variables with mean p, and let $\widehat{\mu}_n = n^{-1}\sum_{t=1}^{n} Y_t$. Then for every $\delta \in (0,1)$,*

$$\Pr\left(\exists n \geq 1 : |\widehat{\mu}_n - p| \geq c(n,\delta)\right) \leq \delta.$$

Proof. For any fixed $n \geq 1$, Hoeffding's inequality gives

$$\Pr\{|\widehat{\mu}_n - p| \geq c(n,\delta)\} \leq 2\exp\{-2nc(n,\delta)^2\} = \frac{\delta}{n(n+1)}.$$

Taking a union bound over all $n \geq 1$ and using $\sum_{n\geq 1} 1/(n(n+1)) = 1$ proves the claim. □

Lemma 3 (Correctness of DUELTEST) *For every pair $u \neq v$ with $p_{u,v} \neq 1/2$, Algorithm 1 terminates almost surely and returns the better arm with probability at least $1-\delta$.*

Proof. Almost-sure termination follows from the strong law of large numbers: $\widehat{\mu}_n \to p_{u,v}$ while $c(n,\delta) \to 0$, so the stopping inequality eventually holds when $p_{u,v} \neq 1/2$. We prove the case $p_{u,v} = 1/2 + \gamma$ with $\gamma > 0$. The other case is symmetric. Let

$$\mathcal{E}_\delta := \{\forall n \geq 1 : |\widehat{\mu}_n - p_{u,v}| < c(n,\delta)\}.$$

By Lemma 2, $\Pr(\mathcal{E}_\delta) \geq 1-\delta$. On $\mathcal{E}_\delta$, the test cannot stop and return v. Indeed, if it returned v at time τ, then $\widehat{\mu}_\tau < 1/2 - c(\tau,\delta)$ by the stopping rule and the decision rule. Hence

$$p_{u,v} - \widehat{\mu}_\tau > \left(\frac{1}{2} + \gamma\right) - \left(\frac{1}{2} - c(\tau,\delta)\right) = \gamma + c(\tau,\delta) > c(\tau,\delta),$$

contradicting $\mathcal{E}_\delta$. Thus the error event is contained in $\mathcal{E}_\delta^c$, whose probability is at most δ. □

The expected number of comparisons made by DUELTEST has the standard inverse-gap form. We record a version with a concrete threshold.

Lemma 4 (Stopping-time bound for one pair) *There is a universal constant $C_{\mathrm{dt}} > 0$ such that, for every pair (u,v) with gap $\gamma = |p_{u,v} - 1/2| \in (0, 1/2]$ and every $\delta \in (0,1)$, the stopping time $\tau(u,v,\delta)$ of Algorithm 1 satisfies*

$$\mathbb{E}[\tau(u,v,\delta)] \leq \frac{C_{\mathrm{dt}}}{\gamma^2}\left(\log\frac{1}{\delta} + \log\frac{1}{\gamma} + 1\right). \tag{1}$$

In particular, since $\gamma \geq \Delta$ for every distinct pair and $\Delta > 0$ under Assumption 1,

$$\mathbb{E}[\tau(u,v,\delta)] \leq \frac{C_{\mathrm{dt}}}{\Delta^2}\left(\log\frac{1}{\delta} + \log\frac{1}{\Delta} + 1\right).$$

Proof. It is enough to consider $p_{u,v} = 1/2 + \gamma$. The other case is symmetric. Define

$$n_\star(\gamma,\delta) := \left\lceil \frac{C_0}{\gamma^2}\left(\log\frac{1}{\delta} + \log\frac{1}{\gamma} + 1\right)\right\rceil,$$

where $C_0 = 32$. Put

$$A := \log\frac{1}{\delta} + \log\frac{1}{\gamma} + 1 \geq 1.$$

The ceiling and $\gamma \leq 1/2$ imply $n_\star \leq 64A/\gamma^2$. Using $\log A \leq A - 1$, we obtain

$$\begin{aligned}\log\frac{4}{\delta} + 2\log n_\star &\leq \log 4 + 2\log 64 + \log\frac{1}{\delta} + 4\log\frac{1}{\gamma} + 2\log A \\ &\leq 6A + (\log 4 + 2\log 64 - 6) \leq 16A \leq \frac{n_\star\gamma^2}{2}.\end{aligned}$$

The function $n \mapsto n\gamma^2/2 - 2\log n - \log(4/\delta)$ is nondecreasing for $n \geq 4/\gamma^2$, and $n_\star$ exceeds this threshold. Therefore, for every $n \geq n_\star$, the bound $n+1 \leq 2n$ gives

$$\log\big(2n(n+1)/\delta\big) \leq \log\frac{4}{\delta} + 2\log n \leq \frac{n\gamma^2}{2}.$$

It follows that $c(n,\delta) \leq \gamma/2$ throughout this range.

If the test has not stopped by time $n \geq n_\star(\gamma,\delta)$, then $|\widehat{\mu}_n - 1/2| \leq c(n,\delta) \leq \gamma/2$. Therefore

$$\{\tau > n\} \subseteq \left\{\widehat{\mu}_n \leq \frac{1}{2} + \frac{\gamma}{2}\right\} = \left\{\widehat{\mu}_n - p_{u,v} \leq -\frac{\gamma}{2}\right\}.$$

A one-sided Hoeffding bound yields

$$\Pr(\tau > n) \leq \exp\{-n\gamma^2/2\}, \qquad n \geq n_\star(\gamma,\delta).$$

Using the tail-sum formula,

$$\mathbb{E}[\tau] = \sum_{n \geq 0} \Pr(\tau > n) \leq n_\star(\gamma, \delta) + \sum_{n \geq n_\star(\gamma,\delta)} e^{-n\gamma^2/2}.$$

Since $1 - e^{-x} \geq x/2$ for $x \in (0, 1]$ and $\gamma \leq 1/2$, the geometric tail is at most $4/\gamma^2$. Absorbing constants into C_{dt} gives (1). □

The test needs a positive gap to terminate. If $p_{u,v} = 1/2$, the confidence event of Lemma 2 keeps $|\widehat{\mu}_n - 1/2| < c(n, \delta)$ for every n, so the test runs forever with probability at least $1 - \delta$. BUDGETEDDUELTEST (Section 5) stops within its budget even on such ties.

B.2 BKT

Lemma 5 (BKT on an arbitrary candidate set) *Under Assumption 1, let $S \subseteq [N]$ be nonempty and let $m = |S|$. For every $\delta \in (0, 1)$, Algorithm 2 satisfies the following two properties. First, if $m \geq 2$, its expected number of comparisons satisfies*

$$\mathbb{E}[\tau_{\mathrm{BKT}}(S, \delta)] \leq \frac{C_{\mathrm{bkt}} m}{\Delta^2} \left(\log \frac{1}{\delta} + \log \frac{1}{\Delta} + 1 \right) \tag{2}$$

for a universal constant $C_{\mathrm{bkt}} > 0$. Second, suppose that S contains an arm w satisfying

$$p_{w,j} > \frac{1}{2} \qquad \textit{for every } j \in S \setminus \{w\}.$$

Then Algorithm 2 returns w with probability at least $1 - \delta$. In particular, if $a^\star \in S$, then $\mathrm{BKT}(S, \delta)$ returns $a^\star$ with probability at least $1 - \delta$.

Proof. If $m = 1$, the algorithm returns the unique arm in S, so the correctness claim is immediate whenever the displayed condition holds. For $m \geq 2$, assume first that such an arm w exists. Track only w through the bracket. In each tournament round r, this arm participates in at most one call to DUELTEST, with confidence $\delta_r = 6\delta/(\pi^2 r^2)$. If w is ever eliminated, one of these calls must return an arm that is worse than w, and hence must be wrong. By Lemma 3 and a union bound,

$$\Pr\{w \text{ is eliminated}\} \leq \sum_{r \geq 1} \delta_r = \sum_{r \geq 1} \frac{6\delta}{\pi^2 r^2} = \delta.$$

If w is never eliminated, it is the final remaining arm.

We now prove the expected-cost bound, which does not require S to contain any Condorcet arm. Let X_{r-1} denote the active set at the beginning of round r, and let K_r be the number of pairwise tests in that round. The halving bound is deterministic: regardless of the duel outcomes, the bracket keeps at most one survivor from each pair and possibly one bye, so

$$|X_{r-1}| \leq \left\lceil \frac{m}{2^{r-1}} \right\rceil, \qquad K_r \leq \left\lfloor \frac{|X_{r-1}|}{2} \right\rfloor.$$

The number of nonempty rounds is at most $R := \lceil \log_2 m \rceil$. By Lemma 4, each test in round r has expected length at most

$$\frac{C}{\Delta^2} \left(\log \frac{1}{\delta} + 2 \log r + \log \frac{1}{\Delta} + 1 \right).$$

Therefore

$$\mathbb{E}[\tau_{\mathrm{BKT}}(S, \delta)] \leq \frac{C}{\Delta^2} \sum_{r=1}^{R} K_r \left(\log \frac{1}{\delta} + 2 \log r + \log \frac{1}{\Delta} + 1 \right).$$

The total number of matches is at most $m - 1$, so the terms not involving $\log r$ contribute at most the right-hand side of (2). For the remaining term, the deterministic halving bound gives

$$\sum_{r=1}^{R} K_r \log r \leq \sum_{r=1}^{R} \left(\frac{m}{2^r} + 1 \right) \log r \leq Cm + CR \log R.$$

Since $R \leq \lceil \log_2 m \rceil$ and $R \log R \leq Cm$ for all $m \geq 2$, the logarithmic-round contribution is also $O(m)$. This proves Equation (2) after increasing the constant. □

When S contains no Condorcet arm, BKT should be interpreted only as returning a bracket representative. The expected-cost part of Lemma 5 still applies to such sets.

C Tree-Guided Identification

This appendix gives the complete proof of the tree-guided identification bound, including the formal-bye treatment for arbitrary $N \geq 2$. The guarantees for TREEASCENT use Assumption 1 and, as in Appendix B, hold conditionally on the history before the pass.

C.1 Tree-Based Tournament Decomposition

If N is not a power of two, embed the N real arms into a completed balanced bracket with

$$M := 2^{\lceil \log_2 N \rceil} < 2N$$

leaves. The additional leaves are formal byes: they are not arms, are not assigned preference probabilities, and never generate stochastic comparisons. If N is a power of two, $M = N$. Let $L := \log_2 M$ and let $\mathcal{T}$ be the completed balanced tree. For a node v, write $\bar{S}(v)$ for all bracket leaves under v and $S(v) := \bar{S}(v) \cap [N]$ for the real arms under v. We do not define a local maximum for arbitrary nonempty $S(v)$, because such a maximum need not exist when nonwinner arms form cycles.

For a starting real arm b, let

$$v_0(b) \subset v_1(b) \subset \cdots \subset v_L(b) = \text{root}$$

be its ancestor chain in the completed tree, with $\bar{S}(v_0(b)) = \{b\}$ and $|\bar{S}(v_h(b))| = 2^h$. Write $U_h(b) := S(v_h(b))$. Let $q_h(b)$ be the sibling of $v_{h-1}(b)$ inside $v_h(b)$ and define

$$Q_h(b) := S(q_h(b)) = U_h(b) \setminus U_{h-1}(b), \qquad \bar{Q}_h(b) := \bar{S}(q_h(b)).$$

The set $Q_h(b)$ may be empty because formal byes are ignored.

Lemma 6 (Sibling-block partition with formal byes) *For every real arm $b \in [N]$,*

$$[N] = \{b\} \sqcup Q_1(b) \sqcup Q_2(b) \sqcup \cdots \sqcup Q_L(b),$$

where the union is disjoint. If $m_h := |Q_h(b)|$ and $\bar{m}_h := |\bar{Q}_h(b)|$, then

$$\sum_{h=1}^{L} m_h = N - 1, \qquad \bar{m}_h = 2^{h-1}, \qquad \sum_{h=1}^{L} \bar{m}_h = M - 1.$$

In particular, when $N = 2^L$, $M = N$ and $m_h = \bar{m}_h = 2^{h-1}$.

Proof. For every $h \geq 1$, the node $v_h(b)$ has two children: $v_{h-1}(b)$ and $q_h(b)$. Restricting to real arms gives

$$U_h(b) = U_{h-1}(b) \sqcup Q_h(b).$$

Iterating this identity up to $h = L$ gives the real-arm partition because $U_0(b) = \{b\}$ and $U_L(b) = [N]$. This also gives $\sum_h m_h = N - 1$. For capacities, the completed bracket is full and balanced, so $|\bar{Q}_h(b)| = |\bar{S}(v_{h-1}(b))| = 2^{h-1}$ and $\sum_{h=1}^{L} 2^{h-1} = M - 1$. □

Definition 1 (Winner-entry level) *For $b \in [N]$, define*

$$h_\star(b) := \begin{cases} 0, & b = a^\star, \\ \text{the unique } h \in \{1, \ldots, L\} \text{ such that } a^\star \in Q_h(b), & b \neq a^\star. \end{cases}$$

The second case is well-defined by the sibling-block partition in Lemma 6.

C.2 Correctness of TREEASCENT

Lemma 7 (Capture and preservation of the Condorcet winner) *Fix a starting arm b and let $h_\star = h_\star(b)$ be the winner-entry level from Definition 1. Suppose that, if $h_\star > 0$, the call $\mathrm{BKT}(Q_{h_\star}(b), \beta_{h_\star})$ returns $a^\star$. Suppose also that every invoked merge call at levels $h \geq \max\{1, h_\star\}$ returns $a^\star$ whenever $a^\star$ is one of its two inputs. Then Algorithm 7 returns $a^\star$.*

Algorithm 7 TREEASCENT$(b, \mathcal{T}, \varepsilon)$: tree-guided identification

Require: Starting arm $b \in [N]$, completed balanced tree $\mathcal{T}$ with M leaves, confidence $\varepsilon \in (0, 1)$.

Let $L \leftarrow \log_2 M$. Compute the ancestor chain $v_0(b) \subset \cdots \subset v_L(b)$ and the sibling blocks $Q_h(b)$ and $\bar{Q}_h(b)$.
$c_0 \leftarrow b$.
for $h = 1, \ldots, L$ **do**
  $m_h \leftarrow |Q_h(b)|$ and $\bar{m}_h \leftarrow |\bar{Q}_h(b)|$.
  Set $\beta_h \leftarrow (\varepsilon/2)\bar{m}_h/(M-1)$ and $\eta_h \leftarrow \varepsilon/(2L)$.
  if $m_h = 0$ **then**
    $c_h \leftarrow c_{h-1}$.
  else
    $g_h \leftarrow \mathrm{BKT}(Q_h(b), \beta_h)$.
    $c_h \leftarrow$ DUELTEST(c_{h-1}, g_h, η_h).
  end if
end for
return c_L.

Proof. If $h_\star = 0$, then $c_0 = a^\star$. Consider any later level h. If $Q_h(b)$ is empty, the algorithm sets $c_h = c_{h-1} = a^\star$. If $Q_h(b)$ is nonempty, the merge duel compares $a^\star$ with the representative g_h, and the stated merge condition returns $a^\star$. Thus $c_h = a^\star$ for all h.

If $h_\star > 0$, we make no claim about the champion before level $h_\star$. Since $a^\star \in Q_{h_\star}(b)$, this block is nonempty. By the stated BKT condition, $g_{h_\star} = a^\star$. The merge duel at level $h_\star$ then compares $a^\star$ with the previous champion and returns $a^\star$, so $c_{h_\star} = a^\star$. The preservation argument from the previous paragraph applies from level $h_\star + 1$ onward, and hence $c_L = a^\star$. □

Theorem 6 (Correctness of one identification pass) *For every starting arm b and every $\varepsilon \in (0, 1)$,*

$$\Pr\{\text{TREEASCENT}(b, \mathcal{T}, \varepsilon) = a^\star\} \geq 1 - \varepsilon.$$

Proof. Let $h_\star = h_\star(b)$. If $h_\star > 0$, then $a^\star \in Q_{h_\star}(b)$, and $a^\star$ is a Condorcet arm within this block. By Lemma 5,

$$\Pr\{\mathrm{BKT}(Q_{h_\star}(b), \beta_{h_\star}) \neq a^\star\} \leq \beta_{h_\star}.$$

If $h_\star = 0$, no capture tournament is needed.

For each level h, define F_h to be the event that the merge call at level h is invoked with $a^\star$ as one of its two inputs but returns the other arm. Conditional on the history before that fresh merge test, Lemma 3 gives failure probability at most η_h. Hence $\Pr(F_h) \leq \eta_h$. On the complement of the capture failure and of all events F_h, Lemma 7 implies that the final output is $a^\star$. Therefore,

$$\Pr\{\text{TREEASCENT}(b, \mathcal{T}, \varepsilon) \neq a^\star\} \leq \mathbf{1}\{h_\star > 0\}\beta_{h_\star} + \sum_{h=1}^{L} \eta_h \leq \sum_{h=1}^{L} \beta_h + \sum_{h=1}^{L} \eta_h.$$

By construction and Lemma 6,

$$\sum_{h=1}^{L} \beta_h = \frac{\varepsilon}{2} \frac{\sum_h \bar{m}_h}{M-1} = \frac{\varepsilon}{2}, \qquad \sum_{h=1}^{L} \eta_h = \frac{\varepsilon}{2}.$$

This proves the theorem. □

C.3 Cost of TREEASCENT

Theorem 7 (One-pass identification cost) *There is a universal constant $C_{\mathrm{TA}} > 0$ such that for every starting arm b and every $\varepsilon \in (0, 1)$, the number of comparisons $\tau_{\mathrm{TA}}(b, \varepsilon)$ used by Algorithm 7 satisfies*

$$\mathbb{E}[\tau_{\mathrm{TA}}(b, \varepsilon)] \leq \frac{C_{\mathrm{TA}} N}{\Delta^2} \left(\log \frac{1}{\varepsilon} + \log \frac{1}{\Delta} + 1 \right).$$

The same bound upper-bounds the weak regret incurred during the pass.

Proof. Let T_h^{sub} be the elementary-comparison count of the subtree tournament at level h, and let T_h^{merge} be the count of the merge duel. These quantities are zero if $Q_h(b)$ is empty. Then

$$\tau_{\mathrm{TA}} = \sum_{h=1}^{L} T_h^{\mathrm{sub}} + \sum_{h=1}^{L} T_h^{\mathrm{merge}}.$$

For the subtree terms, the expected-cost part of Lemma 5 gives

$$\mathbb{E}[T_h^{\mathrm{sub}}] \le \frac{C m_h}{\Delta^2}\left(\log\frac{1}{\beta_h} + \log\frac{1}{\Delta} + 1\right), \qquad m_h := |Q_h(b)|.$$

Therefore

$$\sum_{h=1}^{L} \mathbb{E}[T_h^{\mathrm{sub}}] \le \frac{C}{\Delta^2}\left(\sum_{h=1}^{L} m_h \log\frac{1}{\beta_h} + (N-1)\left(\log\frac{1}{\Delta} + 1\right)\right).$$

Since $m_h \le \bar{m}_h$, $\sum_h \bar{m}_h = M - 1$, and $\beta_h = (\varepsilon/2)\bar{m}_h/(M-1)$,

$$\sum_{h=1}^{L} m_h \log\frac{1}{\beta_h} \le \sum_{h=1}^{L} \bar{m}_h \log\frac{1}{\beta_h} = (M-1)\log\frac{2}{\varepsilon} + \sum_{h=1}^{L} \bar{m}_h \log\frac{M-1}{\bar{m}_h}.$$

Using $\bar{m}_h = 2^{h-1}$ and $M = 2^L$,

$$\sum_{h=1}^{L} \bar{m}_h \log\frac{M-1}{\bar{m}_h} \le \sum_{h=1}^{L} 2^{h-1}(L-h+1)\log 2 = M\log 2 \sum_{j=1}^{L} \frac{j}{2^j} \le 2M\log 2.$$

Thus the subtree contribution is bounded by

$$\frac{CM}{\Delta^2}\left(\log\frac{1}{\varepsilon} + \log\frac{1}{\Delta} + 1\right).$$

For the merge terms, Lemma 4 gives

$$\sum_{h=1}^{L} \mathbb{E}[T_h^{\mathrm{merge}}] \le \frac{CL}{\Delta^2}\left(\log\frac{2L}{\varepsilon} + \log\frac{1}{\Delta} + 1\right).$$

The last display is bounded by the same M-linear expression after adjusting the universal constant. Indeed, $L \le M$ and $L\log(2L) \le CM$ for a numerical constant C, while $L\log(1/\varepsilon)$ and $L(\log(1/\Delta)+1)$ are dominated by $M\log(1/\varepsilon)$ and $M(\log(1/\Delta)+1)$. Since $M < 2N$, the comparison bound follows. Finally, weak regret is at most the number of comparisons, proving the final claim. □

C.4 A lower bound for best-arm identification

The bound of Corollary 1 on the expected stopping time of TG-ITE-BAI contains the term $N\Delta^{-2}\log(1/\delta)$. The next proposition shows that this term cannot be avoided when every pairwise gap equals Δ, even by a learner that knows the gap.

Proposition 2 (Homogeneous-gap identification lower bound) *Fix $N \ge 2$, $d \in (0, 1/4]$, and $\delta \in (0, 1/10]$. Let $\mathcal{H}_N(d)$ be the class of N-arm instances that have a Condorcet winner and satisfy $|p_{i,j} - 1/2| = d$ for all $i \ne j$. Consider an algorithm, possibly taking $g = d$ as input, that is δ-correct on every instance in $\mathcal{H}_N(d)$. Then every $P \in \mathcal{H}_N(d)$ satisfies*

$$\mathbb{E}_P[\tau] \ge \frac{(N-1)\ \mathrm{kl}(1-\delta, \delta)}{\mathrm{kl}(1/2 - d, 1/2 + d)} \ge c\,\frac{N}{d^2}\log\frac{1}{\delta}$$

for a universal $c > 0$, where kl *is the Bernoulli relative entropy.*

Proof. Let $a^\star$ be the Condorcet winner of P, and let L_i be the set of arms that beat arm i in P. For each $i \ne a^\star$, let $P^{(i)}$ be obtained from P by reversing every pair $\{i, j\}$ with $j \in L_i$, so that the

probability that i wins rises from $1/2 - d$ to $1/2 + d$. Arm i then beats every other arm, all pairwise gaps still equal d, and the pairs not involving i are unchanged. Hence $P^{(i)} \in \mathcal{H}_N(d)$.

Let $N_{ij}(\tau)$ be the number of comparisons of the unordered pair $\{i, j\}$ before stopping. If $\mathbb{E}_P[\tau] = \infty$, there is nothing to prove. Otherwise, consider the event $A = \{\tau < \infty,\ \widehat{a} = a^\star\}$. By δ-correctness on P and $P^{(i)}$, $\Pr_P(A) \geq 1 - \delta$ and $\Pr_{P^{(i)}}(A) \leq \delta$. Apply the change-of-measure inequality of Kaufmann *et al.* [2016, Lemma 1], with the selected pairs as arms, at the bounded stopping time $\tau \wedge t$ to the event $\{\tau \leq t,\ \widehat{a} = a^\star\}$, and let $t \to \infty$. Monotone convergence on the left and lower semicontinuity of kl on the right give

$$\sum_{j \in L_i} \mathbb{E}_P[N_{ij}(\tau)]\ \mathrm{kl}(1/2 - d, 1/2 + d) \geq \mathrm{kl}\big(\Pr_P(A),\ \Pr_{P^{(i)}}(A)\big) \geq \mathrm{kl}(1 - \delta, \delta).$$

Each unordered pair appears only in the inequality for its losing arm, which is never $a^\star$. Summing over $i \neq a^\star$ and using $\sum_{i \neq a^\star} \sum_{j \in L_i} N_{ij}(\tau) \leq \tau$ proves the first bound.

For the second bound, $d \leq 1/4$ and $\delta \leq 1/10$ give

$$\mathrm{kl}(1/2 - d, 1/2 + d) = 2d \log \frac{1 + 2d}{1 - 2d} \leq 16 d^2,$$

and $\mathrm{kl}(1 - \delta, \delta) \geq \frac{1}{2} \log(1/\delta)$. Since also $N - 1 \geq N/2$, we can take $c = 1/64$. □

Every instance in $\mathcal{H}_N(d)$ satisfies Assumption 1 with $\Delta = \Delta^\star = d$, so TG-ITE-BAI meets the hypothesis without knowing d. Its upper bound matches the lower bound up to constants and an additional $Nd^{-2} \log(1/d)$ term, which comes from the anytime confidence radius of DUELTEST (Appendix G.1). Given $g = d$, the budgeted policy of Theorem 5 also meets the hypothesis, and its budget $CNd^{-2}(\log(1/\delta) + 1)$ matches the lower bound up to constants since $\log(1/\delta) > 1$.

D Weak Regret

This appendix proves the strictly linear infinite-horizon weak-regret result. The proofs use fresh samples in every screening and replacement call. Here "fresh" means that the elementary outcomes observed inside the current call have not been observed or reused before. The selected pairs may still depend on the past and on earlier outcomes inside the same call. The stochastic model only requires that, conditional on each selected pair and the current history, the newly observed Bernoulli outcome is independent of previous outcomes with the corresponding mean. This is the reason that the analysis remains valid even though the incumbent entering an epoch was selected adaptively from earlier data.

D.1 SCREENANDREPLACE

Lemma 8 (One-step correction probability) *Under Assumption 1, let $\rho, \kappa \in (0, 1)$, and let $\mathcal{H}$ denote the history immediately before a call to Algorithm 4. This history includes the incumbent c already chosen by the algorithm. Let c^+ be the arm returned by the call. Then, almost surely,*

$$\Pr\big(c^+ \neq a^\star \mid \mathcal{H}\big) \leq \mathbf{1}\{c \neq a^\star\}(\rho + \kappa) + \mathbf{1}\{c = a^\star\}\kappa.$$

Consequently,

$$\Pr\big(c^+ \neq a^\star \mid c \neq a^\star\big) \leq \rho + \kappa, \qquad \Pr\big(c^+ \neq a^\star \mid c = a^\star\big) \leq \kappa$$

whenever the conditioning events have positive probability, and $\Pr(c^+ \neq a^\star) \leq \rho + \kappa$ unconditionally.

Proof. Fix any realized pre-call history, including the realized value of c. After this conditioning, the only randomness left is the fresh feedback generated inside the current call. Therefore Lemma 3 and the correctness part of Lemma 5 can be applied to the tests and tournaments run inside this call, even though the pairs tested inside the call may be chosen adaptively from earlier fresh feedback in the same call.

If $c \neq a^\star$, then the screening phase compares c against $a^\star$ exactly once. Since $a^\star \succ c$, DUELTEST$(c, a^\star, \rho)$ inserts $a^\star$ into B with conditional probability at least $1 - \rho$. Conditional on $a^\star \in B$,

the replacement tournament is run on a set containing $a^\star$, and Lemma 5 implies that it returns $a^\star$ with conditional probability at least $1-\kappa$. A union bound gives conditional failure probability at most $\rho+\kappa$.

If $c = a^\star$, then either $B = \emptyset$ and the algorithm returns $a^\star$ deterministically, or $B \neq \emptyset$ and the replacement tournament is run on a set containing $a^\star$. In the latter case the tournament fails with conditional probability at most κ. This proves the displayed conditional inequality. The remaining claims follow by taking expectations and conditioning on $\{c \neq a^\star\}$ or $\{c = a^\star\}$ when those events have positive probability. □

Lemma 9 (Expected weak regret in one screen-and-replace epoch) *Under Assumption 1, let $\rho, \kappa \in (0,1)$, and let $\mathcal{H}$ denote the history immediately before a call to Algorithm 4. This history includes the incumbent c already chosen by the algorithm. Let $W(c,\rho,\kappa)$ be the weak regret incurred during the call. Then, almost surely,*

$$\mathbb{E}[W(c,\rho,\kappa) \mid \mathcal{H}] \le \frac{CN}{\Delta^2}\left[\mathbf{1}\{c \neq a^\star\}\left(\log\frac{1}{\rho}+\log\frac{1}{\kappa}+\log\frac{1}{\Delta}+1\right) + \rho\left(\log\frac{1}{\kappa}+\log\frac{1}{\Delta}+1\right)\right]. \tag{3}$$

Consequently, if $p = \Pr(c \neq a^\star)$, then

$$\mathbb{E}[W(c,\rho,\kappa)] \le \frac{CN}{\Delta^2}\left[p\left(\log\frac{1}{\rho}+\log\frac{1}{\kappa}+\log\frac{1}{\Delta}+1\right) + \rho\left(\log\frac{1}{\kappa}+\log\frac{1}{\Delta}+1\right)\right]. \tag{4}$$

Proof. Fix any realized pre-call history, including the realized value of c. All comparisons inside the call use fresh feedback, so after conditioning on each selected pair and the current within-call history, the expected stopping-time bounds apply exactly as in the nonadaptive case.

During screening, if $c = a^\star$, every screening comparison is of the form $(a^\star, j)$ and has zero weak regret. If $c \neq a^\star$, the comparison against $a^\star$ itself also has zero weak regret. Nevertheless, we use the simpler upper bound by the total number of screening comparisons. There are at most $N-1$ such tests, each with confidence ρ, so Lemma 4 gives conditional screening contribution at most

$$\mathbf{1}\{c \neq a^\star\}\frac{CN}{\Delta^2}\left(\log\frac{1}{\rho}+\log\frac{1}{\Delta}+1\right).$$

If $c \neq a^\star$ and the replacement tournament is invoked, its weak regret is bounded by its total number of comparisons. Since the candidate set has size at most N, Lemma 5 gives conditional contribution at most

$$\mathbf{1}\{c \neq a^\star\}\frac{CN}{\Delta^2}\left(\log\frac{1}{\kappa}+\log\frac{1}{\Delta}+1\right).$$

On the event $c = a^\star$, a non-best arm j enters B only if DUELTEST$(a^\star, j, \rho)$ returns j, which has conditional probability at most ρ. Therefore

$$\mathbb{E}[|B| \mid \mathcal{H}, c = a^\star] \le (N-1)\rho.$$

Conditional on the realized beater set B, if $B \neq \emptyset$, the replacement tournament is run on $|B|+1$ arms and its weak regret is at most its comparison count T_{repl}. This tournament uses fresh samples after the screening phase, so Lemma 5 applies to the inner conditional expectation. By the tower property,

$$\mathbb{E}[T_{\mathrm{repl}} \mid \mathcal{H}, c = a^\star] = \mathbb{E}[\mathbb{E}[T_{\mathrm{repl}} \mid B, \mathcal{H}, c = a^\star] \mid \mathcal{H}, c = a^\star].$$

The inner expectation is at most

$$\frac{C(|B|+1)\mathbf{1}\{|B|>0\}}{\Delta^2}\left(\log\frac{1}{\kappa}+\log\frac{1}{\Delta}+1\right).$$

Since $(|B|+1)\mathbf{1}\{|B|>0\} \le 2|B|$, taking expectation over B gives a false-positive contribution at most

$$\frac{CN\rho}{\Delta^2}\left(\log\frac{1}{\kappa}+\log\frac{1}{\Delta}+1\right).$$

Adding the three conditional contributions proves (3). Taking expectations proves (4). □

D.2 TG-ITE-WEAK

Theorem 8 (Tail bound for infinite-horizon weak regret) *Under Assumption 1, fix any starting incumbent c_0 chosen before the online tail starts, and run the* SCREENANDREPLACE *epochs with $\rho_s = \kappa_s = 4^{-(s+2)}$. If $R^{\mathrm{weak}}_{\mathrm{tail},\infty}$ denotes the weak regret incurred by these epochs only, then*

$$\mathbb{E}[R^{\mathrm{weak}}_{\mathrm{tail},\infty}] \le \frac{CN}{\Delta^2}\left(\log\frac{1}{\Delta}+1\right)$$

for a universal constant $C>0$. In particular, $R^{\mathrm{weak}}_{\mathrm{tail},\infty} < \infty$ almost surely.

Proof. Let $\mathcal{H}_s$ denote the history after epoch s, and let W_s be the weak regret incurred during epoch s. Define

$$p_s := \Pr(c_{s-1} \neq a^\star).$$

The starting incumbent can be arbitrary, so $p_1 \le 1$. By Lemma 8, conditionally on the history before epoch s,

$$\Pr(c_s \neq a^\star \mid \mathcal{H}_{s-1}) \le \mathbf{1}\{c_{s-1}\neq a^\star\}(\rho_s+\kappa_s) + \mathbf{1}\{c_{s-1}=a^\star\}\kappa_s \le \rho_s+\kappa_s.$$

Taking expectations gives

$$p_{s+1} = \Pr(c_s \neq a^\star) \le \rho_s+\kappa_s = 2\cdot 4^{-(s+2)}.$$

Equivalently, for every $s \ge 2$,

$$p_s \le 2\cdot 4^{-(s+1)},$$

while $p_1 \le 1$. Thus p_s is geometrically summable from $s=2$ onward.

Applying Lemma 9 conditionally on the history before epoch s gives

$$\mathbb{E}[W_s \mid \mathcal{H}_{s-1}] \le \frac{CN}{\Delta^2}\Bigg[\mathbf{1}\{c_{s-1}\neq a^\star\}\left(\log\frac{1}{\rho_s}+\log\frac{1}{\kappa_s}+\log\frac{1}{\Delta}+1\right) + \rho_s\left(\log\frac{1}{\kappa_s}+\log\frac{1}{\Delta}+1\right)\Bigg].$$

Taking expectations yields the same bound with $\mathbf{1}\{c_{s-1}\neq a^\star\}$ replaced by p_s. Because $\rho_s=\kappa_s=4^{-(s+2)}$, we have $\log(1/\rho_s)=\log(1/\kappa_s)=(s+2)\log 4$. Using $p_1\le 1$, $p_s \le 2\cdot 4^{-(s+1)}$ for $s\ge 2$, and $\rho_s = 4^{-(s+2)}$,

$$\sum_{s\ge 1}(p_s+\rho_s)(s+1) < \infty, \qquad \sum_{s\ge 1}(p_s+\rho_s) < \infty,$$

because both are bounded by universal multiples of $1+\sum_{s\ge1} s4^{-s}$. The first series controls the factors $\log(1/\rho_s)$ and $\log(1/\kappa_s)$, while the second controls the common factor $\log(1/\Delta)+1$. Summing over s yields

$$\sum_{s\ge1}\mathbb{E}[W_s] \le \frac{CN}{\Delta^2}\left(\log\frac{1}{\Delta}+1\right).$$

Each epoch contains finitely many fresh DUELTEST and BKT calls, and each call terminates almost surely under $\Delta>0$. Hence all epochs terminate almost surely and the infinite comparison stream is well-defined. Since the epochs are consecutive and disjoint in time,

$$R^{\mathrm{weak}}_{\mathrm{tail},\infty} = \sum_{s\ge1} W_s,$$

and monotone convergence gives $\mathbb{E}[R^{\mathrm{weak}}_{\mathrm{tail},\infty}] = \sum_s \mathbb{E}[W_s]$, proving the expected-regret bound.

Finally, a nonnegative random variable with finite expectation is finite almost surely. Hence $R^{\mathrm{weak}}_{\mathrm{tail},\infty} < \infty$ almost surely. □

Corollary 3 (Constant-confidence warm start) *Under Assumption 1, Algorithm 5 with the constant-confidence warm start in line 1 satisfies*

$$\mathbb{E}[R_\infty^{\mathrm{weak}}] \le \frac{CN}{\Delta^2}\left(\log\frac{1}{\Delta}+1\right)$$

for a universal constant $C > 0$. In particular, $R_\infty^{\mathrm{weak}} < \infty$ almost surely.

Proof. Let W_0 be the weak regret incurred by the warm start. Since weak regret is at most the number of comparisons, Theorem 7 with confidence $1/4$ gives

$$\mathbb{E}[W_0] \le \frac{CN}{\Delta^2}\left(\log\frac{1}{\Delta}+1\right).$$

By Theorem 8, the tail contributes at most the same order uniformly over the starting incumbent. Since

$$R_\infty^{\mathrm{weak}} = W_0 + R_{\mathrm{tail},\infty}^{\mathrm{weak}},$$

the claim follows after absorbing constants. The almost-sure statement follows because both nonnegative terms on the right-hand side are finite almost surely. □

Corollary 4 (High-confidence warm start) *Under Assumption 1, if line 1 of Algorithm 5 is replaced by $c_0 \leftarrow \textsc{TreeAscent}(b_0, \mathcal{T}, \delta)$ and c_0 is reported as the BAI answer when this pass finishes, at time τ_{Dual}, then $\Pr(c_0 = a^\star) \ge 1-\delta$ and*

$$\max\left\{\mathbb{E}[\tau_{\mathrm{Dual}}],\ \mathbb{E}[R_\infty^{\mathrm{weak}}]\right\} \le \frac{CN}{\Delta^2}\left(\log\frac{1}{\delta}+\log\frac{1}{\Delta}+1\right).$$

Proof. The BAI correctness and warm-start comparison bound follow from Theorems 6 and 7. The weak regret of the warm start is bounded by its number of comparisons, and Theorem 8 adds a tail term of order $CN\Delta^{-2}(\log(1/\Delta)+1)$, which is absorbed into the displayed bound. The warm start is exactly the TREEASCENT call of TG-ITE-BAI, and screening makes no decision before it ends. With the same inputs and feedback, both policies thus report the same arm at the same time. □

D.3 From gap-weighted to binary regret

In each round, checking whether zero, one, or both selected arms equal $a^\star$ shows that the gap-weighted losses of Remark 1 satisfy

$$\Delta^\star r_t^{\mathrm{weak}} \le r_t^{\mathrm{weak,gw}} \le \tfrac{1}{2} r_t^{\mathrm{weak}}, \qquad \tfrac{\Delta^\star}{2} r_t^{\mathrm{str}} \le r_t^{\mathrm{str,gw}} \le \tfrac{1}{2} r_t^{\mathrm{str}}. \tag{5}$$

Summing over rounds and taking expectations preserves these relations. By the left inequality for weak regret, the gap-weighted bound of Saad *et al.* [2024, Theorem 4.2] implies that WR-TINF attains binary weak regret $O(N/(\Delta^\star)^2)$ on every Condorcet instance. The next corollary shows that this order is unavoidable in the worst case.

Corollary 5 (Binary weak-regret lower bound) *Let $N \ge 6$, $d \in (0, 1/4)$, and let $\mathcal{D}(d)$ be the class of N-arm instances with a Condorcet winner $a^\star$ satisfying $p_{a^\star i} - 1/2 = d$ for every $i \ne a^\star$ and $|p_{ij} - 1/2| \le d$ for nonwinners i, j. There are universal constants $c, c' > 0$ such that, for every policy and $T \ge c'N/d^2$,*

$$\sup_{P \in \mathcal{D}(d)} \mathbb{E}_P[R_T^{\mathrm{weak}}] \ge cN/d^2.$$

Proof. All winner gaps in $\mathcal{D}(d)$ equal d, so in every round the gap-weighted weak loss is exactly d times the binary one, including self-comparisons. Saad *et al.* [2024, Theorem 5.1] bound the supremum of the expected gap-weighted weak regret over $\mathcal{D}(d)$ below by cN/d in this horizon regime. Dividing by d proves the claim. □

Algorithm 8 TG-ITE-STRONG-KNOWNHORIZON$(b_0, \mathcal{T}, H)$

Require: Starting arm b_0; fixed balanced tree $\mathcal{T}$; horizon $H \geq 1$.
1: Set $\epsilon_H \leftarrow 1/(H+1)$.
2: Run TREEASCENT$(b_0, \mathcal{T}, \epsilon_H)$ until the pass terminates or the horizon H expires.
3: **if** the pass terminates by time H with output $\widehat{a}$ **then**
4: For every remaining round up to time H, play the self-comparison $(\widehat{a}, \widehat{a})$.
5: **else**
6: Halt at time H; all rounds up to H were spent inside the identification pass.
7: **end if**

E Strong Regret

This appendix proves the strong-regret guarantees. We use the self-play convention from the problem setting: a self-comparison (i, i) produces dummy feedback and has zero strong regret only when $i = a^\star$. Every non-dummy comparison made by TREEASCENT is between two distinct arms, and therefore contributes at most one unit of binary strong regret.

E.1 TG-ITE-STRONG-KNOWNHORIZON

Theorem 9 (Known-horizon strong regret) *Under Assumption 1, for every horizon $H \geq 1$, Algorithm 8 satisfies*

$$\mathbb{E}[R_H^{\mathrm{str}}] \leq \frac{CN}{\Delta^2}\left(\log(H+1) + \log\frac{1}{\Delta} + 1\right) + 1.$$

Proof. Let τ_H and $\widehat{a}$ be the stopping time and eventual output of the full TREEASCENT pass if it were allowed to run until termination. This is a counterfactual definition relative to the finite horizon: when $\tau_H > H$, the actual algorithm never computes or uses $\widehat{a}$ before time H. The pair $(\tau_H, \widehat{a})$ is nevertheless well-defined because the pass terminates almost surely. During the identification pass, every realized comparison is a distinct-arm comparison and hence contributes at most one unit of binary strong regret. In the event $\tau_H > H$, the algorithm spends all H rounds inside the pass and then halts, so $\widehat{a}$ has no effect on the realized regret. If $\tau_H \leq H$, the remaining self-play rounds contribute strong regret only on the event $\{\widehat{a} \neq a^\star\}$. Pathwise,

$$R_H^{\mathrm{str}} \leq \min\{\tau_H, H\} + (H - \tau_H)_+ \mathbf{1}\{\widehat{a} \neq a^\star\} \leq \tau_H + H\mathbf{1}\{\widehat{a} \neq a^\star\}.$$

Taking expectations and applying Theorems 6 and 7 with $\varepsilon_H = 1/(H+1)$ gives

$$\mathbb{E}[R_H^{\mathrm{str}}] \leq \frac{CN}{\Delta^2}\left(\log\frac{1}{\varepsilon_H} + \log\frac{1}{\Delta} + 1\right) + H\varepsilon_H.$$

Since $\log(1/\varepsilon_H) = \log(H+1)$ and $H\varepsilon_H \leq 1$, the theorem follows. □

E.2 TG-ITE-STRONG-ANYTIME

Theorem 10 (Horizonless strong regret) *Under Assumption 1, for every horizon $T \geq 2$, Algorithm 6 satisfies*

$$\mathbb{E}[R_T^{\mathrm{str}}] \leq \frac{CN}{\Delta^2}\left[\log(T+1) + (\log\log(T+2)+1)\left(\log\frac{1}{\Delta} + 1\right)\right] + C'(\log\log(T+2)+1)$$

for universal constants $C, C' > 0$. For every fixed $\Delta > 0$, the leading term as $T \to \infty$ is $O(N\Delta^{-2}\log T)$. The additional $O(N\Delta^{-2}\log\log T\ \log(1/\Delta))$ term is explicitly retained and may dominate when the instance gap is very small.

The $C'(\log\log(T+2)+1)$ term below comes from the S_T many self-play error costs $B_s\varepsilon_s = 1$, and is independent of N and Δ.

Proof. We first justify the phase variables used below. Conditional on any history before a phase starts, the next TREEASCENT call is a finite composition of fresh pairwise tests and tournaments, and

therefore terminates almost surely under Assumption 1. By induction over phases, for every fixed s, phase s starts and terminates almost surely. Hence the full identification cost τ_s and the output $\widehat{a}_s$ of phase s are well-defined. Here τ_s counts only the distinct-arm comparison rounds used by that identification pass. For a phase that has started but not finished by time T, $\widehat{a}_s$ is counterfactual only relative to the truncated horizon, and using it below only produces an upper bound.

Let K_T be the number of phases that have started by time T. If phase s starts at time t_s, then

$$t_s = 1 + \sum_{r=1}^{s-1} (\tau_r + B_r) \geq \sum_{r=1}^{s-1} B_r.$$

Thus any phase s that has started by time T satisfies $\sum_{r=1}^{s-1} B_r \leq T$. For $s \geq 2$, the left side is at least $B_{s-1} = 2^{2^{s-1}}$. Hence any such started phase satisfies $2^{2^{s-1}} \leq T$, and, after harmless endpoint slack handled by $T + 2$,

$$s \leq 2 + \lceil \log_2 \log_2 (T + 2) \rceil =: S_T.$$

Therefore $K_T \leq S_T$ deterministically. Identification passes can only delay later phases and cannot increase this bound.

Strong regret before time T is upper-bounded pathwise by the full identification costs and full self-play-block losses of all phases that have started:

$$R_T^{\mathrm{str}} \leq \sum_{s=1}^{K_T} \tau_s + \sum_{s=1}^{K_T} B_s \mathbf{1}\{\widehat{a}_s \neq a^\star\}.$$

This may count parts of a pass or a self-play block beyond time T, but only as a nonnegative overcount. Since $\tau_s \geq 0$ and $B_s \mathbf{1}\{\widehat{a}_s \neq a^\star\} \geq 0$, the deterministic bound $K_T \leq S_T$ gives

$$R_T^{\mathrm{str}} \leq \sum_{s=1}^{S_T} \tau_s + \sum_{s=1}^{S_T} B_s \mathbf{1}\{\widehat{a}_s \neq a^\star\}.$$

Because S_T is deterministic, we can exchange the finite summation and expectation after taking expectations.

Let $\mathcal{F}_{s-1}$ denote the history before phase s starts, including the starting arm c_{s-1} chosen by the algorithm. Conditional on $\mathcal{F}_{s-1}$, this starting arm is fixed, and the s-th identification pass uses fresh samples. Since the TREEASCENT guarantees are uniform over the starting arm, they apply conditionally:

$$\Pr(\widehat{a}_s \neq a^\star \mid \mathcal{F}_{s-1}) \leq \varepsilon_s, \qquad \mathbb{E}[\tau_s \mid \mathcal{F}_{s-1}] \leq \frac{CN}{\Delta^2} \left(\log \frac{1}{\varepsilon_s} + \log \frac{1}{\Delta} + 1 \right).$$

Taking expectations and summing over $s \leq S_T$ yields

$$\mathbb{E}[R_T^{\mathrm{str}}] \leq \frac{CN}{\Delta^2} \sum_{s=1}^{S_T} \left(\log \frac{1}{\varepsilon_s} + \log \frac{1}{\Delta} + 1 \right) + \sum_{s=1}^{S_T} B_s \varepsilon_s.$$

By construction $B_s \varepsilon_s = 1$ and $\log(1/\varepsilon_s) = 2^s \log 2$, so

$$\mathbb{E}[R_T^{\mathrm{str}}] \leq \frac{CN}{\Delta^2} \left(\sum_{s=1}^{S_T} 2^s + S_T \left(\log \frac{1}{\Delta} + 1 \right) \right) + S_T.$$

Finally, the definition of S_T gives $S_T \leq 3 + \log_2 \log_2 (T + 2)$ and therefore

$$\sum_{s=1}^{S_T} 2^s = 2^{S_T + 1} - 2 \leq 16 \log_2 (T + 2) \leq C \log(T + 1), \qquad S_T \leq C(\log \log(T + 2) + 1).$$

Substituting these estimates proves the theorem. The additive $C'(\log \log(T + 2) + 1)$ term is the absorbed form of $S_T = \sum_{s=1}^{S_T} B_s \varepsilon_s$. The induction argument at the beginning of the proof also shows that the anytime policy is well-defined as an infinite comparison stream. □

E.3 A lower bound for strong regret

Proposition 3 (Homogeneous-gap strong-regret lower bound) *Fix $N \geq 2$ and $d \in (0, 1/2)$, and let P be the homogeneous total-order instance with $p_{i,j} = 1/2 + d$ for $i < j$, so that $\Delta = \Delta^\star = d$. Consider a policy that is strongly consistent over strict total orders, that is, $\mathbb{E}_Q[R_T^{\mathrm{str}}] = o(T^a)$ for every strict-total-order instance Q and every $a > 0$. Then*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_P[R_T^{\mathrm{str}}]}{\log T} \geq \frac{N-1}{\mathrm{kl}(1/2-d, 1/2)} \geq \frac{N-1}{4d^2}.$$

Proof. By (5), strong consistency for binary regret implies strong consistency for gap-weighted regret. Let $N_{ij}(T)$ be the number of comparisons of the pair $\{i, j\}$, $i \neq j$, up to time T. Each arm $i > 1$ is beaten exactly by the arms $1, \ldots, i-1$, and Lemma 1(ii) of Komiyama *et al.* [2015], applied to the class of strict total orders, gives

$$\mathrm{kl}(1/2-d, 1/2) \sum_{j<i} \mathbb{E}_P[N_{ij}(T)] \geq (1 - o(1)) \log T.$$

Summing over $i = 2, \ldots, N$ counts each pair once, and every comparison of two distinct arms incurs one unit of binary strong regret. Hence

$$\mathbb{E}_P[R_T^{\mathrm{str}}] \geq \sum_{i=2}^{N} \sum_{j<i} \mathbb{E}_P[N_{ij}(T)] \geq \frac{(N-1)(1-o(1)) \log T}{\mathrm{kl}(1/2-d, 1/2)}.$$

Dividing by $\log T$ and letting $T \to \infty$ proves the first bound. For the second, $\log x \leq x - 1$ gives $\mathrm{kl}(p, q) \leq (p-q)^2/(q(1-q))$, so $\mathrm{kl}(1/2-d, 1/2) \leq 4d^2$. □

Every strict total order satisfies Assumption 1, so by Theorem 10 the anytime policy is strongly consistent over this class. On P, its bound grows as $O(Nd^{-2} \log T)$ and therefore matches the lower bound up to constants as $T \to \infty$.

F Known-Gap Budgeted Extension

This appendix proves the guarantees stated in Section 5. Throughout this section, $g \leq \Delta^\star$ is the known lower bound, ties among nonwinners are allowed, and all calls use fresh samples. The proofs mirror the gap-oblivious analysis, but the pairwise cost is now a deterministic fixed budget.

Algorithm 9 BUDGETEDDUELTEST(u, v, g, δ)

Require: Arms $u \neq v$; budget parameter $g \in (0, 1/2]$; confidence level $\delta \in (0, 1)$.

Set $M \leftarrow \left\lceil \frac{1}{2g^2} \log \frac{4}{\delta} \right\rceil$ and $S \leftarrow 0$.
for $n = 1, 2, \ldots, M$ **do**
  Observe $Y_n \sim \mathrm{Bernoulli}(p_{u,v})$, where $Y_n = 1$ means that u wins.
  $S \leftarrow S + Y_n$, $\widehat{\mu}_n \leftarrow S/n$, and $b_n(\delta) \leftarrow \sqrt{\frac{\log(8n^2/\delta)}{2n}}$.
  if $|\widehat{\mu}_n - 1/2| > b_n(\delta) - g$ **then**
    if $\widehat{\mu}_n \geq 1/2$ **then**
      return u.
    else
      return v.
    end if
  end if
end for
if $\widehat{\mu}_M \geq 1/2$ **then**
  return u.
else
  return v.
end if

Proof of Lemma 1. The deterministic budget is immediate from the loop length. We prove correctness for the case $p_{u,v} = 1/2 + \gamma$ with $\gamma \geq g$. Let $M = \lceil (2g^2)^{-1} \log(4/\delta) \rceil$ and $b_n = b_n(\delta)$.

Consider first an erroneous early return. Since ties are broken in favor of u, the algorithm can return v only if, for some $n \leq M$, $\widehat{\mu}_n < 1/2$ and $|\widehat{\mu}_n - 1/2| > b_n - g$. If $b_n \geq g$, then

$$p_{u,v} - \widehat{\mu}_n = \gamma + \left(\frac{1}{2} - \widehat{\mu}_n\right) > g + b_n - g = b_n.$$

If $b_n < g$, then $p_{u,v} - \widehat{\mu}_n > g > b_n$. Hence every erroneous early return is contained in

$$\{\exists n \leq M : p_{u,v} - \widehat{\mu}_n > b_n\}.$$

A one-sided Hoeffding bound and a union bound give

$$\Pr\left(\exists n \leq M : p_{u,v} - \widehat{\mu}_n > b_n\right) \leq \sum_{n \geq 1} \exp\bigl(-2nb_n^2\bigr) = \sum_{n \geq 1} \frac{\delta}{8n^2} \leq \frac{\delta}{4}.$$

If no early return occurs and the final majority vote is wrong, then $\widehat{\mu}_M < 1/2$, and therefore $p_{u,v} - \widehat{\mu}_M > \gamma \geq g$. By Hoeffding's inequality,

$$\Pr(\widehat{\mu}_M < 1/2) \leq \exp\bigl(-2Mg^2\bigr) \leq \frac{\delta}{4}.$$

Thus the error probability is at most $\delta/2$, and in particular at most δ. The proof for $p_{u,v} = 1/2 - \gamma$ is identical with the inequalities reversed, except that an empirical tie now returns the wrong arm. Hoeffding's inequality bounds these non-strict events in the same way. □

For the remaining proofs, let BKT_g, $\textsc{TreeAscent}_g$, and $\textsc{ScreenAndReplace}_g$ denote the natural variants obtained by using $\textsc{BudgetedDuelTest}$ instead of $\textsc{DuelTest}$ everywhere.

Lemma 10 (Budgeted knockout tournament) *Let $S \subseteq [N]$ be nonempty and $m = |S|$. For every $\delta \in (0,1)$, $\mathrm{BKT}_g(S, \delta)$ uses at most*

$$\frac{Cm}{g^2}\left(\log\frac{1}{\delta} + 1\right)$$

comparisons deterministically for a universal constant $C > 0$. Moreover, if S contains an arm w such that $p_{w,j} \geq 1/2 + g$ for every $j \in S \setminus \{w\}$, then $\mathrm{BKT}_g(S, \delta)$ returns w with probability at least $1 - \delta$.

Proof. The correctness proof only tracks w. In tournament round r, w participates in at most one pairwise call with confidence $\delta_r = 6\delta/(\pi^2 r^2)$. Whenever w is one of the two inputs, the pairwise gap is at least g, so Lemma 1 bounds the conditional probability of eliminating w by δ_r. A union bound gives $\sum_{r \geq 1} \delta_r = \delta$.

For the deterministic budget, let K_r be the number of matches in round r. The halving structure gives $\sum_r K_r \leq m - 1$ and $\sum_r K_r \log r \leq Cm$, exactly as in the proof of Lemma 5. Each round-r call uses at most

$$\left\lceil \frac{1}{2g^2} \log \frac{4}{\delta_r} \right\rceil \leq \frac{C}{g^2}\left(\log\frac{1}{\delta} + \log r + 1\right)$$

comparisons. Summing over rounds and using the two halving bounds proves the budget bound. □

Lemma 11 (Budgeted tree ascent) *If $g \leq \Delta^\star$, then for every starting arm b, tree $\mathcal{T}$, and confidence $\varepsilon \in (0,1)$,*

$$\Pr\{\textsc{TreeAscent}_g(b, \mathcal{T}, \varepsilon) = a^\star\} \geq 1 - \varepsilon,$$

and the number of comparisons in the pass is deterministically bounded by

$$\tau_{\mathrm{TA},g}(b, \varepsilon) \leq \frac{CN}{g^2}\left(\log\frac{1}{\varepsilon} + 1\right).$$

The same deterministic bound upper-bounds the weak or strong regret incurred by the pass.

Proof. The capture-and-preservation proof is the same as Theorem 6. If $a^\star$ first enters the ancestor path through the critical block $Q_{h_\star}(b)$, then within this block it beats every other arm by at least g. Lemma 10 therefore returns $a^\star$ from the critical block with failure probability at most $\beta_{h_\star}$. Once $a^\star$ is the current champion, every later merge call involving $a^\star$ has pairwise gap at least g, so Lemma 1 bounds its failure probability by η_h. Thus the same union bound gives $\sum_h \beta_h + \sum_h \eta_h \le \varepsilon$.

For cost, the subtree tournaments satisfy, by Lemma 10,

$$\sum_{h=1}^{L} T_h^{\mathrm{sub}} \le \frac{C}{g^2}\left(\sum_{h=1}^{L} m_h \log\frac{1}{\beta_h} + N\right) \le \frac{CN}{g^2}\left(\log\frac{1}{\varepsilon} + 1\right),$$

where the final inequality is the same size-weighted entropy calculation as in Theorem 7. The merge calls contribute at most

$$\frac{CL}{g^2}\left(\log\frac{2L}{\varepsilon} + 1\right) \le \frac{CN}{g^2}\left(\log\frac{1}{\varepsilon} + 1\right),$$

because $L \le M < 2N$ and $L\log(2L) \le CN$. Adding the two displays proves the deterministic comparison budget. Each comparison costs at most one unit of weak or strong regret, so the same bound applies to regret incurred during identification. □

Lemma 12 (Budgeted screen-and-replace epoch) *Let $\mathcal{F}$ be the history before a call to* ScreenAndReplace$_g(c,\rho,\kappa)$*, and suppose that c is $\mathcal{F}$-measurable. Let c^+ be the returned arm and let $W(c,\rho,\kappa)$ be the weak regret incurred during the call. If $g \le \Delta^\star$, then, almost surely,*

$$\Pr\big(c^+ \ne a^\star \mid \mathcal{F}\big) \le \mathbf{1}\{c \ne a^\star\}(\rho+\kappa) + \mathbf{1}\{c = a^\star\}\kappa,$$

and

$$\mathbb{E}[W(c,\rho,\kappa) \mid \mathcal{F}] \le \frac{CN}{g^2}\left[\mathbf{1}\{c \ne a^\star\}\left(\log\frac{1}{\rho} + \log\frac{1}{\kappa} + 1\right) + \rho\left(\log\frac{1}{\kappa} + 1\right)\right].$$

Proof. The correction probability is identical to Lemma 8. If $c \ne a^\star$, then the screening phase compares c with $a^\star$, and this comparison inserts $a^\star$ into the beater set with failure probability at most ρ. Conditional on $a^\star$ being present, the replacement tournament contains an arm that beats every other candidate by at least g, so Lemma 10 returns $a^\star$ with failure probability at most κ. If $c = a^\star$, only the replacement tournament can lose $a^\star$, again with probability at most κ.

For the regret bound, the screening part costs at most $(N-1)\lceil (2g^2)^{-1}\log(4/\rho)\rceil$ comparisons, but contributes weak regret only when $c \ne a^\star$. If $c \ne a^\star$, the replacement tournament has size at most N and costs at most $CNg^{-2}(\log(1/\kappa)+1)$ by Lemma 10. If $c = a^\star$, a nonwinner enters the beater set only after a false positive against $a^\star$, which has conditional probability at most ρ. Hence $\mathbb{E}[|B| \mid \mathcal{F}, c = a^\star] \le (N-1)\rho$. Applying the deterministic tournament budget conditionally on B, and using $(|B|+1)\mathbf{1}\{|B|>0\} \le 2|B|$, yields the false-positive term $CN\rho g^{-2}(\log(1/\kappa)+1)$. Summing the contributions proves the display. □

Proof of Theorem 5. The BAI statement is Lemma 11 with $\varepsilon = \delta$. For the constant-confidence weak-regret algorithm, the warm start costs at most CN/g^2 regret by Lemma 11. For the tail, use Lemma 12 with $\rho_s = \kappa_s = 4^{-(s+2)}$. The same recursion as in Theorem 8 gives $p_1 \le 1$ and $p_s \le 2\cdot 4^{-(s+1)}$ for $s \ge 2$. Since $\log(1/\rho_s) = \log(1/\kappa_s) = (s+2)\log 4$, taking expectations in Lemma 12 bounds the weak regret of epoch s by $CNg^{-2}(p_s+\rho_s)(s+1)$. The series $\sum_{s\ge1}(p_s+\rho_s)(s+1)$ is finite, so summing the epoch bounds gives $\mathbb{E}[R_\infty^{\mathrm{weak}}] \le CN/g^2$. With a δ-confidence warm start, the warm-start contribution becomes $CNg^{-2}(\log(1/\delta)+1)$, and the tail is absorbed.

For known-horizon strong regret, the pathwise inequality

$$R_H^{\mathrm{str}} \le \tau_H + H\mathbf{1}\{\widehat{a} \ne a^\star\}$$

from Theorem 9 is unchanged. Applying Lemma 11 with $\varepsilon_H = 1/(H+1)$ bounds τ_H by $CNg^{-2}(\log(H+1)+1)$ and $H\Pr(\widehat{a} \ne a^\star)$ by $H\varepsilon_H \le 1$, which gives the stated bound.

For the anytime strong-regret policy, the phase-counting argument of Theorem 10 is unchanged: by time T, at most $S_T = O(\log\log(T+2)+1)$ phases have started, and $\sum_{s\le S_T}\log(1/\varepsilon_s) =$

$O(\log(T+1))$. Lemma 11 replaces each phase cost by $CNg^{-2}(\log(1/\varepsilon_s)+1)$, while the self-play error term remains $\sum_{s\le S_T} B_s\varepsilon_s = S_T$. Therefore

$$\mathbb{E}[R_T^{\mathrm{str}}] \le \frac{CN}{g^2}\left(\log(T+1)+\log\log(T+2)+1\right) + C'(\log\log(T+2)+1),$$

which completes the proof. □

G Nonwinner-Gap Dependence of the Weak-Regret Policy

We prove Proposition 1 in two steps. We first show that DUELTEST at confidence $1/64$, the screening confidence of the first SCREENANDREPLACE call, needs in expectation at least order $\gamma^{-2}\log(1/\gamma)$ comparisons on a pair with gap γ. We then build a four-arm instance on which the constant-confidence warm start returns a nonwinner u with probability independent of γ. The first screen then runs this test between u and a nonwinner v at gap γ, and each comparison costs one unit of weak regret.

G.1 Duration of DUELTEST near a tie

Lemma 13 (Near-tie duration of DUELTEST) *Let $\delta_0 = 1/64$ and let τ_γ be the stopping time of Algorithm 1 with confidence δ_0 on i.i.d. Bernoulli$(1/2+\gamma)$ observations. For every sufficiently small $\gamma > 0$, put*

$$m_\gamma = \left\lfloor \frac{\gamma^{-2}\log(1/\gamma)}{100} \right\rfloor.$$

Then $\Pr(\tau_\gamma > m_\gamma) \ge 63/64$, and hence $\mathbb{E}[\tau_\gamma] = \Omega(\gamma^{-2}\log(1/\gamma))$.

Proof. Write $c_n = c(n,\delta_0)$. For real $n \ge 1$, the derivative of c_n^2 has the sign of

$$1 + \frac{n}{n+1} - \log\frac{2n(n+1)}{\delta_0} < 2 - \log 256 < 0,$$

so c_n decreases in n. Since $2n(n+1)/\delta_0 \ge n^2$, we also have $c_n^2 \ge \log n/n$. For sufficiently small γ, $m_\gamma \ge \gamma^{-1}$. Using this bound in the numerator and the definition of m_γ in the denominator gives

$$c_{m_\gamma}^2 \ge \frac{\log m_\gamma}{m_\gamma} \ge \frac{\log(1/\gamma)}{\gamma^{-2}\log(1/\gamma)/100} = 100\gamma^2.$$

Since c_n decreases, $\gamma \le c_n/10$ for every $n \le m_\gamma$. If the test stops at such an n, then

$$|\widehat{\mu}_n - (1/2+\gamma)| \ge |\widehat{\mu}_n - 1/2| - \gamma > c_n - \gamma \ge \tfrac{9}{10}c_n,$$

and Hoeffding's inequality bounds the probability of this event by

$$2\exp\{-2n(9c_n/10)^2\} = 2\left(\frac{\delta_0}{2n(n+1)}\right)^{81/100}.$$

For $n \le 16$, stopping is impossible on every observation sequence, because $c_n \ge c_{16} > 1/2$ while $|\widehat{\mu}_n - 1/2| \le 1/2$. For larger n, a union bound, the bound $n(n+1) \ge n^2$, and the integral test give

$$\begin{aligned}
\Pr(\tau_\gamma \le m_\gamma) &\le 2\sum_{n=17}^{m_\gamma}\left(\frac{\delta_0}{2n(n+1)}\right)^{81/100} \\
&\le 2\left(\frac{1}{128}\right)^{81/100}\int_{16}^{\infty} x^{-162/100}\,dx \\
&= \frac{100}{31}\,2^{-163/20} < 4\cdot 2^{-8} = \frac{1}{64}.
\end{aligned}$$

The expected-duration bound follows from $\mathbb{E}[\tau_\gamma] \ge m_\gamma \Pr(\tau_\gamma > m_\gamma)$ and $m_\gamma \ge \gamma^{-2}\log(1/\gamma)/200$ for sufficiently small γ. □

G.2 A wrong warm start forces a near-tied comparison

Proof of Proposition 1. Use four arms $a^\star, u, v, w$ with

$$p_{a^\star,x} = \tfrac{1}{2} + d \quad (x \in \{u, v, w\}), \qquad p_{u,v} = \tfrac{1}{2} + \gamma, \qquad p_{u,w} = p_{v,w} = \tfrac{1}{2} + d,$$

together with $p_{j,i} = 1 - p_{i,j}$ and $p_{i,i} = 1/2$. Then $\Delta^\star = d$ and $\Delta = \gamma$ for $0 < \gamma < d$. Arrange the leaves as $(u, w, a^\star, v)$ and start from u. The constant-confidence warm start runs TREEASCENT with $\varepsilon = 1/4$ and $L = 2$, so each merge has confidence $\eta_h = 1/16$. The pass first merges u with w. It then runs BKT on $\{a^\star, v\}$ with confidence $\beta_2 = 1/12$, whose single match has confidence $6\beta_2/\pi^2 = 1/(2\pi^2)$, and finally merges the two winners.

For $\eta \in (0, 1)$, let $k(\eta) = \min\{n \geq 1 : c(n, \eta) < 1/2\}$. This is the stopping time of DUELTEST at confidence η when every observation favors the same arm. Let E be the event that the first merge sees $k(1/16)$ wins of u over w, the BKT match sees $k(1/(2\pi^2))$ wins of $a^\star$ over v, and the final merge sees $k(1/16)$ wins of u over $a^\star$. On E, the warm start returns u. These three tests use fresh samples on pairs with gap d, so

$$q_d := \Pr(E) = (1/2 + d)^{k(1/16)+k(1/(2\pi^2))}(1/2 - d)^{k(1/16)} > 0,$$

which does not depend on γ.

The first SCREENANDREPLACE call starts from the incumbent u. Its screening confidence is $\rho_1 = 4^{-3} = \delta_0$, and its incumbent stays u until every challenger has been screened. Hence DUELTEST(u, v, δ_0) is called whatever the earlier screening outcomes are. The earlier screening tests compare u with $a^\star$ or w. They have gap d and terminate almost surely, so the u–v test is reached. Given their history and E, the u–v comparisons are fresh Bernoulli$(1/2 + \gamma)$ observations, and each incurs one unit of binary weak regret because neither arm is $a^\star$. Lemma 13 therefore gives

$$\mathbb{E}[R_\infty^{\mathrm{weak}}] \geq q_d\, \mathbb{E}[\tau_\gamma] \geq \frac{63 q_d}{64}\, m_\gamma \geq c_d\, \gamma^{-2} \log(1/\gamma)$$

for a constant $c_d > 0$ depending only on d. Since $N = 4$ and $\Delta^\star = d$ stay fixed while this bound diverges as $\gamma \to 0$, Algorithm 5 has no bound depending only on N and $\Delta^\star$. □

H Experimental Protocols and Additional Results

The experiments reported here replace those of the submitted version and of the discussion period. We reran all experiments after correcting numerical oversights in our earlier experimental code and re-checking the baselines against their source papers (Appendix H.2). The code for all experiments is included in the supplementary material.

H.1 Setup

We report cumulative binary weak and strong regret as defined in Section 2. Every identification, exploration, and self-play round counts toward the horizon, and regret averages include runs that are still identifying at the horizon. At the horizon, each online identification is completed-correct (C), completed-wrong (W), or unfinished (U). Unless stated otherwise, $\pm$ denotes the sample standard deviation over seeds. Intervals for a mean are 95% Student-t intervals, and intervals for an event rate are 95% Wilson intervals. Seeds are paired across algorithms within a configuration.

All matrices satisfy $p_{ii} = 1/2$ and $p_{ji} = 1 - p_{ij}$. Unless stated otherwise, the winner is arm 0, the starting arm is $N - 1$, and the balanced tree follows label order.

Cyclic Condorcet family. Set $p_{0j} = 1/2 + \Delta^\star$ for $j > 0$. For nonwinners $1 \leq i < j \leq m$ with $m = N - 1$, set $p_{ij} = 1/2 + d$ if $(j - i) \bmod m \leq m/2$ and $p_{ij} = 1/2 - d$ otherwise. Since N is a power of two, m is odd and no antipodal tie arises. The nonwinners therefore form cycles, and $\Delta = \min\{\Delta^\star, d\}$.

Perturbed cyclic family. The main weak-regret instance uses this construction with $\Delta^\star = 0.26$ and $d = 0.40$. It then adds independent Uniform$[-0.025, 0.025]$ noise to each nonwinner probability p_{ij} with $i < j$ and restores reciprocity. With matrix seed 123, the nonwinner gaps lie in $[0.375, 0.425]$.

Table 2: Experiment configurations. Seeds are per algorithm and cell, and d denotes the nonwinner gap of the cyclic family.

Experiment	Configuration	Seeds
Main BAI	$N = 8, 16, 32, 64, 128, 256$; $(\Delta^\star, d) = (.32, .38)$; $\delta = .1$	100
Main weak	$N = 256$; perturbed cyclic; $T = 50{,}000$ and $200{,}000$	30
Main strong	$N = 256$; $(\Delta^\star, d) = (.30, .38)$; $T = 100{,}000$	30
Gap sweeps	$N = 256$; $\Delta^\star = .30$; $d = .38, .20, .10, .05, .025$; online $T = 10^6$	100
Small winner gap	$N = 256$; $(\Delta^\star, d) = (.05, .20)$; $T = 10^6$	30
Fresh/cache	$N = 256$; $(\Delta^\star, d) = (.30, .38)$; $T = 10^6$	100 pairs
Horizon scaling	$N = 256$; $T = 10^4, 3{\cdot}10^4, 10^5, 3{\cdot}10^5, 10^6$	100
Arm scaling	$N = 64, 128, 256, 512, 1024$; $T = 10^6$; $(\Delta^\star, d) = (.30, .38)$	10
Additional BAI scaling	$N = 8, 16, 32, 64, 128, 256$; $\Delta^\star = .30$; $d = .38, .05$; $\delta = .1$	30
Joint objective	$N = 32, 64, 128, 256$; cyclic and total order; $T = 50{,}000$; $\delta = .1$	100
Anytime tail	$N = 256$; $(\Delta^\star, d) = (.30, .38)$; $T = 10^5, 10^6$	100
Failure stress	$N = 32$; $\Delta = \Delta^\star = g = .05$; $\delta = .5$; $T = 10^6$	5,000

Equal-gap total order. The joint experiment also uses $p_{ij} = 1/2 + 0.30$ for $i < j$, so arm 0 is the winner and $\Delta = \Delta^\star = 0.30$. In the main online panels and the anytime-tail study, the cyclic winner is relabeled to 128, with start 255.

Figure 2(a) uses the first 30 of the 100 BAI seeds, and the native WS-W comparator in the joint study uses the first 30 seeds.

H.2 Algorithms, side information, and source assumptions

WS-W, WR-TINF, WR-EXP3-IX, and Versatile-DB are adapted from the supplementary code of Saad *et al.* [2024], available at `https://openreview.net/forum?id=dY4YGqvfgW`. We corrected the last three to follow their source algorithms, as detailed below, and implemented the rest ourselves.

TG-ITE and BKT. Every gap-oblivious test uses the boundary $c(n, \delta)$ and the strict stopping rule of Algorithm 1, computed in the log domain to handle very small phase confidences. BKT is Algorithm 2 on all arms. The tree allocation, byes, warm-start confidence $1/4$, screening schedule, known-horizon confidence $1/(T + 1)$, and anytime schedule $\epsilon_s = 2^{-2^s}$, $B_s = 2^{2^s}$ follow the stated algorithms. The known-gap variants use Algorithm 9 with the oracle input $g = \Delta^\star$.

SeqElim and Knockout. SeqElim implements Algorithm 1 of Falahatgar *et al.* [2017a], and Knockout implements the tournament and comparison primitives of Falahatgar *et al.* [2017b]. Both target approximate maximum selection. We give them $\delta = 0.1$ and $\epsilon = 0.95\Delta^\star = 0.304$, which uses knowledge of $\Delta^\star$. SeqElim assumes strong stochastic transitivity, and Knockout also assumes the stochastic triangle inequality with constant 1. These assumptions fail on the cyclic matrices, so we report how often the two methods return the exact winner.

WS-W and WS-S. We implement Algorithms 1–2 of Chen and Frazier [2017]. WS-S uses $\beta = 1.1$ and plays $\lfloor \beta^\ell \rfloor$ self-play rounds after its ℓth exploration round. Its condition $1 < \beta < p_{\min}/(1-p_{\min})$, where $p_{\min} = \min_{p_{ij} > 1/2} p_{ij}$, holds on all our matrices, since $p_{\min} \geq 0.525 > 1.1/2.1$.

WR-TINF and WR-EXP3-IX. We implement both algorithms of Saad *et al.* [2024]. WR-TINF uses Tsallis exponent $2/3$, learning rate $2N^{-1/6}/\sqrt{t}$, and challenger weights proportional to $p_i^{2/3}$. After each duel, the sampled challenger, placed on the right, receives the importance-weighted loss Y/q, where q is its conditional sampling probability and $Y = 1$ means that the left arm wins. WR-EXP3-IX uses the source incumbent–challenger construction with rates $\eta_n = \sqrt{\log M/(Mn)}$ and $\gamma_n = \eta_n/2$ for $M = N - 1$ challengers. Both algorithms target weak regret, and we also report their binary strong regret.

Versatile-DB. We use the two Tsallis-regularized learners of Algorithm 3 and Appendix B of Saha and Gaillard [2022], with exponent $1/2$ and learning rate $4/\sqrt{t}$. With $Y_t = 1$ meaning that the left arm wins, the two loss estimates are $(1 - Y_t)/p_{i_t}^{\text{left}}$ and $Y_t/p_{j_t}^{\text{right}}$.

WSW-PE and its continuation. WSW-PE follows Algorithm 3 of Ma *et al.* [2025], keeping cumulative pair counts and scores, with pair radius

$$\sqrt{\frac{\log 8 + \log N - \log \delta + 2 \log n_{ij}}{2 n_{ij}}}.$$

Its $O(N \log N)$ guarantee assumes a total order, so we report total-order results next to the cyclic runs. Since WSW-PE stops after returning an arm, its online runs continue by repeatedly comparing that arm with a companion drawn from the algorithm seed, without knowledge of the winner (WSW-PE+ETC). The joint study also reports the WS-W process embedded in the source.

H.3 Results

Main comparison. Table 3 reports the full 100-seed BAI study behind Figure 2(a). Every method returned the correct winner in all 100 runs at each arm count. BKT uses fewer comparisons than TG-ITE-BAI throughout, and Knockout, which receives the gap-derived tolerance, is faster still. SeqElim has zero spread because its comparison budgets are deterministic.

Table 4 reports the endpoints of Figure 2(b,c) and the weak regret at $2 \cdot 10^5$ rounds. TG-ITE-Weak has a lower mean than WR-TINF, WR-EXP3-IX, and WS-W. WSW-PE+ETC has the lowest mean but a much larger spread. TG-ITE-KH has the lowest strong regret, and TG-ITE-Anytime ends above Versatile-DB, WR-TINF, and WS-S at 10^5 rounds. WR-EXP3-IX always compares distinct arms, so its binary strong regret equals the horizon.

Table 3: Main BAI comparison: comparisons in thousands, $\delta = .1$. All outputs are correct. SeqElim and Knockout receive $\epsilon = .304$.

N	TG-ITE	BKT	SeqElim	Knockout	WSW-PE
8	0.36 ± 0.04	0.33 ± 0.04	0.77 ± 0.00	0.25 ± 0.03	0.52 ± 0.07
16	0.77 ± 0.06	0.69 ± 0.06	1.88 ± 0.00	0.52 ± 0.04	1.21 ± 0.15
32	1.57 ± 0.08	1.36 ± 0.08	4.34 ± 0.00	1.06 ± 0.06	2.67 ± 0.26
64	3.15 ± 0.12	2.71 ± 0.10	9.77 ± 0.00	2.11 ± 0.08	5.75 ± 0.47
128	6.34 ± 0.17	5.36 ± 0.16	21.59 ± 0.00	4.20 ± 0.13	11.95 ± 0.67
256	12.73 ± 0.22	10.65 ± 0.18	47.18 ± 0.00	8.36 ± 0.17	25.07 ± 1.39

Table 4: Main online comparisons: binary regret in thousands. The weak and strong columns use the perturbed and the cyclic matrix of Figure 2, respectively.

Algorithm	Weak, $5 \cdot 10^4$	Weak, $2 \cdot 10^5$	Strong, 10^5
TG-ITE-Weak	10.01 ± 0.18	10.01 ± 0.18	–
WS-W	20.63 ± 20.09	30.00 ± 37.52	–
WR-TINF	18.22 ± 6.44	20.55 ± 8.00	55.20 ± 1.79
WR-EXP3-IX	17.19 ± 6.21	17.66 ± 8.77	100.00 ± 0.00
WSW-PE+ETC	5.54 ± 4.59	5.54 ± 4.59	–
TG-ITE-KH	–	–	22.02 ± 0.27
TG-ITE-Anytime	–	–	62.76 ± 0.37
Versatile-DB	–	–	25.89 ± 1.49
WS-S	–	–	56.35 ± 20.78

Global-gap sweeps. The sweeps fix $\Delta^\star = .30$ and lower the nonwinner gap d, so that Δ falls from .30 to .025. The known-gap variants receive $g = .30$. For BAI (Table 5), the cost of TG-ITE-BAI grows sharply as Δ shrinks, and so does that of BKT, which stays lower. The known-gap variant

remains cheap because its capped tests need not resolve small nonwinner gaps. All three methods are correct in every run.

Online, at $\Delta = .05$ and .025, none of the 100 gap-oblivious warm starts or known-horizon identification passes finishes by $T = 10^6$ (Tables 6 and 7). These runs are unfinished, so weak regret is nearly or exactly T and strong regret is exactly T. At $\Delta = .10$, identification finishes but dominates the cost. Mean weak regret is 229,326 against 15,061 for WR-TINF, and mean strong regret is 398,969 against 78,359 for WS-S. With $g = .30$, every warm start and identification pass finishes correctly at every gap. Mean weak regret then stays between 3,279 and 8,257, and mean strong regret between 7,335 and 24,928.

Table 5: BAI gap sweep: comparisons in thousands, $N = 256$, $\Delta^\star = .30$, $\delta = .1$. All outputs are correct. The g variant receives $g = \Delta^\star = .30$.

d	Δ	TG-ITE-BAI	BKT	TG-ITE-g-BAI
0.38	0.3	12.80 ± 0.23	10.74 ± 0.20	3.73 ± 0.06
0.2	0.2	52.16 ± 1.09	44.93 ± 1.12	6.95 ± 0.15
0.1	0.1	243.64 ± 5.92	216.08 ± 5.48	8.94 ± 0.11
0.05	0.05	$1{,}117.85 \pm 22.81$	$1{,}008.63 \pm 23.27$	9.49 ± 0.09
0.025	0.025	$5{,}075.97 \pm 100.86$	$4{,}648.08 \pm 100.42$	9.62 ± 0.07

Table 6: Online weak-regret gap sweep: regret in thousands, $N = 256$, $\Delta^\star = .30$, $T = 10^6$. C/W/U refers to the warm starts. Oracle-g receives $g = .30$ and is 100/0/0 in every row.

Δ	Gap-oblivious	Oracle-g	WR-TINF	Default C/W/U
0.3	11.12 ± 0.21	3.28 ± 0.07	15.08 ± 2.45	100/0/0
0.2	48.10 ± 1.12	6.08 ± 0.15	14.93 ± 2.26	100/0/0
0.1	229.33 ± 4.84	7.71 ± 0.09	15.06 ± 2.35	100/0/0
0.05	999.76 ± 0.06	8.15 ± 0.07	15.09 ± 2.44	0/0/100
0.025	$1{,}000.00 \pm 0.00$	8.26 ± 0.06	15.02 ± 2.34	0/0/100

Table 7: Online strong-regret gap sweep: regret in thousands, $N = 256$, $\Delta^\star = .30$, $T = 10^6$. C/W/U refers to the identification passes. Oracle-g receives $g = .30$ and is 100/0/0 in every row.

Δ	Gap-oblivious	Oracle-g	WS-S	Default C/W/U
0.3	24.36 ± 0.29	7.33 ± 0.09	75.78 ± 43.83	100/0/0
0.2	92.08 ± 1.45	13.78 ± 0.26	90.55 ± 121.21	100/0/0
0.1	398.97 ± 6.63	20.93 ± 0.29	78.36 ± 62.79	100/0/0
0.05	$1{,}000.00 \pm 0.00$	24.09 ± 0.20	85.21 ± 80.83	0/0/100
0.025	$1{,}000.00 \pm 0.00$	24.93 ± 0.15	82.55 ± 58.84	0/0/100

Small winner gap. With $\Delta^\star = .05$ and nonwinner gap .20 (Table 8), both TG-ITE warm starts finish correctly in all 30 runs, after $85{,}102 \pm 4{,}623$ comparisons for the gap-oblivious policy and $40{,}177 \pm 1{,}201$ for the known-gap policy. These counts exceed the weak regrets because comparisons with the winner incur no weak loss. Both policies incur far less weak regret than WR-TINF.

Cost of fresh samples. The ablation compares TG-ITE-Anytime with a heuristic that keeps win counts for every pair across phases and reuses them in later calls, using the same matrix, tree, start, and 100 paired seeds. This heuristic has no guarantee. At $T = 10^6$, the fresh implementation uses 95,091 new identification comparisons on average and the cache heuristic 32,339. Their ratio is 2.940, with paired 95% bootstrap interval $[2.935, 2.946]$ from 20,000 resamples. Both complete five identification passes in every run, and all their outputs are correct (Table 9). Phases 2–5 begin at the winner in every run and account for 87.59% of the fresh comparisons. The first phase costs the same in both, so the saving comes from cheaper repeated confirmation.

Table 8: Small-winner-gap experiment: binary weak regret in thousands at $T = 10^6$, $N = 256$, $\Delta^\star = .05$, $d = .20$. The g variant receives $g = .05$.

Algorithm	Mean $\pm$ sample SD	95% mean interval
TG-ITE-Weak	48.09 ± 1.11	$[47.68, 48.51]$
TG-ITE-g-Weak	32.19 ± 0.73	$[31.92, 32.46]$
WR-TINF	467.70 ± 102.55	$[429.41, 505.99]$

Table 9: New identification comparisons by phase, in thousands, over 100 paired seeds. The last column counts runs that enter the phase with the correct incumbent, which is the same for both.

Phase	Fresh samples	Cache heuristic	Start at winner
1	11.80 ± 0.28	11.80 ± 0.28	0/100
2	13.22 ± 0.23	1.61 ± 0.07	100/100
3	16.13 ± 0.24	2.70 ± 0.10	100/100
4	21.62 ± 0.27	5.50 ± 0.15	100/100
5	32.32 ± 0.33	10.73 ± 0.19	100/100
Total	95.09 ± 0.64	32.34 ± 0.37	–

Scaling in the horizon and the number of arms. Horizon scaling evaluates prefixes of one 10^6-round trajectory per seed for TG-ITE-Weak and TG-ITE-Anytime, and it reruns TG-ITE-KH for each horizon (Table 10). Weak regret stays constant from $3 \cdot 10^4$ rounds on. TG-ITE-KH grows slowly, from 22,088 at 10^5 to 24,332 at 10^6. TG-ITE-Anytime rises to 95,214 by $3 \cdot 10^5$ and then stays flat, because both later endpoints fall in the self-play block of its fifth phase. At $T = 10^4$, no KH pass finishes, so its regret is T, while at the other horizons every pass finishes correctly.

Arm scaling compares the three online policies with WS-W and WS-S over 10 seeds per cell (Table 11). From $N = 64$ to 1024, mean regret divided by N stays between 40.1 and 45.2 for TG-ITE-Weak, 94.5 and 97.3 for TG-ITE-KH, and 370.2 and 379.8 for TG-ITE-Anytime, so all three grow linearly in N. WS-W has lower weak regret only at $N = 64$. TG-ITE-KH has lower strong regret than WS-S at every N, whereas TG-ITE-Anytime does not. In the additional 30-seed BAI sweep (Table 12), TG-ITE-BAI and BKT are correct in every run, and BKT is cheaper throughout. At $N = 256$, lowering Δ from .30 to .05 raises the mean cost from 12,784 to 1,112,906 for TG-ITE-BAI and from 10,722 to 1,005,190 for BKT.

Table 10: Horizon scaling at $N = 256$, $(\Delta^\star, d) = (.30, .38)$: binary regret in thousands. Weak reports weak regret, and KH and Anytime report strong regret.

T	Weak	KH	Anytime
10,000	9.82 ± 0.05	10.00 ± 0.00	10.00 ± 0.00
30,000	11.13 ± 0.20	20.92 ± 0.28	29.98 ± 0.00
100,000	11.13 ± 0.20	22.09 ± 0.30	62.84 ± 0.48
300,000	11.13 ± 0.20	23.16 ± 0.29	95.21 ± 0.56
1,000,000	11.13 ± 0.20	24.33 ± 0.31	95.21 ± 0.56

Tree and starting arm. With $N = 256$ and $\delta = .1$, we ran TG-ITE-BAI on 50 random leaf orders with 10 shared seeds each. The mean cost per tree ranges from 12,744 to 13,047, a max/min ratio of 1.0238, and all 500 outputs are correct. Varying the winner-entry level $h_\star$, the first ancestor level that contains the winner, over $0, \dots, 8$ with 30 seeds each gives mean costs between 12,700 and 12,755, again with all outputs correct. The cost is thus insensitive to the tree and the starting arm.

Joint output. At horizon $H = 50{,}000$, the joint study compares TG-ITE-Dual, which runs a δ-confidence TreeAscent and then the weak-regret policy, with TG-ITE-Weak, which keeps its confidence-$1/4$ warm start, and with WSW-PE+ETC. In Table 13, every identification and warm start finishes with the correct winner. On the cyclic instance with $N = 256$, Dual identifies after 12,732 comparisons on average against 27,999 for WSW-PE, but its weak regret is 12,035 against

Table 11: Online arm-count scaling: binary regret in thousands, $T = 10^6$, $(\Delta^\star, d) = (.30, .38)$. Weak and WS-W report weak regret, and KH, Anytime, and WS-S report strong regret.

N	Weak	WS-W	KH	Anytime	WS-S
64	2.56 ± 0.09	1.09 ± 1.17	6.23 ± 0.12	24.31 ± 0.31	13.68 ± 1.20
128	5.38 ± 0.14	7.05 ± 9.14	12.37 ± 0.21	47.93 ± 0.41	32.40 ± 9.20
256	11.15 ± 0.25	20.09 ± 34.50	24.35 ± 0.34	95.28 ± 0.64	70.84 ± 34.18
512	22.80 ± 0.35	100.56 ± 167.82	48.57 ± 0.37	190.08 ± 1.03	200.75 ± 165.25
1024	46.24 ± 0.33	138.47 ± 118.15	96.78 ± 0.63	379.04 ± 1.17	336.06 ± 114.73

Table 12: Additional BAI arm-count scaling: comparisons in thousands, $\Delta^\star = .30$, $\delta = .1$. All outputs are correct. Column labels give Δ, with nonwinner gap .38 or .05.

N	TG-ITE, .30	BKT, .30	TG-ITE, .05	BKT, .05
8	0.40 ± 0.06	0.38 ± 0.05	17.12 ± 2.86	15.62 ± 2.53
16	0.81 ± 0.06	0.75 ± 0.06	47.41 ± 3.45	43.93 ± 3.63
32	1.61 ± 0.08	1.43 ± 0.08	111.55 ± 7.99	101.72 ± 5.42
64	3.21 ± 0.13	2.78 ± 0.12	250.51 ± 10.86	228.80 ± 7.73
128	6.40 ± 0.15	5.44 ± 0.16	535.00 ± 17.07	482.54 ± 12.92
256	12.78 ± 0.19	10.72 ± 0.19	$1{,}112.91 \pm 26.52$	$1{,}005.19 \pm 22.20$

4,009 for WSW-PE+ETC. On the total order, the weak regrets are 20,547 and 560. WSW-PE+ETC makes more comparisons yet incurs less weak regret, because comparisons with the winner are free. Over 30 seeds at $N = 256$, the native WS-W process has weak regret $18{,}063 \pm 18{,}422$ on the cyclic instance and 526 ± 472 on the total order. In 800 further paired runs, with 100 seeds at each of four arm counts in both families, standalone TreeAscent and the identification stage of Dual return the same correct arm after the same number of comparisons.

Table 13: Joint evaluation in thousands, $H = 50{,}000$. τ is the $\delta = .1$ identification time for Dual and WSW-PE and the confidence-$1/4$ warm-start time for TG-ITE-Weak. The total-order assumption of WSW-PE holds only in the lower block.

N	Statistic	TG-ITE-Dual	TG-ITE-Weak	WSW-PE+ETC
Cyclic Condorcet				
32	$\tau \wedge H$	1.62 ± 0.10	1.48 ± 0.09	3.03 ± 0.27
	R_H	1.20 ± 0.07	1.10 ± 0.07	0.25 ± 0.34
64	$\tau \wedge H$	3.21 ± 0.12	2.96 ± 0.12	6.50 ± 0.34
	R_H	2.71 ± 0.10	2.49 ± 0.10	0.81 ± 0.90
128	$\tau \wedge H$	6.39 ± 0.16	5.90 ± 0.16	13.45 ± 0.54
	R_H	5.78 ± 0.15	5.33 ± 0.13	1.79 ± 1.56
256	$\tau \wedge H$	12.73 ± 0.22	11.74 ± 0.21	28.00 ± 1.03
	R_H	12.04 ± 0.22	11.08 ± 0.20	4.01 ± 3.68
Equal-gap total order				
32	$\tau \wedge H$	2.47 ± 0.15	2.29 ± 0.15	2.92 ± 0.19
	R_H	2.06 ± 0.14	1.91 ± 0.13	0.07 ± 0.07
64	$\tau \wedge H$	5.13 ± 0.21	4.76 ± 0.22	6.27 ± 0.26
	R_H	4.62 ± 0.20	4.29 ± 0.21	0.15 ± 0.12
128	$\tau \wedge H$	10.49 ± 0.31	9.71 ± 0.28	13.09 ± 0.37
	R_H	9.88 ± 0.29	9.15 ± 0.28	0.28 ± 0.21
256	$\tau \wedge H$	21.24 ± 0.45	19.70 ± 0.45	27.33 ± 0.50
	R_H	20.55 ± 0.44	19.04 ± 0.44	0.56 ± 0.36

Anytime tails and failures. We followed 100 TG-ITE-Anytime trajectories to 10^6 rounds on the cyclic instance of Figure 2(c). The observed maximum is only 1.018 times the mean at 10^5 and 1.017 times at 10^6 (Table 14 and Figure 3). None of the 500 completed identification passes returns a

wrong arm. Treating trajectories as independent trials, the one-sided exact 95% upper bound on the probability that a run contains a wrong identification is 2.95%.

To observe failures, a stress test uses $N = 32$, a cyclic matrix with $\Delta = \Delta^\star = g = .05$, loose confidence $\delta = .5$, and 5,000 seeds. Each run performs one TreeAscent$_g$ pass and then self-plays its output to $T = 10^6$, so a wrong output incurs strong regret T. All 5,000 passes finish, and 27 return a wrong arm, a rate of 0.54% with Wilson interval $[0.371\%, 0.785\%]$. Their regret 10^6 is 39.50 times the mean 25,316 over correct runs. These 27 runs alone account for the gap between the all-run mean 30,579 and the median 25,372 (Table 15).

Table 14: Tail statistics of TG-ITE-Anytime: strong regret in thousands over 100 trajectories, $N = 256$, $(\Delta^\star, d) = (.30, .38)$. Both horizons use the same trajectories.

T	Mean $\pm$ sample SD	Median	95th percentile	Maximum
100,000	62.81 ± 0.45	62.79	63.51	63.92
1,000,000	95.18 ± 0.56	95.18	96.03	96.77

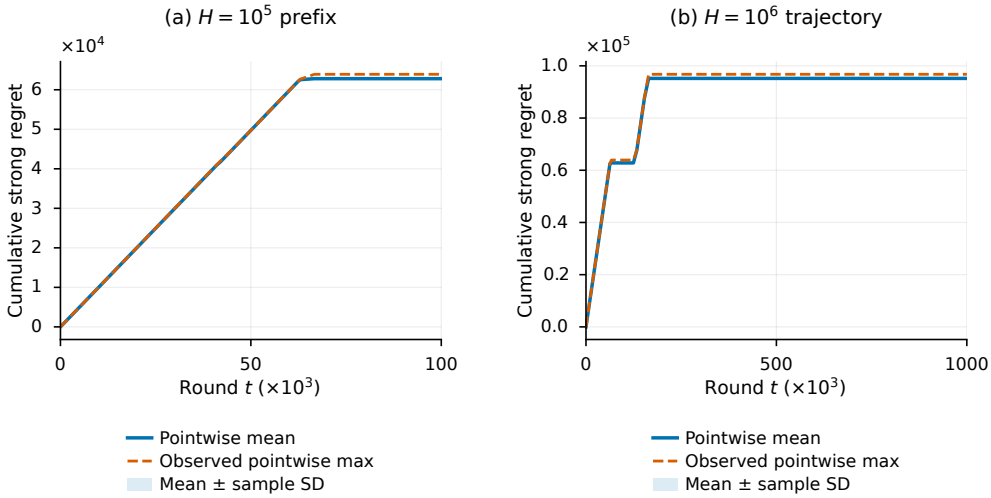


Figure 3: Strong regret of the same 100 TG-ITE-Anytime trajectories up to 10^5 and 10^6. The solid line is the mean, the band is mean $\pm$ sample SD, and the dashed line is the pointwise maximum.

Table 15: Failure stress test: strong regret in thousands, grouped by the output of the pass.

Condition	Runs	Mean $\pm$ SD	Median	95th pct.	Maximum
All runs	5,000	30.58 ± 71.45	25.37	27.21	1,000.00
Correct output	4,973	25.32 ± 1.14	25.37	27.15	29.26
Wrong output	27	$1{,}000.00 \pm 0.00$	1,000.00	1,000.00	1,000.00

H.4 Computational resources

Our experiments are conducted on a server with an AMD Ryzen Threadripper 1950X CPU and 256GB RAM. With 32 parallel workers, all experiments complete within 5 minutes.

I Potential Societal Impact

This work is primarily theoretical, and its experiments use only synthetic preference matrices. In ranking, recommendation, and human-preference learning, where people make the comparisons,

needing fewer comparisons reduces the effort asked of them. The identified arm, however, is the Condorcet winner of the preferences actually collected, so it can favor some groups over others if the people who provide the comparisons do not represent those whom a system serves.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, we have clearly introduced the problem we study, the progress and limitations of the existing methods, and our contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations and future work in Section 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We have clearly introduced the assumptions in Sec. 2. The complete proofs are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have included detailed introductions to the algorithmic processes of the proposed approaches. The detailed experimental setups are also included in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code for all experiments is included in the supplementary material. The script `reproduce.sh` runs every experiment and summarizes the results, using `uv` with Python 3.12 and the package versions pinned in `requirements.txt`. Appendix H specifies the instances, parameters, and seeds.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The detailed experimental setups are included in Sec. 6 and Appendix H.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We have conducted repeated experiments and report variances in our results to ensure statistical significance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have discussed the computational resources in Appendix H.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We ensure our work follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss potential positive and negative societal impacts in Appendix I.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We believe that this paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have discussed the assets used in our paper in Appendix H.2.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new asset is the experiment code in the supplementary material. Its usage is documented in `reproduce.sh`, `run.py`, and `analyze.py`, and Appendix H describes the experiments it runs.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.