TIE-SAFE STABLE LEARNING FOR TWO-SIDED UNKNOWN MATCHING MARKETS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper characterizes, up to universal constants, the class-wise asymptotic logarithmic regret of two-sided matching markets with unknown preferences and possible exact ties. Under aggregate regret against the coordinate-wise lowest player payoffs among weakly stable matchings, learning is governed by comparisons that certify stability or rule out unstable outcomes, rather than complete preference recovery. A signed certificate view isolates two complementary structural margins. The certificate margin Δ measures how strongly some stable matching rejects every deviation, while, when $\Delta = 0$, the preservation margin Δ_{pres} measures the smallest blocking strength of any unstable matching. For N players and $K \geq N$ arms, both regimes obey the same core–tail complexity law

$$\Psi(h) = \frac{N^2}{h^2} + \frac{N(K-N)}{h},$$

where the two terms correspond to resolving core comparisons and screening outside arms. Upper and policy-uniform lower bounds match, establishing the sharp class-wise rate $\Theta(\Psi(h)\log T)$, with $h = \Delta$ when $\Delta > 0$ and $h = \Delta_{\mathrm{pres}}$ at the zero-certificate boundary. Neither the leading logarithmic term nor the horizon-independent remainder depends on the global minimum positive preference gap. Aggregate Core–Tail Learning (ACTL) attains both guarantees without knowing either margin, the tie pattern, or the underlying regime. A communication-aware decentralized variant preserves the same class-wise $O(\Psi(h)\log T)$ order over connected bounded-bandwidth local networks when communication consumes physical time.

1 INTRODUCTION

Two-sided matching models online platforms that repeatedly pair participants with initially unknown preferences, including labor, recommendation, and dating markets. Because learning is coupled with allocation, observing an edge requires forming a match and can displace another participant (Liu et al., 2020; Su et al., 2022; Li et al., 2025; Basu, 2025). Repeated exploration can therefore impose market-wide losses before a stable assignment—one with no mutually preferred deviation—is found. We study N players and $K \geq N$ arms over a horizon T, allowing exact ties. Under weak preferences, a player-optimal weakly stable matching need not exist (Manlove, 2002); we instead use each player's lowest payoff across weakly stable matchings. This coordinate-wise baseline is always defined, requires no matching-selection rule, and gives zero regret to every stable outcome. Summing positive losses prevents gains to one player from offsetting losses to another (Section 3). We ask *how much such loss is unavoidable when both sides' preferences are unknown, and whether one policy can attain that cost without knowing the regime*.

Learning every preference comparison can be wasteful (Athanasopoulos et al., 2026). Suppose player i slightly prefers arm a to her current partner, but a strictly prefers its current partner to i. The pair cannot block no matter how small the player's preference gap. Thus stability can be certified by one rejecting endpoint per deviation without recovering complete rankings. This motivates the *certificate margin* Δ: the largest margin by which some matching rejects every deviation (Definition 1). In strict markets, Δ equals the admissible gap of super-stability analysis (Proposition 2; Basu, 2025) and can stay bounded away from zero even as the global minimum positive preference gap $\Delta_{\min}$ tends to zero.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Exact ties under weak preferences (Irving, 1994; Manlove et al., 2002) change the learning task. If a player is indifferent between her partner and an arm that prefers her to its partner, the pair does not block but has zero rejection margin. If every stable matching has such a boundary pair, then $\Delta = 0$. Unstable matchings nevertheless contain blocking pairs with strictly positive gains. The *preservation margin* Δ_{pres} measures the smallest blocking strength over unstable matchings (Definition 2); it is the largest threshold at which partial information still identifies exactly the weakly stable set (Proposition 1). Thus Δ and Δ_{pres} measure opposite sides of the signed certificate $g(W)$: the former supports positive certification, while the latter supports learning when $\Delta = 0$.

Let h denote the relevant margin: $h = \Delta$ when $\Delta > 0$ and $h = \Delta_{\mathrm{pres}}$ when $\Delta = 0$. Both cases obey the same core–tail law

$$\Psi(h) = \frac{N^2}{h^2} + \frac{N(K-N)}{h}.$$

The two terms reflect distinct exploration costs. In the lower-bound families, resolving $\Theta(N^2)$ core comparisons requires $\Omega(\log T/h^2)$ observations at constant loss per test. The $N(K-N)$ outside edges require resolution at the same scale, but each probe costs only $\Theta(h)$, producing the second term. ACTL matches both charges, so the class-wise asymptotic order $\Theta(\Psi(h)\log T)$ of Section 5.3 describes an unavoidable cost rather than a feature of one algorithm. The lower bounds hold for every policy, even if it knows the hard family and its scale; ACTL's leading order and horizon-independent remainder do not depend on $\Delta_{\min}$.

Aggregate Core–Tail Learning (ACTL) attains this law without knowing the regime through three ingredients. A pilot chooses the baseline and decides whether positive certification is statistically plausible. A single core–tail validation engine, shared by both cases, then resolves the core through cyclic matchings and screens outside arms with shortfall-adaptive probes, retaining all observations. Certificate tests commit the market to a stable matching as soon as the evidence suffices; if positive certification remains unresolved, ACTL refines threshold comparisons and uses a UCB–Gale–Shapley fallback.

The same core–tail law persists when endpoint rewards remain local. A communication-aware variant synchronizes quantized summaries over connected local links while matching continues, preserving the leading $O(\Psi(h)\log T)$ order even when communication consumes physical time (Section 5.4).

Contributions.

- **Signed certificate characterization.** A signed certificate unifies positive certification and preservation, and characterizes the partial information sufficient for each regime (Section 3).
- **Sharp core–tail minimax law.** Policy-uniform lower bounds and adaptive upper bounds match up to universal constants in both margin classes (Sections 4–5.3).
- **One regime-adaptive policy.** ACTL attains both rates without knowing Δ, Δ_{pres}, the tie pattern, or the regime, with a horizon-independent additive remainder and logarithmic individual regret (Section 5.2).
- **Communication-aware decentralization.** LC-ACTL preserves the class-wise core–tail order over connected bounded-bandwidth local networks even when communication consumes physical time. In-band signaling retains logarithmic regret without an auxiliary channel (Section 5.4).

2 Related Work

Matching bandits and ties. Matching theory (Gale & Shapley, 1962; Gusfield & Irving, 1989; Roth & Sotomayor, 1990) has bandit extensions (Das & Kamenica, 2005; Liu et al., 2020). Published player-optimal bounds for strict markets with known or structured opposite-side priorities use a minimum player-side gap (Kong & Li, 2023; Zhang et al., 2022; Kong et al., 2024). Under our coordinate-wise stable baseline, the two-sided regret bound is instead governed by certificate margins and is independent of global $\Delta_{\min}$. With ties, Kong et al. (2025) obtain $O(NK\log T/\Delta_{\min}^2)$ player stable regret, while Lin et al. (2025) give worker-optimal approximations and a lower bound across large-gap and near-tied regimes. For full-information stability notions and other learning variants, see Appendix J (Irving, 1994; Manlove et al., 2002; Manlove, 2013; Li et al., 2025).

Super-stability and two-sided learning. Basu (2025) obtains admissible-gap logarithmic pessimal stable regret; in strict markets, their gap equals our certificate margin (Proposition 2). They also give a binary stable regret lower bound and shared-flag decentralization. Under strict preferences, Athanasopoulos et al. (2026) use pervasive stable matchings (Rastegari et al., 2014) to identify an optimal stable matching from partial profiles without global $\Delta_{\min}$ in mismatch counts, with a regret extension (see also Athanasopoulos et al., 2025).

Decentralized matching bandits. Prior work coordinates through outcomes (Liu et al., 2021), shared boards or flags (Pagare & Ghosh, 2024; Basu, 2025), or market-action signaling (Sankararaman et al., 2021; Basu et al., 2021); communication-free learning is also studied (Maheshwari et al., 2022). We charge synchronization to regret with two-sided unknown rewards and exact ties (Appendix J). Table 1 compares guarantees.

Table 1: Logarithmic stable regret guarantees under different regret criteria.

Work	Setting	Regret criterion	Main guarantee
Kong & Li (2023); Kong et al. (2024)	One-sided; strict	Player-optimal stable reward	$\Delta_{\min}$-dependent upper bounds
Kong et al. (2025)	One-sided; ties	Player-pessimal stable reward	$O(NK \log T/\Delta^2_{\min})$ per player
Basu (2025)	Two-sided; strict	Player-pessimal stable reward	$O(K \log T/\Delta^2)$ per player
Athanasopoulos et al. (2026)	Two-sided; strict	Optimal-stable mismatch count	Logarithmic; no global $\Delta_{\min}$ dependence
This paper	**Two-sided; ties**	**Aggregate stable reward**	$\Theta(\Psi(\Delta)\log T)$, $\Delta > 0$; $\Theta(\Psi(\Delta_{\text{pres}})\log T)$, $\Delta = 0$

3 SETTING

There are N players $\mathcal{I} = [N]$, $K \geq N$ arms $\mathcal{A} = [K]$, and $T \geq 1$ rounds; write $\mathcal{M}$ for the injections from $\mathcal{I}$ to $\mathcal{A}$, and denote a market instance by ν. At round t, a policy selects $M_t \in \mathcal{M}$ based on the observed history, possibly at random. Write $M(i)$ for player i's partner and $M^{-1}(a)$ for arm a's partner, with $M^{-1}(a) = \perp$ if it is unmatched. A partner is *real* when it is not the unmatched option $\perp$; all real partners are acceptable, meaning strictly preferred to being unmatched.

When M_t is played, each matched edge (i, a) reveals two endpoint rewards with means $\mu^P_{i,a}, \mu^A_{a,i} \in (0, 1]$; conditional on the history and chosen matching, each centered observation is 1-sub-Gaussian. Set $\mu^A_{a,\perp} = 0$. Preferences are induced by the corresponding mean rewards: a strict market has no ties among real partners on any preference list, while a weak market allows ties. This repeated-matching model is standard (Liu et al., 2020; 2021; Basu et al., 2021; Pokharel & Das, 2023; Zhang & Fang, 2024b; Basu, 2025). Section 5.4 extends it to decentralized physical time, with communication slots included in the horizon and regret.

Stable rewards and aggregate regret. A pair (i, a) blocks M if $\mu^P_{i,a} > \mu^P_{i,M(i)}$ and $\mu^A_{a,i} > \mu^A_{a,M^{-1}(a)}$. Call M *weakly stable* if no pair blocks it; such matchings always exist, and weak stability is the standard concept under indifference (Irving, 1994; Manlove et al., 2002; Erdil & Ergin, 2017). Let $\mathcal{W}^\star(\nu)$ be the set of weakly stable matchings and define

$$\beta_i := \min_{W \in \mathcal{W}^\star} \mu^P_{i,W(i)}, \qquad \alpha_a := \min_{W \in \mathcal{W}^\star} \mu^A_{a,W^{-1}(a)}.$$

The vector $(\beta_i)_i$ is a coordinate-wise stable baseline and need not be jointly attained by a single weakly stable matching. For $(z)_+ := \max\{z, 0\}$, the regret processes are

$$R_i(T) := \sum_{t=1}^{T} (\beta_i - \mu^P_{i,M_t(i)})_+, \qquad R^A_a(T) := \sum_{t=1}^{T} (\alpha_a - \mu^A_{a,M_t^{-1}(a)})_+, \tag{1}$$

with $R_P^{\Sigma}(T) := \sum_{i=1}^{N} R_i(T)$ the aggregate player regret, which is our main objective. Unlike in strict matching markets, a player-optimal weakly stable matching need not exist under indifference (Manlove, 2002). Thus a benchmark tied to a single player-optimal weakly stable outcome is not generally available without an additional selection rule. The quantity β_i is instead defined for every weak market and is no larger than player i's payoff in any weakly stable matching; hence every weakly stable outcome incurs zero positive-part regret for every participant. Appendix I shows a learning barrier to simultaneously requiring player-optimal player regret and arm-pessimal arm regret on nearby strict markets without a known gap lower bound.

Positive reward gaps. Let $\Delta_{\min}$ be the smallest positive difference between two real-partner means on the same preference list, with $\Delta_{\min} = 1$ if none exists. Exact ties and comparisons involving the unmatched option $\perp$ are excluded. For stability, the relevant question is instead whether at least one endpoint rejects each proposed deviation.

Definition 1 (Stability-certificate margin)**.** *For $W \in \mathcal{M}$ and $a \neq W(i)$, define*

$$c_W(i,a) := \begin{cases} \max\{\mu^P_{i,W(i)} - \mu^P_{i,a},\ \mu^A_{a,W^{-1}(a)} - \mu^A_{a,i}\}, & a \in W(\mathcal{I}), \\ \mu^P_{i,W(i)} - \mu^P_{i,a}, & a \notin W(\mathcal{I}). \end{cases}$$

The margin of W and the best margin in the market are

$$g(W) := \min\big(\{1\} \cup \{c_W(i,a) : a \neq W(i)\}\big), \qquad \Delta := \max_{W \in \mathcal{M}} g(W).$$

For an arm unmatched by W only the player's rejection enters the certificate. When $c_W(i,a) < 0$, the pair (i,a) blocks W, and $-c_W(i,a)$ is the minimum improvement among the required endpoints. The value $c_W(i,a)$ measures how strongly at least one endpoint rejects the deviation, and $g(W)$ is the weakest rejection certificate. Consequently, W is weakly stable exactly when $g(W) \geq 0$; for an unstable matching, $-g(W) > 0$ is its blocking strength. Also $0 \leq \Delta \leq 1$: strict markets satisfy $\Delta \geq \Delta_{\min} > 0$, while weak markets can have $\Delta = 0$. Thus $g(W)$ provides a signed certificate of stability.

Definition 2 (Preservation margin)**.** *The preservation margin of the market is*

$$\Delta_{\text{pres}} := \min\big(\{1\} \cup \{-g(W) : W \in \mathcal{M},\ g(W) < 0\}\big).$$

The set convention encodes the case in which no unstable matching exists: then $\Delta_{\text{pres}} = 1$. Every unstable W satisfies $g(W) \leq -\Delta_{\text{pres}}$.

Reading $g(W)$. The sign of $g(W)$ separates unstable matchings ($g(W) < 0$), the stability boundary ($g(W) = 0$), and matchings with a positive certificate ($g(W) > 0$). The margin $\Delta = \max_W g(W)$ is the largest certificate value, while Δ_{pres} is the smallest blocking strength among unstable matchings. They measure the two sides of the boundary and need not be ordered.

The two cases for Δ. If $\Delta > 0$, a stable matching attains $g(W) = \Delta > 0$, so learning can focus on positively certifying one outcome. If $\Delta = 0$, every stable matching has $g(W) = 0$: stability gives $g(W) \geq 0$, while the market-wide maximum is zero. Positive certification is then impossible, but every unstable matching satisfies $g(W) \leq -\Delta_{\text{pres}}$. Learning must distinguish the stable boundary from the closest unstable outcome, which explains the role of Δ_{pres}.

Why thresholded partial information? Suppose we retain only comparisons whose reward difference is at least h and leave smaller ones unresolved. Which matchings can still appear stable under a strict completion? Proposition 1 gives an exact answer and an operational interpretation of Δ_{pres}.

Proposition 1 (Preservation characterization)**.** *Let π_h be the population threshold profile of Appendix E, retaining exactly the comparisons with reward difference at least h, and let $\mathcal{W}_h$ be the set of matchings that are weakly stable under some strict linear extension of π_h. Then $\mathcal{W}_h = \{W \in \mathcal{M} : g(W) > -h\}$ for every $h \in (0,1]$, and consequently $\Delta_{\text{pres}} = \max\{h \in (0,1] : \mathcal{W}_h = \mathcal{W}^\star\}$.*

If $g(W) \leq -h$, some blocking pair improves by at least h at every required endpoint, survives thresholding, and rules out W under every completion. If $g(W) > -h$, no blocking pair is forced at

resolution h; the linear-extension argument of Appendix F yields a completion under which W is stable. Thus the threshold profile excludes exactly the matchings with $g(W) \le -h$, and Δ_{pres} is the coarsest resolution that preserves the true weakly stable set.

The positive side recovers a familiar quantity in strict markets. Basu (2025) defines the admissible gap $\Delta_{\mathcal{A}}$ as the largest minimum reward gap over partial ranks consistent with the true ranking that admit a super-stable matching.

Proposition 2 (Strict markets: the certificate margin is the admissible gap)**.** *In a strict market, $\Delta = \Delta_{\mathcal{A}}$. Both quantities equal the largest $h \in (0, 1]$ for which the threshold profile (16) admits a super-stable matching.*

Thresholding at h converts $g(W) \ge h$ into super-stability under the corresponding partial profile, and maximizing h yields the equality; see Appendix E. The equivalence is specific to strict markets, since the admissible gap is defined relative to a strict full ranking.

Both margins are governed by a single rate: for $0 < h \le 1$, write $\Psi(h) := \frac{N^2}{h^2} + \frac{N(K-N)}{h}$. Then $\Psi(h) \le N^2/h^2 + NK/h \le 2\Psi(h)$: the two forms of the rate differ by at most a factor of two.

Having identified the two structural margins, we next show that neither can be improved statistically: both induce the same unavoidable core–tail complexity.

4 Lower Bound

The two terms in $\Psi(d)$ reflect equally informative tests with different losses: constant loss when a core player is displaced, order-d loss when probing an outside arm; d is the construction scale, $C, c > 0$ universal constants.

4.1 Positive Certificate Margin

Theorem 1 (Aggregate lower bound)**.** *Fix any $N = 2n + 1 \ge 3$, $K \ge N$, and $d \in (0, 1/64]$. There exist explicit finite families $\mathcal{H}^{\mathrm{str}}_{N,K,d}$ and $\mathcal{H}^{\mathrm{tie}}_{N,K,d}$ of strict instances and of instances with exact ties. Every instance admits a unique weakly stable matching with $d/2 \le \Delta \le 2d$, where $\Delta_{\min} = d$ in the tied family and $\Delta_{\min} \ge d/(64K)$ in the strict family. There is a threshold $T_0(N, K, d)$ independent of the policy such that, for either regime $s \in \{\mathrm{str}, \mathrm{tie}\}$, every horizon-indexed learning policy (possibly randomized) satisfies*

$$\sup_{\nu \in \mathcal{H}^{s}_{N,K,d}} \mathbb{E}_{\nu} R^{\Sigma}_{P}(T) \ge c\Psi(d) \log T, \qquad T \ge T_0. \tag{2}$$

The bound is policy-uniform: it holds even for policies that know the families and d, and extends to even $N \ge 4$ by a private dummy pair with different universal constants.

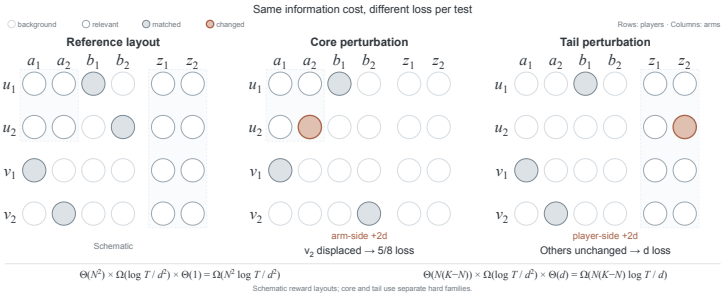

Figure 1: **Core and tail lower-bound mechanisms.** The reward layouts are schematic; the formal core and tail constructions are separate hard families given in Appendix C. At construction scale d, each alternative changes one endpoint mean by $2d$. Core probes incur constant aggregate loss, whereas outside probes incur order-d loss.

Each alternative in Figure 1 changes one mean by $2d$, and with it which matching is uniquely stable; common tags remove ties.

Core: the displaced player pays. In the core base, u_i holds its home b_i and guard v_j holds a_j. A probe (u_i, a_j) can benefit u_i, but displaces v_j at constant loss; the spare arm only transfers that loss to the anchor. Thus even parallel probes pay a constant per edge across $\Theta(N^2)$ arm-side changes.

Tail: only the probing player pays. In a separate family, a player's home exceeds each outside arm by d, and raising one outside player-side mean by $2d$ changes that player's stable partner; an outside probe costs the probing player only d, with no other player displaced, across $N(K-N)$ changes.

From information to regret. Each changed unit-variance Gaussian coordinate carries $2d^2$ KL divergence per observation, so beating the claimed bound on every instance requires $\Omega(\log T/d^2)$ samples of every changed edge, contradicting the regret charged to that edge; the maximum expected regret over the union is at least half the sum of the two charges. The finite-family testing lemma gives a policy-independent threshold (Appendix D).

4.2 Zero Certificate Margin

The tied families above retain positive certificates. A 2×2 tie module instead yields hard instances in $\mathcal{C}_{\mathrm{pres}}(h) := \{\nu : \Delta(\nu) = 0,\ \Delta_{\mathrm{pres}}(\nu) \geq h\}$.

Theorem 2 (Zero-certificate lower bound)**.** *Fix $N \geq 5$, $K \geq N$, and $0 < h \leq 1/64$. There is a threshold $T_0(N, K, h)$, independent of the policy, such that every horizon-indexed learning policy satisfies*

$$\sup_{\nu \in \mathcal{C}_{\mathrm{pres}}(h)} \mathbb{E}_\nu R_P^\Sigma(T) \geq c \left(\frac{N^2}{h^2} + \frac{N(K-N)}{h} \right) \log T, \qquad T \geq T_0.$$

With $d = h$, append the module to either tied family on $N-2$ players and $K-2$ arms: its unused, doubly tied pair forces $\Delta = 0$, and the module and cross edges add neither information nor cheaper probes, preserving $\Delta_{\mathrm{pres}} = h$, the $K-N$ outside arms, and the rate up to constants (Appendix C.4).

5 Upper Bound

The lower bound dictates what an optimal algorithm must do: resolve the core at constant loss, screen outside edges by shortfall. One adaptive policy, ACTL, does exactly that; Sections 5.1–5.4 describe it, state and match its guarantees, and present the paid local-communication extension.

5.1 Aggregate Core–Tail Learning

ACTL (Algorithm 1) has three phases: a pilot selects a baseline and target mode; core–tail validation resolves relevant comparisons; and, once targets are complete, the algorithm tests for safe commitment or replays a cached Gale–Shapley (GS) matching computed from the current snapshot of upper-confidence-bound (UCB) estimates. We call this matching the UCB–GS cache. Certify tries to prove that one candidate is stable with a positive certificate. Preserve instead learns enough threshold comparisons to exclude unstable outcomes when positive certification cannot be established. Appendix E gives the executable specification.

The common validation engine. Fix a baseline W_c and its assigned-arm set $A_c := W_c(\mathcal{I})$. *Core validation* cycles through N matchings over $\mathcal{I} \times A_c$, sampling every player against every baseline arm. An *outside probe* moves only player i to $a \notin A_c$, keeping every other player at her baseline partner; when W_c is stable, only the moved player can incur positive regret. A persistent cursor selects an action whose cumulative target is unmet and returns Complete only when all targets are met (Appendix E.4; line 12).

The target sizes mirror the structural scale h. Each core comparison needs $O(\log(eT)/h^2)$ observations and may incur constant player loss. For an outside edge, let $s_{i,a} := (\beta_i - \mu^P_{i,a})_+$ be its player-side regret shortfall. It needs $O(\log(eT)/\max\{h, s_{i,a}\}^2)$ probes, each costing $s_{i,a}$. The inequality $s_{i,a}/\max\{h, s_{i,a}\}^2 \leq 1/h$ yields the $N(K-N)/h$ tail charge, while the core contributes N^2/h^2.

Two target-setting modes. The cyclic pilot samples every edge $m_0 = \lceil \log(eT) \rceil$ times and freezes clipped, quantized means $\bar{\mu}$ with accuracy radius ε (line 1). Let $\bar{g}(W)$ denote the certificate on these means and $\bar{\Delta} = \max_W \bar{g}(W)$. If $4\varepsilon \geq 1$, no empirical margin can pass the branch gate, so the exact maximum-certificate search is skipped.

If the search runs and returns $\bar{\Delta} > 4\varepsilon$, ACTL enters CERTIFY (line 3): W_c is the empirical maximum-certificate matching, $\bar{\Delta}$ sets the resolution scale, and fixed outside-edge targets reflect their pilot shortfalls relative to W_c. Otherwise it enters PRESERVE (lines 5–6): W_c is the tie-broken empirical player-proposing Gale–Shapley (GS) matching (Gale & Shapley, 1962), and the running estimates begin at $\widehat{\beta}_i = \bar{\mu}^P_{i,W_c(i)}$ and $\widehat{h} = 1$.

In PRESERVE, ACTL evaluates one candidate matching Z on the frozen pilot matrix per uncommitted round; generating Z does not require playing it (line 10). The iterator stores only Z and advances it in $O(NK)$ time; it never stores the matching space (Proposition 3). A candidate can lower either the payoff estimates $\widehat{\beta}_i$ or the scale estimate $\widehat{h}$; both updates only enlarge the adaptive targets, and sample counts are never reset. The pilot selects a target-setting mode without having to identify the true regime. Appendix E gives the exact search and targets.

Fallback and certificate tests. When targets are complete, ACTL replays the current UCB–GS cache (line 17); later PRESERVE updates may enlarge targets, so validation resumes without resetting counts. When $\Delta = 0$, positive commitment is impossible, but sufficiently resolved threshold comparisons make each newly constructed cache truly weakly stable. Thus the fallback permits stable exploitation without commitment (Appendix E.7).

A horizon-wide certificate test commits a passing matching for the remaining rounds and leaves the state unchanged on failure. ACTL tests the baseline and a temporary UCB–GS candidate at new snapshots, the baseline when fixed CERTIFY targets complete, and every new fallback cache (lines 11, 16, and 17). Core–tail targets govern sample allocation; certificate tests provide a separate early-stopping step (Appendix E.7).

Algorithm 1 Aggregate Core–Tail Learning (ACTL)

Require: N, K, T; public labels and deterministic tie-breaking
1: **Pilot:** sample every edge $m_0 = \lceil \log(eT) \rceil$ times; freeze the quantized pilot matrix; compute ε; if $4\varepsilon < 1$, run the maximum-certificate search for $(\bar{W}, \bar{\Delta})$.
2: **if** $4\varepsilon < 1$ **and** $\bar{\Delta} > 4\varepsilon$ **then**
3: mode $\leftarrow$ CERTIFY; $W_c \leftarrow$ empirical maximum-certificate matching; fix targets (19).
4: **else**
5: mode $\leftarrow$ PRESERVE; $W_c \leftarrow$ tie-broken empirical player-proposing GS matching; $\widehat{\beta}_i \leftarrow \bar{\mu}^P_{i,W_c(i)}$; $\widehat{h} \leftarrow 1$.
6: Initialize the candidate stream and adaptive targets (20).
7: **end if**
8: Initialize the validation cursor, an empty fallback cache, and the certificate tests.
9: **for** each remaining reward round t **do**
10: If PRESERVE: $Z \leftarrow$ next candidate; UPDATE-PRESERVATION(Z).
11: On the first post-pilot round and each new snapshot version: TRY-COMMIT(W_c); compute a temporary UCB–GS matching W_{UCB}; TRY-COMMIT(W_{UCB}).
12: $a \leftarrow$NEXT-VALIDATION-ACTION.
13: **if** $a \neq$COMPLETE **then**
14: Play a.
15: **else**
16: If CERTIFY and its fixed targets just completed, TRY-COMMIT(W_c).
17: Refresh the cache if empty or its snapshot version or dyadic index changed; on refresh, TRY-COMMIT(M_{cache}); replay M_{cache} unless committed.
18: **end if**
19: Update rewards and counts; refresh quantized snapshots when a count doubles.
20: **end for**
Every stage stops at T; if the horizon ends during the pilot, no later stage begins.

5.2 Regret Guarantees

The same structural rate governs both branches; only the relevant margin changes.

Theorem 3 (Unified upper bounds)**.** *There is a universal constant $C > 0$ such that ACTL (Section 5.1), which knows only N, K, T, satisfies*

$$\mathbb{E}R_P^{\Sigma}(T) \le C\Psi(h_\star)\log(eT) + C\mathcal{L}_\star, \qquad \limsup_{T\to\infty} \frac{\mathbb{E}R_P^{\Sigma}(T)}{\log T} \le C\Psi(h_\star), \tag{3}$$

where

$$h_\star := \begin{cases} \Delta, & \Delta > 0, \\ \Delta_{\mathrm{pres}}, & \Delta = 0, \end{cases} \qquad \mathcal{L}_\star := \begin{cases} N^2K\left(\dfrac{1}{\Delta^4} + \dfrac{1}{\Delta_{\mathrm{pres}}^2\Delta^2}\right)\log\dfrac{eNK}{\Delta}, & \Delta > 0, \\ \left(\dfrac{N^3K}{\Delta_{\mathrm{pres}}^2} + \dfrac{N^2K}{\Delta_{\mathrm{pres}}^4}\right)\log\dfrac{eNK}{\Delta_{\mathrm{pres}}}, & \Delta = 0. \end{cases}$$

Both the leading logarithmic term and $\mathcal{L}_\star$ are free of $\Delta_{\min}$. With $B_T = \log(8NK(T+1)^5)$, every participant also satisfies

$$\max\Big\{\max_i \mathbb{E}R_i(T), \max_a \mathbb{E}R_a^A(T)\Big\} \le \frac{CNKB_T}{\delta_{\mathrm{safe}}^2}, \qquad \delta_{\mathrm{safe}} := \begin{cases} \Delta_{\mathrm{pres}}, & \Delta = 0, \\ \min\{\Delta, \Delta_{\mathrm{pres}}\}, & \Delta > 0, \end{cases} \tag{4}$$

The term $\Psi(h_\star)\log T$ is the class-wise sharp contribution. The horizon-independent $\mathcal{L}_\star$ collects unresolved-pilot, pre-discovery fallback, and confidence-failure costs.

Proof sketch. The main difficulty is the $\Delta = 0$ case, where no stable matching has a positive certificate and the relevant scale $h = \Delta_{\mathrm{pres}}$ is unknown. On the good pilot event, PRESERVE keeps $\widehat{h}$ above a constant fraction of h, while its baseline estimates keep each adaptive outside-edge target within a constant factor of the ideal $\max\{h, s_{i,a}\}^{-2}$ target. Consequently, validation incurs $O(\Psi(h)B_T)$ regret even before the matchings that determine the true margins have been examined. Later updates of $\widehat{h}$ or $\widehat{\beta}$ only enlarge cumulative targets; edge counts are retained, so sampling never restarts.

The matchings needed to determine the true margins appear by the end of one pass through the injections, but the iterator evaluates only one candidate per round and uses polynomial work for that step. Generating a candidate does not require playing it. The regret analysis uses the discovery index τ only through $\log(1+\tau) = O(N\log(eK))$. Before discovery, each confidence-good unstable UCB–GS cache is charged to a played edge still uncertain at scale h; replay samples that edge until a snapshot doubling invalidates the cache. This charging places the pre-discovery cost in the horizon-independent remainder.

Once the required witness matchings (at most $N+1$: one attaining each β_i and one attaining $-\Delta_{\mathrm{pres}}$) have appeared, the running estimates become two-sided accurate. Target completion then screens unresolved outside edges below the stable payoff baseline while preserving every threshold-h comparison on the remaining edges. By Proposition 1, these comparisons admit a truly weakly stable reference matching. That reference is used in the analysis: ACTL plays the UCB–GS cache, and proposer optimality together with the preserved comparisons rules out any blocking pair in each newly established cache. Stable exploitation therefore requires neither finding the reference nor recovering the full preference profile. In CERTIFY, by contrast, fixed targets give $O(\Psi(\Delta)B_T)$ validation regret and certify the baseline once the pilot resolves $\Delta > 0$. Confidence failures and unresolved pilots enter $\mathcal{L}_\star$; Appendix F gives the full proof.

5.3 Matching the Lower Bound

To compare the upper and lower bounds over margin classes, let $\Pi_{\log}$ denote the policies whose expected regret for every participant (players and arms) is $O_\nu(\log(eT))$ on every fixed market ν, with finite instance-dependent coefficients independent of T; they know N, K, T, but not the true margins or tie pattern (formal policy-class discussion in Appendix H). For $s \in \{\mathrm{str}, \mathrm{tie}\}$, let $\mathcal{C}^s(h)$ contain the size-N, K markets with $\Delta \ge h$ that are strict or contain exact ties, respectively. Recall the class $\mathcal{C}_{\mathrm{pres}}(h)$ of Section 4.2. Write $R_P^{\Sigma,\pi}(T)$ for the aggregate player regret induced by policy π.

With $K \geq N$, Theorems 1–3 give the first equality below for $N \geq 3$ and $0 < h \leq 1/128$, and the second for $N \geq 5$ and $0 < h \leq 1/64$:

$$\inf_{\pi \in \Pi_{\log}} \sup_{\nu \in \mathcal{C}^s(h)} \limsup_{T \to \infty} \frac{\mathbb{E}_\nu R_P^{\Sigma,\pi}(T)}{\log T} = \Theta(\Psi(h)), \quad \inf_{\pi \in \Pi_{\log}} \sup_{\nu \in \mathcal{C}_{\mathrm{pres}}(h)} \limsup_{T \to \infty} \frac{\mathbb{E}_\nu R_P^{\Sigma,\pi}(T)}{\log T} = \Theta(\Psi(h)). \tag{5}$$

The first equality covers $\Delta > 0$ markets, with either strict preferences or exact ties; the second covers $\Delta = 0$ markets. In each class, the lower bound identifies difficult markets, while the same ACTL policy gives the upper bound throughout the class. Thus (5) characterizes the minimax logarithmic cost of these margin classes.

For the $\Delta = 0$ class the same rate holds with the supremum inside the horizon limit, since the remainder is uniformly bounded (Appendix H). The same class-wise order $\Theta(\Psi(h) \log T)$ is attainable over every fixed connected local graph covered by Theorem 4. The statistical lower bound carries over: messages and routing use past observations and protocol randomness, so they add no reward KL divergence.

5.4 Communication-Aware Decentralized Learning

In LC-ACTL, each player and arm keeps its own reward observations and shares rounded summaries needed to coordinate later decisions. LC-ACTL runs a quantized version of ACTL, denoted ACTL^q in Appendix G.5. It retains ACTL's search, two modes, validation, and fallback, with a different quantizer and validation constant. Communication uses physical slots, and their rewards count in regret. The appendix proves the statistical guarantee for this version.

Theorem 4 (Communication-aware decentralized regret)**.** *Assume cooperative participants with public labels and reliable synchronized communication.* Local network: *on any fixed connected network with one-bit neighbor links where matching continues during communication, LC-ACTL satisfies*

$$\limsup_{T \to \infty} \frac{\mathbb{E} R_P^{\Sigma}(T)}{\log T} \leq C \Psi(h_\star). \tag{6}$$

In-band: *without an auxiliary channel, signaling through proposals and acceptances adds* $O((N + K)NK \log T)$ *regret. Here* $h_\star$ *is defined in Theorem 3,* C *is universal, and the finite-time remainders are independent of* $\Delta_{\min}$.

The local-network result uses communication alongside ordinary matching. After a sufficiently accurate pilot resolves a truly weakly stable baseline, participants can keep playing it while messages travel, so those slots add no regret. If the pilot remains unresolved or its estimates are inaccurate, the resulting communication loss enters the horizon-independent remainder. In-band messages instead use proposals and acceptances, which can disrupt baseline play and yield the additional logarithmic term. Appendix G proves both cases through Theorem 5. The statistical core–tail term is minimax-sharp; optimizing the additional in-band signaling term remains an orthogonal communication problem.

6 Conclusion

Stable learning is governed by stability-relevant comparisons. The certificate margin Δ governs positive certification; the preservation margin Δ_{pres} governs learning when $\Delta = 0$. One adaptive policy attains the sharp class-wise core–tail law $\Psi(h) \log T$ in both cases without depending on the global $\Delta_{\min}$. Appendix B tests the core–tail scaling, $\Delta_{\min}$ independence, and zero-certificate learning.

The same order is achievable over connected local networks with communication charged to physical time, and logarithmic regret remains achievable through in-band signaling without an auxiliary channel. In Preserve, ACTL traverses the matching space one candidate at a time, with polynomial work per interaction round and only logarithmic dependence on the witness-discovery index in regret. The decentralized result assumes cooperative agents, synchronized slots, reliable local control, and public identities.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing, literature and citation checking, proof checking, and experimental design. AI-generated suggestions were independently reviewed and verified by the authors. All mathematical claims, proofs, algorithms, experiments, citations, and conclusions are the responsibility of the authors.

REPRODUCIBILITY STATEMENT

Complete lower-bound constructions and proofs are provided in Appendices C–D, the executable specification and computational complexity of ACTL in Appendix E, the full upper-bound proofs in Appendix F, and the decentralized protocol and proofs in Appendix G. Appendix B reports the experimental protocol, calibration procedure, seeds, statistical fitting procedure, and all structural settings used in the empirical evaluation.

REFERENCES

Atila Abdulkadiroğlu and Tayfun Sönmez. School choice: A mechanism design approach. *American economic review*, 93(3):729–747, 2003.

Atila Abdulkadiroğlu, Parag A Pathak, and Alvin E Roth. Strategy-proofness versus efficiency in matching with indifferences: Redesigning the nyc high school match. *American Economic Review*, 99(5):1954–1978, 2009.

Andreas Athanasopoulos, Anne-Marie George, and Christos Dimitrakakis. Probably correct optimal stable matching for two-sided market under uncertainty. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, pp. 152–160, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269.

Andreas Athanasopoulos, Anne-Marie George, and Christos Dimitrakakis. Probably correct optimal stable matching under two-sided uncertainty. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 228–254. PMLR, 2026. URL https://arxiv.org/abs/2607.04824.

Soumya Basu. Competing bandits in matching markets via super stability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3226–3250. PMLR, 2025. URL https://proceedings.mlr.press/v267/basu25a.html.

Soumya Basu, Karthik Abinav Sankararaman, and Abishek Sankararaman. Beyond $\log^2$ (t) regret for decentralized bandits in matching markets. In *International Conference on Machine Learning*, pp. 705–715. PMLR, 2021.

Sarah H Cen and Devavrat Shah. Regret, stability & fairness in matching markets with bandit learners. In *International Conference on Artificial Intelligence and Statistics*, pp. 8938–8968. PMLR, 2022.

Ágnes Cseh and Attila Juhos. Pairwise preferences in the stable marriage problem. *ACM Transactions on Economics and Computation*, 9(1), 2021. doi: 10.1145/3434427. URL https://hpi.de/friedrich/docs/publications/2021/TEAC.pdf.

Xiaowu Dai and Michael Jordan. Learning in multi-stage decentralized matching markets. *Advances in Neural Information Processing Systems*, 34:12798–12809, 2021a.

Xiaowu Dai and Michael I Jordan. Learning strategies in decentralized matching markets under uncertain preferences. *The Journal of Machine Learning Research*, 22(1):11806–11855, 2021b.

Sanmay Das and Emir Kamenica. Two-sided bandits and the dating market. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pp. 947–952, 2005.

Aytek Erdil and Haluk Ergin. What's the matter with tie-breaking? improving efficiency in school choice. *American Economic Review*, 98(3):669–689, 2008.

Aytek Erdil and Haluk Ergin. Two-sided matching with indifferences. *Journal of Economic Theory*, 171:268–292, 2017. ISSN 0022-0531. doi: https://doi.org/10.1016/j.jet.2017.07.002. URL https://www.sciencedirect.com/science/article/pii/S0022053117300789.

David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

Mirco Gelain, Maria Silvia Pini, Francesca Rossi, Kristen Brent Venable, and Toby Walsh. Local search for stable marriage problems with ties and incomplete lists. In *Pacific Rim International Conference on Artificial Intelligence*, pp. 64–75. Springer, 2010.

Ian P Gent, Robert W Irving, David F Manlove, Patrick Prosser, and Barbara M Smith. A constraint programming approach to the stable marriage problem. In *International Conference on Principles and Practice of Constraint Programming*, pp. 225–239. Springer, 2001.

Ian P Gent, Patrick Prosser, Barbara Smith, and Toby Walsh. Sat encodings of the stable marriage problem with ties and incomplete lists. *SAT*, 2002:133–140, 2002.

Avishek Ghosh, Abishek Sankararaman, Kannan Ramchandran, Tara Javidi, and Arya Mazumdar. Decentralized competing bandits in non-stationary matching markets. *arXiv preprint arXiv:2206.00120*, 2022.

Dan Gusfield and Robert W. Irving. *The Stable Marriage Problem: Structure and Algorithms*. MIT Press, Cambridge, MA, 1989. ISBN 9780262071185.

Magnús M Halldórsson, Robert W Irving, Kazuo Iwama, David F Manlove, Shuichi Miyazaki, Yasufumi Morita, and Sandy Scott. Approximability results for stable marriage problems with ties. *Theoretical Computer Science*, 306(1-3):431–447, 2003.

Magnús M Halldórsson, Kazuo Iwama, Shuichi Miyazaki, and Hiroki Yanagisawa. Improved approximation results for the stable marriage problem. *ACM Transactions on Algorithms (TALG)*, 3(3):30–es, 2007.

Hadi Hosseini, Sanjukta Roy, and Duohan Zhang. Putting gale & shapley to work: Guaranteeing stability through learning. *arXiv preprint arXiv:2410.04376*, 2024.

Robert W. Irving. Stable marriage and indifference. *Discrete Applied Mathematics*, 48(3):261–272, 1994. ISSN 0166-218X. doi: https://doi.org/10.1016/0166-218X(92)00179-P. URL https://www.sciencedirect.com/science/article/pii/0166218X9200179P.

Robert W Irving and David F Manlove. An 8/5-approximation algorithm for a hard variant of stable marriage. In *International Computing and Combinatorics Conference*, pp. 548–558. Springer, 2007.

Robert W Irving, David F Manlove, and Sandy Scott. Strong stability in the hospitals/residents problem. In *Annual Symposium on Theoretical Aspects of Computer Science*, pp. 439–450. Springer, 2003.

Kazuo Iwama and Shuichi Miyazaki. A survey of the stable marriage problem and its variants. In *International conference on informatics education and research for knowledge-circulating society (ICKS 2008)*, pp. 131–136. IEEE, 2008.

Kazuo Iwama, David Manlove, Shuichi Miyazaki, and Yasufumi Morita. Stable marriage with incomplete lists and ties. In Jirí Wiedermann, Peter van Emde Boas, and Mogens Nielsen (eds.), *Automata, Languages and Programming*, pp. 443–452, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg. ISBN 978-3-540-48523-0.

Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael Jordan, and Jacob Steinhardt. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34:3323–3335, 2021.

Zoltán Király. Better and simpler approximation algorithms for the stable marriage problem. *Algorithmica*, 60(1):3–20, 2011.

Fang Kong and Shuai Li. Player-optimal stable regret for bandit learning in matching markets. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1512–1522. SIAM, 2023.

Fang Kong and Shuai Li. Improved bandits in many-to-one matching markets with incentive compatibility. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:13256–13264, 03 2024. doi: 10.1609/aaai.v38i12.29226.

Fang Kong, Junming Yin, and Shuai Li. Thompson sampling for bandit learning in matching markets. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 3164–3170. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/439. URL https://doi.org/10.24963/ijcai.2022/439. Main Track.

Fang Kong, Zilong Wang, and Shuai Li. Improved analysis for bandit learning in matching markets. *Advances in Neural Information Processing Systems*, 2024.

Fang Kong, Jingqi Tang, Mingzhu Li, Pinyan Lu, John C.S. Lui, and Shuai Li. Bandit learning in matching markets with indifference. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7ENakslm9J.

Shuai Li, Zilong Wang, and Fang Kong. A survey on bandit learning in matching markets. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 10546–10554. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/1171. URL https://doi.org/10.24963/ijcai.2025/1171. Survey Track.

Yuantong Li, Chi-hua Wang, Guang Cheng, and Will Wei Sun. Rate-optimal contextual online matching bandit. *arXiv preprint arXiv:2205.03699*, 2022.

Yuantong Li, Guang Cheng, and Xiaowu Dai. Two-sided competing matching recommendation markets with quota and complementary preferences constraints. In *Forty-first International Conference on Machine Learning*, 2024.

Shiyun Lin, Simon Mauras, Nadav Merlis, and Vianney Perchet. Stable matching with ties: Approximation ratios and learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=HPmDgKOlM3.

Lydia T Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, pp. 1618–1628. PMLR, 2020.

Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *J. Mach. Learn. Res.*, 22:211–1, 2021.

Chinmay Maheshwari, Shankar Sastry, and Eric Mazumdar. Decentralized, communication-and coordination-free learning in structured matching markets. In *Advances in Neural Information Processing Systems*, 2022.

David Manlove. *Algorithmics of matching under preferences*, volume 2. World Scientific, 2013.

David F Manlove. Stable marriage with ties and unacceptable partners. Technical report, Technical Report TR-1999-29, University of Glasgow, Department of Computing . . . , 1999.

David F. Manlove. The structure of stable marriage with indifference. *Discrete Applied Mathematics*, 122(1–3):167–181, 2002. doi: 10.1016/S0166-218X(01)00322-5.

David F Manlove, Robert W Irving, Kazuo Iwama, Shuichi Miyazaki, and Yasufumi Morita. Hard variants of stable marriage. *Theoretical Computer Science*, 276(1):261–279, 2002. ISSN 0304-3975. doi: https://doi.org/10.1016/S0304-3975(01)00206-7. URL https://www.sciencedirect.com/science/article/pii/S0304397501002067.

Dániel Marx and Ildikó Schlotter. Parameterized complexity and local search approaches for the stable marriage problem with ties. *Algorithmica*, 58(1):170–187, 2010.

Yifei Min, Tianhao Wang, Ruitu Xu, Zhaoran Wang, Michael I Jordan, and Zhuoran Yang. Learn to match with no regret: Reinforcement learning in markov matching markets. *arXiv preprint arXiv:2203.03684*, 2022.

Deepan Muthirayan, Chinmay Maheshwari, Pramod P Khargonekar, and Shankar Sastry. Competing bandits in time varying matching markets. *arXiv preprint arXiv:2210.11692*, 2022.

Tejas Pagare and Avishek Ghosh. Explore-then-commit algorithms for decentralized two-sided matching markets. *arXiv preprint arXiv:2408.08690*, 2024.

Gaurab Pokharel and Sanmay Das. Converging to stability in two-sided bandits: The case of unknown preferences on both sides of a matching market. *arXiv preprint arXiv:2302.06176*, 2023. URL https://arxiv.org/abs/2302.06176.

Baharak Rastegari, Anne Condon, Nicole Immorlica, Robert W. Irving, and Kevin Leyton-Brown. Reasoning about optimal stable matchings under partial information. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, pp. 431–448. ACM, 2014. doi: 10.1145/2600057.2602884.

Alvin E Roth and Marilda Sotomayor. Two-sided matching. *Handbook of game theory with economic applications*, 1:485–541, 1992.

Alvin E. Roth and Marilda A. Oliveira Sotomayor. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, volume 18 of *Econometric Society Monographs*. Cambridge University Press, Cambridge, 1990. ISBN 9780521390156.

Abishek Sankararaman, Soumya Basu, and Karthik Abinav Sankararaman. Dominate or delete: Decentralized competing bandits in serial dictatorship. In *International Conference on Artificial Intelligence and Statistics*, pp. 1252–1260. PMLR, 2021.

Yi Su, Magd Bayoumi, and Thorsten Joachims. Optimizing rankings for recommendation in matching markets. In *Proceedings of the ACM Web Conference 2022*, pp. 328–338, 2022.

Zilong Wang, Liya Guo, Junming Yin, and Shuai Li. Bandit learning in many-to-one matching markets. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 2088–2097, 2022.

Yirui Zhang and Zhixuan Fang. Decentralized competing bandits in many-to-one matching markets. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 2603–2605, 2024a.

Yirui Zhang and Zhixuan Fang. Decentralized two-sided bandit learning in matching market. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, UAI '24. JMLR.org, 2024b.

Yirui Zhang, Siwei Wang, and Zhixuan Fang. Matching in multi-arm bandit with collision. In *Advances in Neural Information Processing Systems*, 2022.

A NOTATION

Regrets are random positive-part sums; performance bounds take expectations. The last two groups collect notation used only in the decentralized extension and lower-bound constructions.

Table 2: Main notation.

Symbol	Meaning
Market and regret	
$\mathcal{I}, \mathcal{A}; N, K, T$	Players, arms, their sizes, and the reward-round horizon of the central model.
$M(i), M^{-1}(a), \perp$	Assigned partners and the unmatched-arm convention.
$\mu^P_{i,a}, \mu^A_{a,i}$	Player-side and arm-side mean rewards.
$\mathcal{M}, \mathcal{W}^\star$	Feasible injections and weakly stable matchings.
β_i, α_a	Lowest weakly stable player and arm rewards.
$R_i, R^A_a, R^\Sigma_P, R^{\Sigma,\pi}_P$	Individual player, individual arm, aggregate player regret, and aggregate regret under policy π.
$R^{\mathrm{pp}}_i, R^{\mathrm{ind}}_i, R^{\mathrm{ap}}_a$	Strict player-pessimal, indifferent, and strict arm-pessimal regrets.
$M^{\mathrm{po}}, R^{\mathrm{po}}_P, R^{\mathrm{ap}}_A$	Player-optimal stable matching, total player-optimal regret, and total arm-pessimal regret in the benchmark barrier.
Structural margins	
$\Delta_{\min}$	Smallest positive real-partner preference gap, or 1 if none exists.
$\Delta_{\mathcal{A}}$	Admissible gap of Basu (2025); equals Δ in strict markets.
Δ_F	Pervasive-profile gap of Athanasopoulos et al. (2026); $\Delta_F \le \Delta_{\mathcal{A}}$.
$c_W, g(W), \Delta$	Pairwise rejection certificate, signed matching certificate $g(W)$, and maximum certificate margin.
Δ_{pres}	Distance from the signed boundary to the nearest unstable matching: $\min(\{1\} \cup \{-g(W) : g(W) < 0\})$.
δ_{safe}	Individual-safety scale: Δ_{pres} if $\Delta = 0$, else $\min\{\Delta, \Delta_{\mathrm{pres}}\}$.
$\pi_h, \mathcal{W}_h$	Threshold profile at margin h and its possibly stable matchings $\{W : g(W) > -h\}$.
$d, \Psi(h)$	Lower-bound construction scale and core–tail rate function.
ACTL statistics and decisions	
$m_0, \varepsilon, \eta_T, Q_T, H_{\mathrm{val}}$	Central ACTL pilot samples, accuracy radius, quantization step and quantizer, and validation-round envelope.
$\bar{\mu}, \widehat{\mu}$	Frozen clipped and quantized pilot means, and raw all-history empirical means.
$\bar{g}(W), \bar{\Delta}$	Frozen-pilot signed certificate and its maximum.
b_t, B_T, r_t, r_*	Anytime and horizon-wide confidence clocks and radii.
$a_e(t), a^*_e, U_e, L^*_e, U^*_e$	Snapshot radii, public UCB, and commitment intervals.
$\bar{W}, W_c, A_c$	Empirical maximum-certificate matching when $\bar{\Delta} > 0$; fixed validation baseline and its assigned-arm set.
$\mathcal{H}_T, \mathcal{G}_t$	Horizon-wide and establishment-time snapshot confidence events.
$\bar{d}_{i,a}$	Frozen-pilot outside baseline reward difference in the certificate branch.
$\widehat{\beta}_i, \widehat{h}, \widehat{\delta}_{i,a}$	Running estimates of lowest weakly stable player payoffs, preservation-scale estimate, and outside-edge losses.
$q_0, q_{i,a}$	Cumulative core and outside sample targets.
$n_{i,a}(t), L, U$	Edge count and confidence endpoints.
C_s	Core validation matchings of the fixed baseline.
$Z, M_{\mathrm{cache}}, \kappa_s$	Structural candidate injection, cached fallback matching, and its dyadic index.
$h_\star, \mathcal{L}_\star$	Unified scale (Δ if positive, else Δ_{pres}) and finite-time remainder.
Communication-only notation	
t, s, Q_t	Physical slot, retained learning index, and realized assignment; Q_t may be partial during in-band signaling.

Table 2: Main notation (continued).

Symbol	Meaning
ℓ, γ, L_0, A_0	LC-ACTL log horizon, pilot concentration clock, pilot bits per endpoint, and validation constant.
n_e, m_e, v_e, ζ_e	Retained edge count, snapshot count, quantized endpoint value, and grid-error allowance.
$F(u), B(T), C_{\mathrm{post}}$	Resolution remainder $u^{-4}\log^2(eNK/u)$, logical source-bit count, and post-pilot communication slots.
$(\lambda, \kappa), D, G$	Source-bit latency, per-slot baseline disruption, graph diameter bound, and local network.
Policy classes and lower-bound constructions	
$\Pi_{\log}, \mathcal{C}^s(h), \mathcal{C}_{\mathrm{pres}}(h)$	Policies with logarithmic participant regret on each fixed market; regime-s markets with $\Delta \geq h$; zero-certificate markets with $\Delta_{\mathrm{pres}} \geq h$.
$i_0, u_i, v_j; x, b_i, a_j$	Designated player, rows, guards; shared arm, homes, dangerous arms.

In a strict market $R_i^{\mathrm{pp}} = R_i$ and $R_a^{\mathrm{ap}} = R_a^A$; with ties $R_i^{\mathrm{ind}} = R_i$, while $\beta_i^{\mathrm{pp}} = \beta_i^{\mathrm{ind}} = \beta_i$ and $\alpha_a^{\mathrm{ap}} = \alpha_a$.

B Experimental Validation of the Structural Theory

B.1 Experimental Protocol

The experiments empirically validate the structural predictions of the theory on the synthetic families of Appendix C—the strict core family, the strict-tagged tail family (Eq. (9)), the decoupled-$\Delta_{\min}$ family, and the zero-certificate module—together with random strict markets. The positive-certificate experiments use strict versions of the core and tail hard families; exact ties are retained only for the zero-certificate experiments where they are intrinsic to the regime.

The main metric is the aggregate positive-part player regret $R_P^{\Sigma}(T)$ of Section 3; we additionally report the fraction of rounds played at a weakly stable matching. All structural quantities (Δ, Δ_{pres}, $\Delta_{\min}$, and uniqueness of stable matchings) are verified exactly by exhaustive enumeration on every instance for which exact enumeration is computationally tractable, and the exact certificate margin Δ of every positive-certificate instance is recorded in the raw results. Calibration uses validation seeds 0–2 only, and every reported result uses disjoint test seeds $1000+$; no test-seed tuning is performed.

Scaling exponents are fitted by ordinary least squares to the window means in log–log coordinates, and R^2 is the coefficient of determination of that regression. The reported 95% confidence interval is a seed-resampling bootstrap (10^4 resamples of the test seeds within each window point, refitting the exponent), so it propagates the seed-level variability of the means rather than only the residual scatter of the few window points. Error bars on experimental means are ± 1.96 standard errors of the mean over test seeds. Window-edge finite-horizon effects (pilot coverage at large d, horizon censoring at small d) bend the mean curves and are discussed alongside each fit.

For finite-horizon evaluation, we use ACTL-Practical, a calibrated implementation that preserves ACTL's structural algorithm, certificate test, and adaptive allocation mechanism while adjusting only numerical calibration constants. The common settings, selected once on validation seeds and used unchanged for every test run, are $m_0 = \lceil 4\ell \rceil$ pilot samples per edge with $\ell = \log(eT)$ (rather than $\lceil \ell \rceil$), target constants 0.5 (Certify) and 0.0625 (Preserve) in place of 1024 in Section E.5, a 0.05 scale on the pilot radius ε used in the 4ε branch gate and candidate classification, and a 0.2 scale on the fallback UCB–GS bonus. The commitment test always keeps the specification radius.

Table 3 lists the finite-horizon calibration settings used to isolate the corresponding structural effects. Target constants are shown in Certify/Preserve order; "gate free" leaves the branch decision to the pilot.

Table 3: ACTL-Practical finite-horizon calibration settings.

Evaluation	Pilot factor	ε scale	Targets	Branch
Core, positive	48	0.02	0.25/0.2	CERTIFY
Core, preservation	48	0.02	1.0/0.2	PRESERVE
Tail scaling	96	0.02	0.2/0.2	PRESERVE
Tail scaling, longer horizon	96	0.02	0.5/0.5	PRESERVE
Dimensional scaling	24	0.02	0.2/0.2	gate free

The uniform ablation ACTL-NoTailAdapt replaces the per-edge shortfall $\widehat{\delta}_{i,a}$ in the PRESERVE targets by $\widehat{h}$. We implement the two published methods, Basu-SS (Basu, 2025) and E-PSM/EE-PSM (Athanasopoulos et al., 2026), directly from their published specifications and validate the implementations against brute-force structural checks on small markets. Existing package tests cover the centralized implementation invariants. The communication-aware extension is analyzed theoretically in Appendix G; the experiments here focus on validating the statistical core–tail mechanisms.

B.2 INDEPENDENCE FROM THE GLOBAL MINIMUM PREFERENCE GAP

The first experiment isolates the dependence on the global minimum positive preference gap at a fixed certificate margin. The decoupled strict family holds $\Delta = 0.1000$ fixed while $\Delta_{\min}$ shrinks by a factor of 77 across seven instances ($N = K = 5$, $T = 2^{17}$, 30 test seeds); enumeration confirms a Δ spread of 0.00% over the sweep. Figure 2 plots mean regret against $1/\Delta_{\min}$. ACTL-Practical changes by a factor of only 0.998 across the entire sweep ($1{,}788 \to 1{,}785$). The full-profile confidence-elimination control (Full-Profile CE), which must resolve the complete preference ranking before committing, remains at its no-resolution level (199,339–200,375) at every sweep point, including the easiest one: the irrelevant fine comparisons are too small to be separated within the horizon, so the control never certifies. The experiment isolates the distinction between structural and global preference gaps: holding Δ fixed while shrinking $\Delta_{\min}$ has essentially no effect on ACTL-Practical, whereas full-profile recovery remains bottlenecked by the irrelevant fine comparisons.

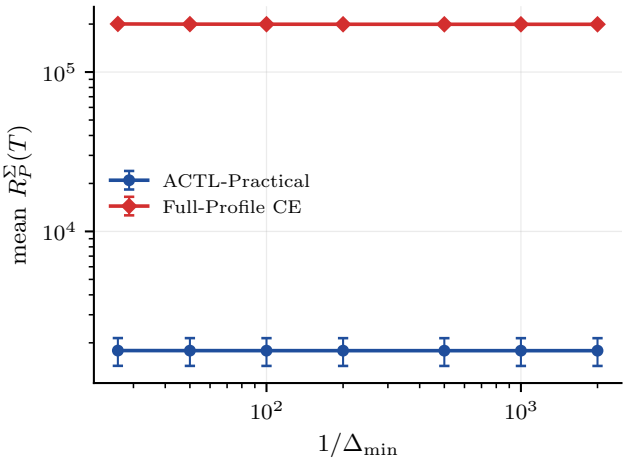

Figure 2: Independence from the global minimum preference gap. At fixed $\Delta = 0.1000$, ACTL-Practical is essentially invariant over the $77\times$ sweep of $\Delta_{\min}$, while the full-profile control remains at its no-resolution level at every point. Means with 95% confidence intervals over 30 test seeds ($N = K = 5$, $T = 2^{17}$).

B.3 Empirical Core–Tail Scaling

The core–tail law $\Psi(h) = N^2/h^2 + N(K-N)/h$ predicts two channels: resolving the N^2 core comparisons requires h^{-2}-scaled samples at constant loss per observation, whereas the shortfall-adaptive screening of the $N(K-N)$ outside edges carries an h^{-1}-type charge. Figure 3 measures both. The core panel uses the strict core family ($N = K = 5$, $T = 2^{17}$, 30 test seeds) and plots $R_{\mathrm{core}}/\log(eT)$ against h for branch-controlled evaluations—which force the corresponding branch so that its scaling is measured without regime mixing—and for an oracle-certificate benchmark that knows the true margins. The tail panel uses the strict-tagged tail family (Eq. (9); $N = 4$, $K = 32$, $T = 2^{20}$, 30 test seeds) and plots $R_{\mathrm{tail}}/[N(K-N)\log(eT)]$ against the exact certificate margin $h = \Delta$ of each instance—not the construction parameter d—for the shortfall-adaptive allocation and for the uniform-target ablation ACTL-NoTailAdapt.

On the uncensored core window $d \geq 0.075$, the fitted exponents are -2.22 (95% CI $[-3.70, -1.23]$, $R^2 = 0.99$) for the positive evaluation and -2.29 ($[-3.42, -1.31]$, $R^2 = 0.96$) for the preservation evaluation, closely tracking the predicted h^{-2} dependence; the oracle benchmark attains the predicted -2.00 ($R^2 = 1.00$). At smaller d, finite-horizon censoring prevents validation from completing within T; we therefore estimate the scaling exponent only on the uncensored window.

On the tail panel, the adaptive curve tracks the predicted h^{-1} charge ($\beta = -1.09$, 95% CI $[-1.76, -0.46]$, $R^2 = 0.94$; normalized values $62.9 \to 54.1 \to 40.2$ over $h = \Delta = 0.0492, 0.0616, 0.0739$). The uniform ablation is substantially flatter ($\beta = -0.30$, $[-0.52, -0.08]$, $R^2 = 0.95$) and lies above the adaptive curve at every grid point, the gap widening from $1.4\times$ to $3.9\times$. A second run of the same evaluations at a wider target constant reproduces the slope (-1.27, $[-1.89, -0.65]$) and the contrast (versus -0.20); the pooled common-slope estimate across both runs (one intercept per target constant) is $\beta = -1.22$ (95% CI $[-1.72, -0.74]$, $R^2 = 0.92$).

To isolate the predicted scaling, we fit the tail exponent on the finite-horizon window where neither pilot dominance nor horizon censoring is substantial. The fit uses $d \in \{0.05, 0.0625, 0.075\}$, where the pilot supplies 16–22% of an edge's samples. The larger points $d = 0.0875$ and $d = 0.10$ are not included in the fit because their pilot shares rise to 30% and 40%, respectively, and already depress the measured charge. On the multi-level tail family ($d = 0.05$, shortfalls $d, 2d, 4d, 8d$), the uniform target incurs $8.2\times$ the tail regret of the adaptive allocation (249,868 versus 30,521) with $3.3\times$ more probes, and weakly stable play collapses from 0.59 to 0.005; this isolates the shortfall-adaptive allocation as the source of the h^{-1} tail charge. Together the two panels exhibit the two terms of $\Psi(h)$: uniformly costly core resolution and shortfall-adaptive outside screening.

B.4 Scaling in Market Size

The two terms of $\Psi(h)$ also carry dimensional factors, N^2 and $K-N$. Figure 4 fixes h and scales each factor. Left panel: core regret grows with fitted exponent 2.13 (95% CI $[1.71, 2.58]$, $R^2 = 1.000$) for $N \in \{3, 5, 7\}$ at $h = 0.09$, against the predicted N^2. Right panel: the Δ_K-normalized tail charge $R_{\mathrm{tail}}\Delta_K/\log(eT)$ grows with fitted exponent 0.96 (95% CI $[0.78, 1.16]$, $R^2 = 0.93$) for $K - N \in \{4, 12, 28, 60, 92\}$ at $N = 4$ and $d = 0.10$, against the predicted $K - N$; normalizing by the exact certificate margin Δ_K of each instance (which varies by less than 0.2% across K here) isolates the dimensional factor. The experiments thus validate the dimensional factors of the two terms in $\Psi(h)$.

B.5 Learning at the Zero-Certificate Boundary

The zero-certificate module has $\Delta = 0$ with $\Delta_{\mathrm{pres}} = h$: no stable matching admits a strictly positive certificate, and only the unstable side remains separated, at resolution h. Figure 5 uses $N = K = 5$, 30 test seeds, and horizons $T \in \{2^{14}, 2^{16}, 2^{18}\}$. In the left panel ($h = 0.04$), the full-profile control grows exactly linearly over the $16\times$ horizon increase (39,526 $\to$ 632,422, ratio 16.0) because it never commits; ACTL-Practical grows strongly sublinearly (2,285 $\to$ 3,253, ratio 1.42) and reaches weakly stable play on 98.5–99.7% of rounds at the largest horizon; the oracle-certificate reference provides a margin-aware benchmark. In the right panel, regret at $T = 2^{18}$ decreases as the preservation margin grows (3,253 $\to$ 1,132 as h goes from 0.04 to 0.12). Positive certification is unavailable at $\Delta = 0$, but preserving enough threshold comparisons still permits stable learning: the relevant finite-horizon behavior tracks Δ_{pres} rather than a positive certificate gap.

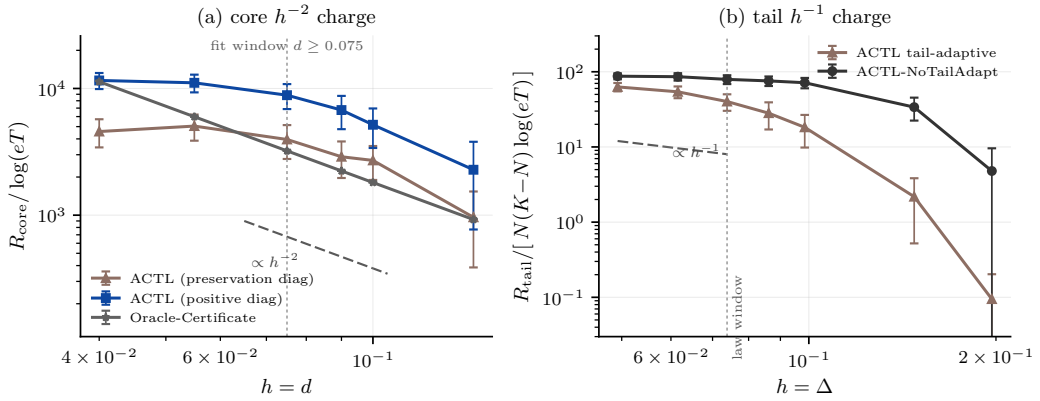

Figure 3: The two channels of the core–tail law. Left: the core validation charge $R_{\text{core}}/\log(eT)$ versus h; on the uncensored window $d \geq 0.075$ the validation-driven curves track the predicted h^{-2} slope, and the oracle-certificate benchmark attains exactly that slope. Right: the tail charge $R_{\text{tail}}/[N(K-N)\log(eT)]$ versus the exact certificate margin $h = \Delta$; the shortfall-adaptive allocation tracks the predicted h^{-1} decay, while the uniform-target ablation ACTL-NoTailAdapt is flatter and higher at every grid point. Vertical dotted lines mark the fitted windows; outside them finite-horizon censoring or pilot coverage prevails. Means with 95% confidence intervals over test seeds.

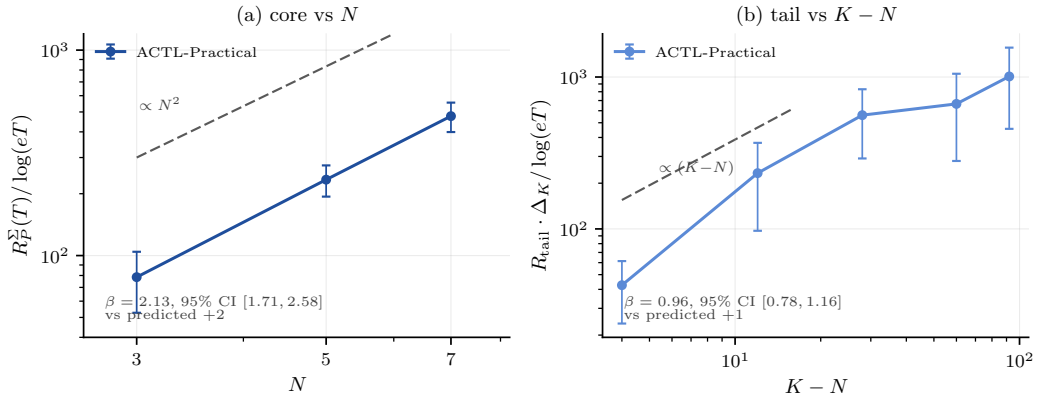

Figure 4: Dimensional factors of the core–tail law. Left: core regret versus N at $h = 0.09$ ($\beta = 2.13$ against the predicted $+2$). Right: the Δ_K-normalized tail charge $R_{\text{tail}}\Delta_K/\log(eT)$ versus $K-N$ at $N = 4$ and $d = 0.10$, where Δ_K is the exact certificate margin of each K instance ($\beta = 0.96$ against the predicted $+1$). Log–log axes with reference-slope guides. Means with 95% confidence intervals over test seeds.

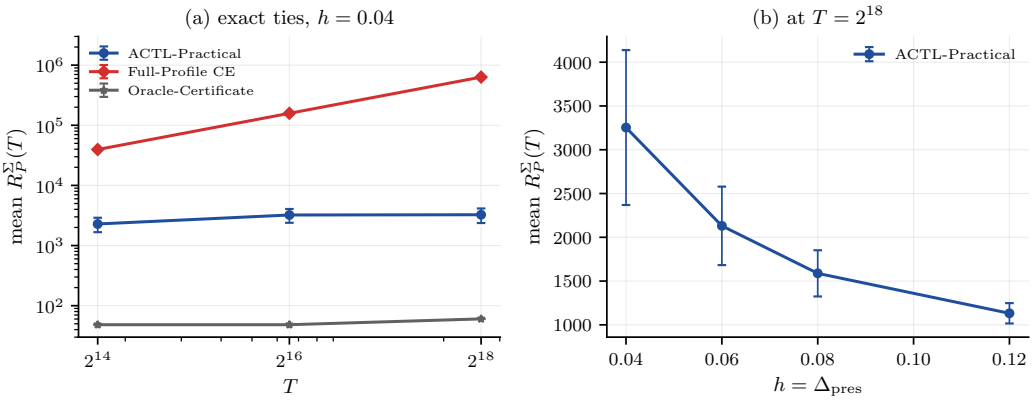

Figure 5: Learning at the zero-certificate boundary ($\Delta = 0$, $\Delta_{\text{pres}} = h$, $N = K = 5$, 30 test seeds). Left: regret versus T at $h = 0.04$; the full-profile control is exactly linear, while ACTL-Practical is strongly sublinear, with the margin-aware oracle-certificate benchmark shown for comparison. Right: regret at $T = 2^{18}$ decreases with the preservation margin h.

B.6 Comparison with Published Two-Sided Strict-Market Methods

Table 4 compares ACTL-Practical with the two closest published two-sided learning methods, Basu-SS (Basu, 2025) and E-PSM (Athanasopoulos et al., 2026). All rows use $T = 2^{15}$ and twelve test seeds; entries are means of $R_P^{\Sigma}(T)$, with the weakly stable round fraction in parentheses. Across the tested strict core, strict-tagged tail, and random instances, ACTL-Practical achieves substantially lower finite-horizon aggregate regret and a higher fraction of weakly stable rounds than the two identification-oriented published methods, whose certification phases remain expensive at the tested horizons: Basu-SS enters its super-stable phase only in the largest-margin market (75% of rounds there, and never elsewhere within T), and E-PSM commits only there (at round ≈7.6k). Because the methods optimize different regret criteria, these numbers should be interpreted as a finite-horizon empirical comparison rather than an asymptotic ordering of their theoretical guarantees. EE-PSM coincides with E-PSM in every cell. All rows use strict-preference instances, so the published methods are evaluated within their stated preference regime.

Setting	ACTL-Practical	Basu-SS	E-PSM
core d=0.0625	4,289 (0.82)	47,522 (0.20)	43,971 (0.01)
core d=0.40	1,140 (0.98)	11,924 (0.80)	8,917 (0.78)
tail K=16	3,454 (0.46)	18,970 (0.06)	18,970 (0.06)
tail K=32	6,470 (0.20)	15,987 (0.03)	15,987 (0.03)
random strict	534 (0.97)	35,812 (0.00)	34,120 (0.00)

Table 4: Finite-horizon comparison with the two closest published two-sided methods ($T = 2^{15}$, twelve test seeds, means over seeds): aggregate player regret $R_P^{\Sigma}(T)$ with the weakly stable round fraction in parentheses. All rows use strict-preference instances. EE-PSM coincides with E-PSM in every cell.

C Complete Aggregate Lower-Bound Constructions

We give the separate core and tail reward tables underlying Figure 1. Every lower-bound reward is an independent unit-variance Gaussian observation, across edges, endpoints, and times. This is a submodel of the conditional sub-Gaussian model. A change of $2d$ in one mean gives one-sample Kullback–Leibler divergence $2d^2$.

C.1 CORE REWARD TABLES

For $N = 2n + 1$, order the players as $p_1 = i_0$, $p_{1+i} = u_i$, $p_{1+n+j} = v_j$. Order arms as $x, b_1, \ldots, b_n, a_1, \ldots, a_n$, followed by the $K - N$ outside arms. Denote the arm with global label k by a_k, so that the dangerous arm $a_j = \mathsf{a}_{1+n+j}$. This distinguishes its local label from the generic arm index.

Let P^0, A^0 be untagged mean tables. On core arms every unspecified player mean is $1/4$; on outside arms every player mean is $1/8$. Override by

$$\begin{aligned} P^0_{i_0,x} &= 7/8, & P^0_{u_i,b_i} &= 1/2, & P^0_{u_i,a_j} &= 3/4, \\ P^0_{v_j,a_j} &= 7/8, & P^0_{v_j,x} &= 7/8. \end{aligned} \tag{7}$$

Every unspecified arm mean is $1/4$, including on outside arms. Override by

$$A^0_{x,i_0} = 7/8, \quad A^0_{b_i,u_i} = 7/8, \quad A^0_{a_j,v_j} = 1/2 + d, \quad A^0_{a_j,u_i} = 1/2. \tag{8}$$

The alternative (i, j) changes only $A^{ij}_{a_j,u_i} = 1/2 + 2d$; player means are unchanged. Write ν^0 and ν^{ij} for the base and alternatives, with the regime understood.

To obtain either regime, let $\tau = d/(64K)$ and use $s = 0$ for ties, $s = 1$ for strict preferences. For every base or alternative table P, A, set

$$\mu^{P,s}_{p_r,\mathsf{a}_k} = P_{p_r,\mathsf{a}_k} + s\tau k, \qquad \mu^{A,s}_{\mathsf{a}_k,p_r} = A_{\mathsf{a}_k,p_r} + s\tau r. \tag{9}$$

The same tags are used in every alternative. They break every original tie when $s = 1$, and each is at most $d/64$. Distinct original levels remain separated. All means lie in $(0, 1)$, the strict gap is at least τ, and the changed coordinate still shifts by exactly $2d$.

Lemma 1 (Core structure and margins)**.** *In either regime, the base has the unique weakly stable matching*

$$W^0(i_0) = x, \quad W^0(u_h) = b_h, \quad W^0(v_l) = a_l.$$

Alternative (i, j) has the unique weakly stable matching

$$W^{ij}(i_0) = x, \quad W^{ij}(u_i) = a_j, \quad W^{ij}(v_j) = b_i,$$

with all other pairs unchanged. Outside arms are unmatched. In the tied regime $\Delta_{\min} = \Delta = d$. In the strict regime $d/2 \le \Delta \le 2d$.

Proof. The pair (i_0, x) is mutual first choice, hence forced. In the base a displaced guard cannot use x. Its own dangerous arm is its best remaining partner, and that arm ranks the guard first, forcing every guard onto its own arm. Each row must then hold its home, which it prefers to every remaining arm and which ranks the row first. Direct inspection shows this matching is stable.

In alternative (i, j), all other guard–dangerous-arm pairs remain forced. Row u_i cannot hold one of those arms. If it does not hold a_j, it prefers a_j to its partner, and the changed arm ranks u_i first. Thus (u_i, a_j) is forced. Every other row holds its home. The remaining guard v_j must hold b_i: it prefers b_i to the outside arms, and an unmatched b_i accepts it. The displayed matching is stable. The tags preserve every strict comparison used here.

In the tied base, row deviations to dangerous arms are rejected by the arm with margin d. All other deviations have a rejecting side with a constant margin greater than d. In particular, x rejects a guard by $5/8$, even though that guard is indifferent between x and its own arm. In an alternative, a_j rejects the displaced guard with margin d; other dangerous deviations have margin d or $2d$. Remaining rejections have margin at least $1/8$. Thus both displayed margins equal d, and the smallest positive list gap is d. Under the tags (9), each mean moves by at most $\tau K = d/64$, so each margin (a comparison of two means) changes by at most $d/32$. Uniqueness implies that each positive stable margin equals the maximum Δ, proving the strict bounds. □

Lemma 2 (Aggregate core charging)**.** *Let $N_{ij}(T) = \sum_{t \le T} \mathbf{1}\{M_t(u_i) = a_j\}$ and $B_j(T) = \sum_{t \le T} \mathbf{1}\{M_t(v_j) = a_j\}$. In either regime,*

$$R^{\Sigma}_P(T; \nu^0) \ge \frac{1}{2} \sum_{i,j} N_{ij}(T), \tag{10}$$

$$R^{\Sigma}_P(T; \nu^0) \ge \tfrac{1}{2}(T - B_j(T)), \qquad R^{\Sigma}_P(T; \nu^{ij}) \ge \tfrac{1}{5} B_j(T). \tag{11}$$

Proof. A round with m row–dangerous-arm tests displaces m distinct guards. At most one of them can receive x. Every other displaced guard receives neither its own arm nor x, losing at least $5/8 - d/64 > 1/2$. If one receives x, player i_0 loses at least the same amount. This proves a loss of at least $m/2$, and also the first inequality of (11) for a specified displaced guard.

In alternative (i, j), suppose v_j holds a_j. Row u_i cannot hold it. If u_i holds a nondangerous arm, its reward is at most $1/2 + d/64$ against a stable reward at least $3/4$, losing more than $1/5$. Otherwise it displaces another guard from a_l, $l \neq j$. That guard either receives a low arm and loses more than $1/2$, or receives x and causes that loss to i_0. Summing establishes the second inequality. □

C.2 Tail reward tables

If $K = N$, omit this family. Otherwise label players $p_1, \dots, p_N$, home arms $x_1, \dots, x_N$, and the $K - N$ outside arms $z_1, \dots, z_{K-N}$, in that order. Set

$$P^0_{p_i,x_h} = \begin{cases} 1/2, & h = i, \\ 1/8, & h \neq i, \end{cases} \qquad P^0_{p_i,z_l} = 1/2 - d, \tag{12}$$

$$A^0_{x_h,p_i} = \begin{cases} 7/8, & i = h, \\ 1/4, & i \neq h, \end{cases} \qquad A^0_{z_l,p_i} = 1/4. \tag{13}$$

Alternative (i, l) changes only P_{p_i,z_l} to $1/2 + d$. Apply (9) using these labels for the strict regime. In particular, the outside arm's priorities are unchanged across alternatives. Denote these instances by $\nu^{\mathrm{tail},0}$ and $\nu^{\mathrm{tail,il}}$.

Lemma 3 (Tail structure, margins, and charging)**.** *The base has the unique stable matching assigning each player its home. In alternative (i, l), its unique stable matching assigns p_i to z_l and all other players to their homes; x_i is unmatched. The tied markets have $\Delta_{\min} = \Delta = d$, and the strict markets have $\Delta_{\min} \geq d/(64K)$ and $d/2 \leq \Delta \leq 2d$. With $N_{il}(T) = \sum_{t \leq T} \mathbf{1}\{M_t(p_i) = z_l\}$ and $B_i(T) = \sum_{t \leq T} \mathbf{1}\{M_t(p_i) = x_i\}$,*

$$R^{\Sigma}_P(T; \nu^{\mathrm{tail},0}) \geq \frac{d}{2} \sum_{i,l} N_{il}(T), \tag{14}$$

$$R^{\Sigma}_P(T; \nu^{\mathrm{tail},0}) \geq \tfrac{d}{2}(T - B_i(T)), \qquad R^{\Sigma}_P(T; \nu^{\mathrm{tail,il}}) \geq \tfrac{d}{2} B_i(T). \tag{15}$$

Proof. Every player and its home are mutual first choices in the base. In an alternative, all unchanged mutual-first pairs remain forced. If p_i does not receive z_l, that outside arm is unmatched and accepts it, while the player strictly prefers it to every available arm. This forces the stated matching; it does not require any special outside-arm priority.

Departing from a home in the base loses at least $d - d/64 > d/2$. Using x_i instead of z_l in the alternative loses at least the same amount. These inequalities imply the charges. Before tagging, outside rejections in the base have margin d, as does the moved player's rejection of its unmatched home in the alternative. All other relevant margins are at least d. The minimum positive list gap is d. Under the tags (9), each mean moves by at most $\tau K = d/64$ and each margin by at most $d/32$; uniqueness then proves the strict assertions. □

C.3 Even-size padding

For even $N \geq 4$, build the tied core on $N - 1$ players and $K - 1$ arms, then add a private dummy pair with means $15/16$ on both sides. Every cross edge involving exactly one dummy endpoint has means $1/16$ on both sides. The dummy pair is forced by mutual first choice. Original players cannot reduce a charging loss by using the dummy arm, and the dummy cannot improve any original stable matching. The new positive gaps exceed d. Apply (9) to the entire padded table with the full N, K labels for the strict regime. New rejection margins exceed d; old margins change by at most $d/32$. Since $(N - 2)^2 = \Theta(N^2)$, the core rate remains unchanged. The tail construction already works for every N.

C.4 THE ZERO-CERTIFICATE MODULE

Fix $N \geq 5$, $K \geq N$, and a scale $d \leq 1/64$. Start from a tied hard instance of Section C on $N-2$ players and $K-2$ arms at scale d, and append two players q_1, q_2 and two arms y_1, y_2. For every $r, s \in \{1, 2\}$, every old player p, and every old arm a, internal edges carry the same mean on both sides, and every cross edge carries $1/64$ in both reward directions:

$$\mu^P_{q_r,y_s} = \mu^A_{y_s,q_r} = \frac{15}{16}, \qquad \mu^P_{p,y_s} = \mu^A_{y_s,p} = \frac{1}{64}, \qquad \mu^P_{q_r,a} = \mu^A_{a,q_r} = \frac{1}{64}.$$

All other means keep their values from the old table. Call the result the enlarged instance.

Lemma 4 (Preservation margin of the tied hard families)**.** *Every tied core or tail hard instance of construction scale d has $\Delta_{\mathrm{pres}} = d$.*

Proof. The tied families have smallest positive real-partner preference gap exactly d, so a pair with both endpoints strictly improving improves each of them by at least d, and for an unmatched arm the player-side strict improvement is at least d. An unstable matching has a blocking pair, whose certificate is at most $-d$; hence every unstable matching satisfies $g(W) \leq -d$.

Conversely, exhibit an unstable matching with $g(W) = -d$. For a core base instance, modify the stable matching by assigning row player u_i to dangerous arm a_j and the displaced guard v_j to row home b_i, leaving all other assignments unchanged; direct inspection shows that (v_j, a_j) is the unique blocking pair, with guard-side gain constant and arm-side gain exactly d, so its pair certificate is exactly $-d$; since it is the unique blocking pair with strength d, every other off-matching pair has certificate at least $-d$, and $g(W) = -d$. For a core alternative (i, j), the original base stable matching is unstable only through (u_i, a_j), whose player-side gain is constant and arm-side gain is d; the same argument gives pair certificate $-d$ and $g(W) = -d$. For a tail base instance, move p_i from its home x_i to an outside arm z_ℓ and leave x_i unmatched: the pair (p_i, x_i) is the only blocker, its pair certificate is $-d$, and no pair has a smaller certificate, so $g(W) = -d$. For the corresponding tail alternative, keep p_i at x_i and leave z_ℓ unmatched: the pair (p_i, z_ℓ) is the only blocker and again its certificate is exactly $-d$. Therefore $\Delta_{\mathrm{pres}} = d$ in all tied hard instances. □

Lemma 5 (Module structure and margins)**.** *In the enlarged instance, every weakly stable matching matches q_1, q_2 to y_1, y_2 internally and restricts to a weakly stable matching of the old market, and conversely every such combination is weakly stable. Every stable matching has $g(W) = 0$, so $\Delta = 0$; and $\Delta_{\mathrm{pres}} = d$.*

Proof. If a matching does not match both new players to the two new arms, then some new player is matched outside the module, and some new arm is either unmatched or matched to an old player. That new player and that new arm each receive at most $1/64$ from their current assignments by the module table, while they value one another at $15/16$, and an unmatched arm accepts every real player. They therefore form a blocking pair. Hence every weakly stable matching matches the module internally, and its restriction to the old market must be weakly stable, since every old blocker remains a blocker. Conversely, combining a weakly stable old matching with either internal perfect matching is weakly stable: the unused internal pair is exactly tied and does not block, and no cross pair blocks because every endpoint strictly prefers its current internal or stable partner to any cross edge.

Fix either combination. The unused internal pair is exactly tied at both endpoints, so its certificate value is 0; hence $g(W) = 0$ for every stable W, and $\Delta = 0$ since unstable matchings have $g(W) < 0$. Every unstable matching with an internally matched module restricts to an unstable old matching, which by Lemma 4 has $g_{\mathrm{old}}(M) \leq -d$, that is, a blocking pair of strength at least d that remains blocking in the enlarged market; its pair certificate is at most $-d$ there. Every matching without an internally matched module has the new-player/new-arm blocker with strength at least $15/16 - 1/64 > d$, so its certificate is at most $-d$ as well. Hence $g(W) \leq -d$ for every unstable W, and $\Delta_{\mathrm{pres}} \geq d$. Taking an old witness matching with $g_{\mathrm{old}}(M) = -d$ and combining it with either internal perfect matching leaves exactly the old blockers, so $g(W) = -d$ and $\Delta_{\mathrm{pres}} \leq d$. Therefore $\Delta_{\mathrm{pres}} = d$. □

Lemma 6 (Charging and information preservation)**.** *For the enlarged core and tail families, the base and every alternative satisfy the divergence bound $\mathrm{KL}(\mathbb{P}^T_0, \mathbb{P}^T_r) \leq 2d^2 \mathbb{E}_0 N_r(T)$ and the charging inequalities (10)–(11) and (14)–(15) with universal constants.*

Proof. The module and all cross edges carry the same means in the base and in every alternative, and the changed old coordinate is unchanged, so the divergence bound is inherited from the old instance. Each charging inequality is a per-round lower bound on aggregate player regret, and the old arguments extend because the new options are never better than the old low arms: in the core, a displaced guard that reaches a module arm earns $1/64 \le 1/4$ and still loses more than $1/2$ against its stable baseline payoff, and in an alternative a row player forced onto a module arm loses at least $3/4 - 1/64$ against its stable reward $3/4$. In the tail, a player assigned to a module arm loses at least $1/2 - 1/64 > d/2$ against its home value, and a new participant displaced from the module loses at least $15/16 - 1/64$ (its reward on any old arm is $1/64$ by the module table). The tail information edge is unchanged and its player-side loss remains exactly d. Hence the charging inequalities hold with the same universal constants. □

D Proof of the Aggregate Lower Bound

We use $\mathrm{KL}(P, Q)$ for Kullback–Leibler divergence between probability laws and $\mathrm{kl}(p, q)$ for Bernoulli divergence. The following finite-family reduction makes the threshold uniform over policies.

Lemma 7 (Finite-family testing)**.** *Let $\nu^0, \dots, \nu^q$ be a fixed finite family and $\mathcal{R}(T) = \max_{0 \le r \le q} \mathbb{E}_{\nu^r} R_P^{\Sigma}(T)$. Let $\mathbb{P}_r^T$ be its full interaction-history law. Suppose events E_r and counts $N_r(T)$, $r \ge 1$, satisfy*

$$\mathbb{P}_0^T(E_r^c) \le A\mathcal{R}(T)/T, \quad \mathbb{P}_r^T(E_r) \le A\mathcal{R}(T)/T, \quad \mathrm{KL}(\mathbb{P}_0^T, \mathbb{P}_r^T) \le \kappa d^2 \mathbb{E}_0 N_r(T),$$

and $R_P^{\Sigma}(T; \nu^0) \ge w \sum_{r=1}^{q} N_r(T)$. For fixed positive A, κ, d, w, q, every policy satisfies

$$\mathcal{R}(T) \ge \frac{wq}{16\kappa d^2} \log T$$

for all sufficiently large T, with a threshold depending only on these constants.

Proof. Suppose the strict reverse inequality holds at a horizon T. For a sufficiently large threshold depending only on the displayed constants, it implies $A\mathcal{R}(T)/T \le T^{-1/2}$. Put $p = \mathbb{P}_0^T(E_r)$ and $q_r = \mathbb{P}_r^T(E_r)$. Then $p \ge 1 - T^{-1/2}$ and $q_r \le T^{-1/2}$. The inequality $\mathrm{kl}(p, q_r) \ge p \log(1/q_r) - \log 2$ gives $\mathrm{kl}(p, q_r) \ge \frac{1}{4} \log T$ for all sufficiently large T. Binary data processing and the KL assumption imply $\mathbb{E}_0 N_r(T) \ge \log T/(4\kappa d^2)$. Summing the charging bound contradicts the assumed reverse inequality by a factor of four. The threshold was chosen using only the hypothetical numerical regret bound and fixed family constants, so it is uniform over policies, including horizon-indexed policies. □

Proof of Theorem 1. In the core family use $E_{ij} = \{B_j(T) \ge T/2\}$. The separation inequalities and Markov's inequality give the event bounds with $A = 10$. The adaptive KL chain rule gives

$$\mathrm{KL}(\mathbb{P}_0^T, \mathbb{P}_{ij}^T) = 2d^2 \mathbb{E}_0 N_{ij}(T),$$

because only one independent Gaussian reward coordinate changes by $2d$. Lemma 7 with $w = 1/2$, $q = n^2$, and $\kappa = 2$ yields $cn^2 \log T/d^2$.

For the tail use $E_{il} = \{B_i(T) \ge T/2\}$. Lemma 3 gives $A = 4/d$, $w = d/2$, and $q = N(K - N)$. Again $\kappa = 2$. This gives $cN(K-N) \log T/d$. In each regime, take the union of core and tail families. The maximum expected regret over this union is at least the larger bound, hence at least half their sum up to a universal constant. Since $n \ge N/3$, the odd-N result follows. With $K = N$ only the core is needed. The constructions establish the gap and uniqueness claims, and the padding argument gives even N.

A decentralized policy and its local transcript are history-dependent. Conditional on the complete past, message and routing kernels are instance-independent functions of available observations and protocol randomness, so they add no KL beyond reward observations. In the connected-graph realization, every physical slot still plays an injection; the same lower bound therefore applies with the physical horizon, irrespective of control traffic. It supplies a statistical lower bound, not a communication lower bound. □

Proof of Theorem 2. Build the enlarged families at scale $d = h$ as in Appendix C.4: the tied hard families of Theorem 1 at size $N' = N - 2$, $K' = K - 2$, including the even-size padding when N' is even, plus the module. By Lemma 5, every enlarged instance has $\Delta = 0$ and $\Delta_{\mathrm{pres}} = d = h$ exactly, so the families lie in $\mathcal{C}_{\mathrm{pres}}(h)$. By Lemma 6, the base and every alternative keep the divergence bound $\mathrm{KL}(\mathbb{P}_0^T, \mathbb{P}_r^T) = 2d^2\mathbb{E}_0 N_r(T)$ and the charging inequalities of the original construction, so Lemma 7 applies to each family with the same constants. Before adjoining the module, the core and tail families therefore force the rates

$$\Omega\left(\frac{(N-2)^2}{d^2}\log T\right), \qquad \Omega\left(\frac{(N-2)(K-N)}{d}\log T\right),$$

respectively, since $K' - N' = (K-2) - (N-2) = K - N$, and the module leaves these charges unchanged. Since $N \geq 5$, $N - 2 = \Theta(N)$ and $(N-2)^2 = \Theta(N^2)$, so the union of the two finite families forces $\Omega(\Psi(d)\log T)$ with $d = h$. The threshold is uniform over policies for the same reason as in the proof of Theorem 1. □

E Complete Specification of ACTL

This appendix gives the complete centralized ACTL policy analyzed in Section 5.1; Appendix F proves its guarantees. Appendix G separately specifies ACTL^q, which shares the structural engine below but changes the pilot and snapshot quantizers and target constant for paid local communication.

Implementation map.

Pilot and exact maximum-certificate search	Section E.1
Runtime state, confidence clocks, and snapshots	Section E.2
Target-mode and baseline initialization	Section E.3
Common core–tail validation engine	Section E.4
Cumulative validation targets	Section E.5
Structural candidate stream and PRESERVE updates	Section E.6
Certificate tests and fallback-cache protocol	Section E.7
Complete executable policy	Section E.8

Throughout, every non-statistical choice is deterministic and reads only public labels: ties in preference and UCB comparisons are broken in favor of the smaller public partner label, arm sets are ordered by increasing label, and the same conventions are used by the pilot, the search below, the snapshot-UCB–GS computations, and ACTL^q.

E.1 Pilot and exact maximum-certificate search

Let $\ell = \log(eT)$, $m_0 = \lceil \ell \rceil$, and $\eta_T = (T+1)^{-2}$, as in Section E.2. The pilot occupies the first Km_0 reward rounds. For $s = 0, \ldots, K-1$ define the cyclic matchings

$$P_s(i) = 1 + \big((i - 1 + s) \bmod K\big), \qquad i \in \mathcal{I},$$

and play $P_{(t-1) \bmod K}$ in pilot round $t = 1, \ldots, Km_0$; that is, the K matchings are repeated m_0 times. Every player–arm edge is sampled exactly once across $P_0, \ldots, P_{K-1}$—player i meets arm a precisely in the matching with $s = (a - i) \bmod K$—hence exactly m_0 times over the pilot. If $T < Km_0$, ACTL plays the first T rounds of the schedule, updates the raw statistics, and stops at T: no post-pilot object is initialized. Only a completed pilot produces the frozen quantized matrix and enters the post-pilot stages (Section E.8).

The pilot accumulates raw reward sums and counts for every endpoint. After a completed pilot, each endpoint empirical mean is clipped to $[0, 1]$ and quantized with the map Q_T of Section E.2, producing the *frozen pilot means* $\bar{\mu}_{i,a}^P, \bar{\mu}_{a,i}^A$, which are never updated again. The signed pilot certificate $\bar{g}(W)$ is defined as $g(W)$ in Definition 1, with the frozen pilot means in place of the true means. The empirical certificate remains signed: $-1 \leq \bar{g}(W) \leq 1$ in general, while its maximum satisfies $0 \leq \bar{\Delta} = \max_W \bar{g}(W) \leq 1$.

For a threshold $h > 0$, use the frozen quantized pilot means to define

$$b \succ_{i,h} a \iff \bar{\mu}_{i,b}^P - \bar{\mu}_{i,a}^P \geq h, \qquad j \succ_{a,h} i \iff \bar{\mu}_{a,j}^A - \bar{\mu}_{a,i}^A \geq h. \tag{16}$$

All real partners rank above unmatched. These relations are asymmetric and transitive: two successive differences of at least h give a difference of at least $2h$. Thus they are strict partial orders.

A matching is super-stable for this partial profile when no unmatched pair has both endpoints failing to strictly prefer their current partners to the challenger. Since the threshold relations are strict partial orders on both sides, they satisfy the poset-on-one-side and asymmetric-on-the-other assumptions of Cseh & Juhos (2021, Theorem 5.2). That result decides existence and constructs a super-stable matching, if one exists, in $O(m^2)$ time for m acceptable edges. Our complete bipartite profile is a special case, with $m = NK$. Any super-stable matching covers all players: otherwise, since $K \geq N$, an unmatched player and an unmatched arm would form a super-blocking pair.

Lemma 8 (Threshold feasibility)**.** *For $0 < h \leq 1$, an injection W with $\bar{g}(W) \geq h$ exists exactly when the threshold profile (16) admits a super-stable matching. Every returned matching then has $\bar{g}(W) \geq h$.*

Proof. For $a = W(j)$ and $a \neq W(i)$, absence of a super-block requires $W(i) \succ_{i,h} a$ or $j \succ_{a,h} i$. This is exactly $\bar{c}_W(i, a) \geq h$, where $\bar{c}$ uses the frozen pilot means in Definition 1. An unmatched arm cannot reject a real player, so an outside pair requires the player's difference to be at least h, again the certificate condition. Enforcing this for all pairs, together with $h \leq 1$, is equivalent to $\bar{g}(W) \geq h$. □

Define the set of positive breakpoints

$$\mathcal{D} = \{1\} \cup \{\bar{\mu}^P_{i,a} - \bar{\mu}^P_{i,b} > 0 : i \in \mathcal{I},\ a, b \in \mathcal{A}\} \cup \{\bar{\mu}^A_{a,i} - \bar{\mu}^A_{a,j} > 0 : a \in \mathcal{A},\ i, j \in \mathcal{I}\}. \tag{17}$$

All breakpoints are at most 1. Write the sorted distinct values as $D_1 < \cdots < D_B$, where $B = O(NK(N + K))$. Every positive $\bar{g}(W)$ is the minimum of the cap 1 and a finite set of empirical differences; hence the positive optimum belongs to $\mathcal{D}$. Feasibility is monotone in the threshold, so binary search finds the exact optimum. Fix once and for all a deterministic implementation of the super-stable-matching subroutine, including a deterministic ordering of all internal choices; centralized ACTL and ACTL^q use this same implementation, with ties resolved in favor of the smaller public player label and then the smaller public arm label.

Algorithm 2 Exact maximum-margin search

Require: Clipped means $\bar{\mu}^P, \bar{\mu}^A$
Construct and sort $\mathcal{D}$; set $l = 1$, $r = B$, $\bar{\Delta} = 0$, $\bar{W} = \varnothing$.
while $l \leq r$ **do**
  $m = \lfloor (l + r)/2 \rfloor$, $h = D_m$; form (16).
  Run the deterministic super-stable matching routine.
  if a super-stable matching W_h exists **then**
    $\bar{\Delta} = h$, $\bar{W} = W_h$, $l = m + 1$.
  else
    $r = m - 1$.
  end if
end while
return $(\bar{W}, \bar{\Delta})$; if $\bar{\Delta} = 0$, the matching is unused.

Sorting costs $O(B \log(eB))$, and $O(\log(eB))$ super-stable calls cost $O(N^2K^2 \log(NK(N + K)))$, so the whole search costs $O((B + (NK)^2) \log(eB))$ with $B = O(NK(N + K))$. Generating the list is lower order. The list uses $O(NK(N + K))$ temporary storage; reward statistics use $O(NK)$. The search is invoked at most once. Every evaluation used by the search, the candidate stream, and the validation targets reads only the frozen quantized pilot matrix.

The search is skipped when the gate cannot trigger. If $4\varepsilon \geq 1$, the implementation does not execute Algorithm 2: since $\bar{\Delta} \leq 1 \leq 4\varepsilon$, the gate $\bar{\Delta} > 4\varepsilon$ of Section E.3 cannot trigger, so the search cannot affect any decision, the implementation proceeds directly to PRESERVE initialization, and no computed value of $\bar{\Delta}$ is read. If $4\varepsilon < 1$, the search returns $(\bar{W}, \bar{\Delta})$, with $\bar{W}$ produced by the fixed deterministic subroutine above. In both cases the symbol $\bar{\Delta}$ denotes the mathematical empirical optimum in every statement and proof; the distinction is only that the implementation computes it solely when $4\varepsilon < 1$.

Population means and the admissible gap. The threshold construction has a population-mean counterpart. Let π_h declare $b \succ_{i,h} a$ exactly when the true means satisfy $\mu^P_{i,b} - \mu^P_{i,a} \geq h$, and symmetrically on the arm side, with all real partners above unmatched. For $0 < h \leq 1$, an injection W is super-stable for π_h if and only if $g(W) \geq h$: the argument is that of Lemma 8 with the clipped pilot means replaced by the true means, because absence of a super-block at pair (i, a) requires one side to strictly prefer its current partner by at least h, which is exactly $c_W(i, a) \geq h$, and an arm unmatched by W can only be rejected by the player's side. In particular, an injection with $g(W) \geq h$ exists exactly when π_h admits a super-stable matching.

Proof of Proposition 2. Write $\gamma(P)$ for the minimum reward gap of a partial rank P, over comparable pairs of real partners on both sides; comparisons involving the unmatched option are excluded, as for $\Delta_{\min}$. By definition, $\Delta_{\mathcal{A}} = \max_P \gamma(P)$ over partial ranks consistent with the true full ranking that admit a super-stable matching. Fix such a P and let $\gamma = \gamma(P)$. Every pair comparable in P has a true difference of magnitude at least γ, and π_γ declares every real-partner pair whose true difference is at least γ; hence every comparison of P is also a comparison of π_γ, with the same direction. Super-stability is preserved when comparisons are added: if no pair has both endpoints weakly preferring the challenger under P, the same holds under the finer π_γ. Let W be super-stable for P; then W is super-stable for π_γ, and the population-mean equivalence above gives $g(W) \geq \gamma$. Hence $\Delta \geq \gamma$ for every admissible P, so $\Delta \geq \Delta_{\mathcal{A}}$. For the reverse inequality, let W attain Δ. Since the market is strict, every comparison declared by π_Δ agrees with the true ranking, so π_Δ is consistent with it, and the population-mean equivalence shows that W is super-stable for π_Δ; hence π_Δ is admissible. Every pair comparable in π_Δ is joined by a directed path of declared comparisons, each of true difference at least Δ, so the pair's true difference is at least Δ as well. Therefore $\gamma(\pi_\Delta) \geq \Delta$ and $\Delta_{\mathcal{A}} \geq \Delta$. □

Relation among structural gaps. Three instance-dependent constants appear in this literature: the smallest positive real-partner gap $\Delta_{\min}$; the admissible gap $\Delta_{\mathcal{A}}$ above; and the admissible gap Δ_F of Athanasopoulos et al. (2026), the largest minimum gap over partial preference profiles consistent with the true ranking that admit a pervasive stable matching, a matching that is optimal under every strict refinement of the profile (Rastegari et al., 2014). In strict markets with minimum gaps taken over comparable pairs of real partners, the three are ordered as

$$\Delta_{\min} \;\leq\; \Delta_F \;\leq\; \Delta_{\mathcal{A}} \;=\; \Delta.$$

Both admissible gaps are at least $\Delta_{\min}$ by definition. For the middle inequality, a profile that admits a pervasive stable matching admits a super-stable matching: the pervasive matching is stable under every strict refinement, and the intersection of the stable matchings over all refinements is the set of super-stable matchings (Basu, 2025, Proposition 2.1). Hence the family defining Δ_F is contained in the family defining $\Delta_{\mathcal{A}}$, and the same functional attains a smaller maximum. The final equality is Proposition 2. The two inequalities need not be tight.

E.2 Runtime state, confidence clocks, and snapshots

Set $\ell = \log(eT)$, $m_0 = \lceil \ell \rceil$, and $\eta_T = (T+1)^{-2}$. The anytime clock $b_t = \log(8NK(t+1)^5)$ drives the UCB scores, targets, and validation decisions; the horizon-wide clock $B_T = b_T$ protects every permanent commitment. For a public endpoint mean v_e based on m_e observations, define

$$a_e(t) = \sqrt{\frac{2b_t}{m_e}} + \eta_T, \qquad a^*_e = \sqrt{\frac{2B_T}{m_e}} + \eta_T, \qquad \gamma = \log\big(4(NK)^3\ell^8\big), \qquad \varepsilon = \sqrt{\frac{2\gamma}{m_0}} + \eta_T. \tag{18}$$

The raw statistical radii are $r_t(m) = \sqrt{2b_t/m}$ and $r_*(m) = \sqrt{2B_T/m}$; the UCB score of an endpoint is its public snapshot mean plus the anytime radius,

$$U_e(t) = v_e + a_e(t),$$

and its commitment interval is $[L^*_e, U^*_e] = [v_e - a^*_e, v_e + a^*_e]$.

Quantization. The quantizer $Q_T(x) = \eta_T \lfloor \mathrm{clip}_{[0,1]}(x)/\eta_T \rfloor$ satisfies $|Q_T(x) - x| \leq \eta_T$ for $x \in [0, 1]$ and takes at most $1 + 1/\eta_T = (T+1)^2 + 1$ values, so a quantized value costs $O(\log(eT))$ bits. Quantization only adds η_T to a statistical error, and clipping cannot increase error because all means lie in $[0, 1]$.

Counts and the strict-before convention. For every edge, n_e is the cumulative number of reward observations collected so far—pilot, validation, and fallback rounds all contribute—and it is never reset. The public snapshot of an edge consists of a snapshot count m_e and a quantized snapshot mean v_e per endpoint. At the beginning of reward round t, all counts n_e, snapshot counts m_e, snapshot means v_e, UCB scores $U_e(t)$, targets, and confidence intervals are computed from observations collected strictly before round t.

Snapshot protocol. After the pilot, $n_e = m_e = m_0$ for every edge and v_e is the frozen quantized pilot mean; the public snapshot version is 0. After a reward round, for every edge whose count has reached $2m_e$, set m_e to the current count, replace v_e by the quantizer applied to the endpoint empirical means over all observations of that edge, and, if at least one edge snapshot changed in this step, increment the single global snapshot version by one. Because a played matching samples each edge at most once, an edge count increases by at most one per reward round, so it reaches $2m_e$ exactly and the equality test cannot be skipped. This convention is deterministic, and ACTL only ever reads whether the version changed.

E.3 Target-mode and baseline initialization

After a completed pilot, ACTL fixes its mode and baseline by the following executable rule: if $4\varepsilon \geq 1$, it skips the search of Section E.1 and initializes PRESERVE directly, since the gate $\bar{\Delta} > 4\varepsilon$ cannot trigger ($\bar{\Delta} \leq 1$); otherwise it runs Algorithm 2 and branches on the returned $\bar{\Delta}$.

- If $4\varepsilon < 1$ and the search returned $\bar{\Delta} > 4\varepsilon$: the mode is CERTIFY, $W_c = \bar{W}$ is the deterministic empirical maximum-certificate matching returned by Algorithm 2, and the fixed targets $q_0^+, q_{i,a}^+$ of Section E.5 are initialized.
- Otherwise—whether because $4\varepsilon \geq 1$ or because the returned optimum did not exceed the gate—the mode is PRESERVE: W_c is the deterministic player-proposing Gale–Shapley matching on the strict profile obtained from the frozen pilot scores by breaking ties in favor of the smaller public partner label on both sides, and
$$\widehat{\beta}_i = \bar{\mu}^P_{i,W_c(i)} \quad (i \in \mathcal{I}), \qquad \widehat{h} = 1.$$
The structural candidate iterator is initialized at $Z = (1, \ldots, N)$ (Section E.6), and the adaptive targets $q_0(t), q_{i,a}(t)$ of Section E.5 are initialized with the current clock.

The guarantees are invariant to the particular deterministic tie-breaking rule: any tie-broken pilot profile is a strict refinement of the empirical weak preferences, and the proofs use only this refinement property (Appendix F); we fix increasing-partner-label everywhere for reproducibility.

E.4 Common core–tail validation engine

Fix the baseline W_c, and let $A_c = W_c(\mathcal{I})$ be its set of assigned arms, ordered by increasing label $a_0 < \cdots < a_{N-1}$. The N cyclic core matchings are
$$C_s(i) = a_{(i-1+s) \bmod N}, \qquad s = 0, \ldots, N-1,$$
and their edge sets partition $\mathcal{I} \times A_c$: player i meets each baseline arm exactly once across $C_0, \ldots, C_{N-1}$. For every pair (i, a) with $a \notin A_c$, the outside probe is the injection
$$Q_{i,a}(i) = a, \qquad Q_{i,a}(i') = W_c(i') \text{ for every } i' \neq i.$$
This is well defined: $a \notin A_c$ means no baseline partner equals a, and player i's baseline arm is freed, so distinct players keep distinct partners; every other player retains her baseline partner. Hence $Q_{i,a}$ is a feasible matching, and it is the probe analyzed in Appendix F.

Fixed action order. The validation action list is the cyclic list
$$C_0, \ldots, C_{N-1}, \qquad (i, a) \text{ with } a \notin A_c \text{ in lexicographic order of } (i, a),$$
where i runs over $\mathcal{I}$ in increasing order and, for each i, a runs over $\mathcal{A} \setminus A_c$ in increasing order.

Deficiency. Write q_0 for q_0^+ in CERTIFY and for $q_0(t)$ in PRESERVE, and $q_{i,a}$ for $q_{i,a}^+$ respectively $q_{i,a}(t)$, as defined in Section E.5. A core action C_s is deficient when at least one of its edges satisfies $n_{(i,C_s(i))} < q_0$. An outside probe (i, a) is deficient when $n_{(i,a)} < q_{i,a}$.

The selector. A single persistent cursor, initialized at C_0, scans the action list cyclically. NEXT-VALIDATION-ACTION returns the first deficient action it encounters and advances the cursor to that action's successor. If a full scan finds no deficient action, it returns COMPLETE and leaves the cursor unchanged. Candidate processing, target refreshes, snapshot updates, and cache refreshes never reset the cursor, and no edge count is ever reset.

E.5 CUMULATIVE VALIDATION TARGETS

Targets are cumulative sample counts: every observation of an edge—from the pilot, from validation, or from fallback replay—counts toward them, no count is reset, and the completion of a target never creates a new stage with fresh counts.

CERTIFY targets (fixed). Let $\bar{d}_{i,a} = \bar{\mu}^P_{i,W_c(i)} - \bar{\mu}^P_{i,a}$ for $a \notin A_c$. The fixed targets are

$$q_0^+ = \left\lceil \frac{1024\, B_T}{\bar{\Delta}^2} \right\rceil, \qquad q_{i,a}^+ = \left\lceil \frac{1024\, B_T}{\bar{d}_{i,a}^2} \right\rceil, \tag{19}$$

initialized once at mode selection and never changed.

PRESERVE targets (adaptive). Let $\widehat{\delta}_{i,a} = (\widehat{\beta}_i - \bar{\mu}^P_{i,a})_+$. The adaptive targets are

$$q_0(t) = \left\lceil \frac{1024\, b_t}{\widehat{h}^2} \right\rceil, \qquad q_{i,a}(t) = \left\lceil \frac{1024\, b_t}{\max\{\widehat{h}, \widehat{\delta}_{i,a}\}^2} \right\rceil. \tag{20}$$

In PRESERVE, the targets are recomputed before every selector invocation, after the current round's candidate has been processed; they are nondecreasing in time (Lemma 17 of Appendix F states this precisely), and an edge that was complete may fall below its target again after a later update, in which case validation resumes with the same cumulative counts and the same persistent cursor.

E.6 STRUCTURAL CANDIDATE STREAM AND PRESERVE UPDATES

Represent an injection as a tuple $Z = (Z(1), \ldots, Z(N))$ with pairwise distinct entries in $[K]$, and order the injections lexicographically among all such tuples. The iterator stores only the current tuple. NEXT-CANDIDATE emits the current tuple and advances it by the immediate lexicographic successor:

- scan positions $j = N, N-1, \ldots, 1$ and take the largest j for which $\{Z(j)+1, \ldots, K\} \setminus \{Z(1), \ldots, Z(j-1)\}$ is nonempty;
- replace $Z(j)$ by the smallest element of that set, and fill positions $j+1, \ldots, N$ with the smallest unused arms in increasing order;
- if no such j exists, replace Z by $(1, \ldots, N)$.

This produces the immediate lexicographic successor of Z: any tuple larger than Z first increases at some position j'', so the largest feasible position j' is at least j''; tuples differing first at $j' > j''$ are smaller, and among tuples differing first at j' the rule takes the smallest feasible increment and the lexicographically smallest completion. After the lexicographically last injection the wrap returns to the first, so one period visits every injection exactly once, in $S = K!/(K-N)!$ steps; the iterator never computes or stores S, uses $O(N \log K)$ bits, and each advance costs $O(NK)$ time.

Timing. Exactly one candidate is processed at the beginning of each uncommitted PRESERVE reward round, before the certificate checks and before the selector is invoked. The candidate is never played merely because it is generated; its signed certificate is evaluated only on the frozen pilot matrix.

UPDATE-PRESERVATION. Given the current candidate Z,

$$\text{if } \bar{g}(Z) > -4\varepsilon: \quad \widehat{\beta}_i \leftarrow \min\{\widehat{\beta}_i, \bar{\mu}^P_{i,Z(i)}\} \text{ for all } i \in \mathcal{I};$$
$$\text{otherwise } (\bar{g}(Z) \le -4\varepsilon): \quad \widehat{h} \leftarrow \min\{\widehat{h}, -\bar{g}(Z)\}.$$

The targets of Section E.5 are then refreshed using the current clock b_t, and no other state changes because of candidate processing. After a permanent commitment, candidate processing stops.

E.7 CERTIFICATE TESTS AND FALLBACK-CACHE PROTOCOL

The commitment test. For a matching W and every $a \neq W(i)$, the horizon-wide test requires

$$L^{P}_{*\,i,W(i)} > U^{P}_{*\,i,a} \quad \text{or} \quad \big[a = W(j) \text{ for some } j \text{ and } L^{A}_{*\,a,j} > U^{A}_{*\,a,i}\big], \tag{21}$$

where for an arm unmatched by W only the player test applies, and $L^*_e = v_e - a^*_e$, $U^*_e = v_e + a^*_e$ are the commitment intervals of Section E.2 (the superscript P or A selects the endpoint).

TRY-COMMIT. TRY-COMMIT(W) applies test (21) to W. On success, W is played from the current round onward, and ACTL permanently commits: candidate generation, validation, and fallback updates stop. On failure, nothing else changes—the targets, the selector cursor, and the fallback cache are untouched.

Invocation schedule. ACTL applies the test

1. to W_c at the initial pilot snapshot;
2. to the temporary player-proposing snapshot-UCB–GS output at the initial pilot snapshot;
3. to both again at every subsequent new snapshot version;
4. to W_c once, when the fixed CERTIFY targets first become complete;
5. to every newly established fallback cache.

Timing protocol. The variable $tested$ (the last tested snapshot version) is initialized to NONE, and the flag $ctested$ (the CERTIFY completion test) to FALSE. In each uncommitted round, after the round's candidate has been processed, if the current snapshot version differs from $tested$, ACTL runs checks (1)–(2), or (3) at a later version, and records the current version as tested. When the selector returns COMPLETE in CERTIFY and $ctested$ = FALSE, ACTL sets $ctested$ = TRUE and runs check (4); if that test fails, it proceeds to the fallback, and the completion test is not repeated on later COMPLETE rounds.

Snapshot-UCB–GS. Rank the edges by the current UCB scores $U_e(t)$ of Section E.2, breaking ties in favor of the smaller public partner label, and run the player-proposing Gale–Shapley algorithm on the resulting strict profile. The output is the temporary W_{UCB} used by checks (2)–(3) and, when a cache is established, the exploitation matching M_{cache}.

Fallback cache. The cache stores a matching M_{cache} together with its establishment snapshot version and dyadic index $\lfloor \log_2 t \rfloor$. It is established or refreshed only when the selector returns COMPLETE, no permanent commitment has occurred, and one of the following holds: the cache is empty; its stored snapshot version differs from the current one; or its stored dyadic index differs from the current $\lfloor \log_2 t \rfloor$. On a refresh, ACTL computes the current player-proposing snapshot-UCB–GS matching, stores it with both keys, and runs TRY-COMMIT on it, then plays it this round unless the test committed to it. If no refresh is required, the existing M_{cache} is replayed. Validation pauses replay but does not erase the cache, and the refresh conditions are rechecked before replay resumes. Temporary UCB–GS computations used only for the snapshot checks never establish or overwrite the cache.

E.8 COMPLETE EXECUTABLE POLICY

At an uncommitted post-pilot round t the chronology is: (1) read the state computed from observations strictly before t; (2) in PRESERVE, process exactly one structural candidate and refresh the targets; (3) run the snapshot-version certificate checks if the current version has not yet been tested; (4) run the validation selector; (5) if it returns a deficient action, play it, and otherwise run the due CERTIFY completion test and then establish or replay the fallback cache; (6) observe both endpoint rewards on every played edge; (7) update raw sums and counts; (8) process doubling-snapshot updates; (9) continue with round $t + 1$. For LC-ACTL, this is the virtual-slot order, with paid synchronization inserted after (8).

Complete executable specification of ACTL.

Require: N, K, T; public player and arm labels; deterministic tie-breaking (increasing public partner label).

Precompute: $\ell, m_0, \eta_T, b_t, B_T, \gamma, \varepsilon, Q_T$ (Section E.2).

Pilot: play $P_{(t-1) \bmod K}$ for $t = 1, \ldots, \min\{T, Km_0\}$; if $T < Km_0$, update raw statistics only and stop at T with no later stage; otherwise clip and quantize every endpoint mean, set $n_e = m_e = m_0$, and let the snapshot version be 0 (Section E.1).

Certificate search: if $4\varepsilon < 1$, run Algorithm 2 for $(\bar{W}, \bar{\Delta})$; otherwise do not search.

Mode:

if $4\varepsilon < 1$ **and** $\bar{\Delta} > 4\varepsilon$ **then**
  mode $\leftarrow$ CERTIFY; $W_c \leftarrow \bar{W}$; initialize the targets (19).
else
  mode $\leftarrow$ PRESERVE; $W_c \leftarrow$ tie-broken player-proposing GS matching on the frozen pilot scores; $\widehat{\beta}_i \leftarrow \bar{\mu}^P_{i,W_c(i)}$; $\widehat{h} \leftarrow 1$; $Z \leftarrow (1, \ldots, N)$; initialize the targets (20) with the current b_t.
end if

Online state: cursor $\leftarrow C_0$; cache $\leftarrow$ empty; tested version $\leftarrow$ NONE; completion-tested $\leftarrow$ FALSE.

for $t = Km_0 + 1, \ldots, T$ **do**
  If PRESERVE: $Z \leftarrow$ NEXT-CANDIDATE; UPDATE-PRESERVATION(Z); refresh the targets with b_t (Section E.6).
  If the current snapshot version differs from the tested version: TRY-COMMIT(W_c); compute the temporary player-proposing snapshot-UCB–GS matching; TRY-COMMIT(W_{UCB}); record the current version as tested (Section E.7).
  $a \leftarrow$ NEXT-VALIDATION-ACTION (Section E.4).
  if $a \neq$ COMPLETE **then**
    Play a.
  else
    If CERTIFY and completion-tested = FALSE: set it TRUE; TRY-COMMIT(W_c).
    If not committed and the cache is empty or its version or dyadic index differs from the current ones: compute $M_{\mathrm{cache}} \leftarrow$ player-proposing snapshot-UCB–GS; store both keys; TRY-COMMIT(M_{cache}).
    If not committed: play M_{cache}.
  end if
  Observe both endpoint rewards on every played edge; update raw sums and counts n_e.
  For every edge with $n_e = 2m_e$: set $m_e \leftarrow n_e$; quantize its snapshot means; increment the snapshot version once if any snapshot changed.
end for

On success of any TRY-COMMIT, the accepted matching is played from the current round onward, steps (2)–(5) of the chronology are skipped, and observations continue to update counts and snapshots. Every operation stops at T; if the horizon ends during the pilot, the pilot is truncated and no later stage begins.

Determinism. All non-statistical choices above are deterministic functions of public labels and synchronized state, once a deterministic implementation of the super-stable-matching subroutine is fixed (Section E.1): pilot matchings, baseline GS, arm order, action list, candidate order, snapshot processing, and cache keys. Sections E.1–E.8 determine centralized ACTL; Appendix G specifies the modified quantized state and shows equality of its retained action prefix with ACTL^q.

Proposition 3 (Computation and storage)**.** *In the arithmetic/comparison model for observed rewards, the one-time maximum-certificate search and every interaction round of centralized ACTL use polynomial work in $N, K, \log(eT)$. After that search, the policy stores $O(NK)$ scalar quantities plus an $O(N \log K)$-bit candidate iterator. Advancing the iterator in* PRESERVE *takes $O(NK)$ time; no step constructs or stores the full set of $K!/(K-N)!$ injections.*

Proof. Section E.1 bounds the one-time search and its temporary storage. The successor rule in Section E.6 stores one tuple and scans at most N positions and K arms per position. Computing $\bar{g}(Z)$ scans the $O(NK)$ possible player–arm pairs. Target refreshes, snapshot updates, and certificate checks inspect polynomially many stored edge statistics. The validation selector scans at most $N + N(K-N)$ actions, each involving at most N edges; a fallback refresh ranks finite preference

lists and runs player-proposing Gale–Shapley, which makes at most NK proposals. Counts, snapshots, targets, the action list, and the cache occupy $O(NK)$ scalar entries after initialization; the candidate tuple needs $O(N \log K)$ bits. Counts and quantized values have $O(\log(eT))$-bit representations. None of these operations depends on the number of injections except through the number of rounds actually executed. □

F Proof of the Upper Bounds

Write $h := \Delta_{\mathrm{pres}}$ throughout this appendix. Constants C below are universal and may change between displays. This appendix proves the upper bounds for the policy specified in Appendix E: the pilot and exact search of Section E.1, the clocks, radii, quantizer, and snapshot protocol of E.2, the targets of E.5, and the commitment and fallback-cache protocol of E.7. No policy state is redefined here. The proof first establishes concentration and snapshot coverage, then bounds cumulative validation cost. Witness discovery and pointwise preservation show when later fallback caches are stable; separate charging bounds control earlier caches and confidence failures. The final subsection combines these ingredients for both certificate regimes.

F.1 Concentration for the clocks and pilot

Recall from Section E.2 the parameters of (18), the clocks b_t and B_T, the radii $a_e(t)$ and a_e^*, and the quantizer Q_T. Write $\mathcal{P}$ for the event that every frozen pilot endpoint value lies within ε of its true mean; $\mathcal{H}_T$ for the event that every realized endpoint empirical mean, at every positive count $m \le T$, lies within $r_*(m)$ of its true mean; and $\mathcal{G}_t$ for the event that every realized endpoint empirical mean, at every positive count $m \le t$, lies within $r_t(m)$ of its true mean. The event $\mathcal{H}_T$ protects every irreversible commitment, while the events $\mathcal{G}_t$ protect the snapshots used at cache establishment times.

Lemma 9 (Concentration and two-sided snapshot coverage)**.** *With $p_T := 1/((NK)^2\ell^8)$,*

$$\Pr(\mathcal{P}^c) \le p_T, \qquad \Pr(\mathcal{H}_T^c) \le \frac{T}{2(T+1)^5}, \qquad \Pr(\mathcal{G}_t^c) \le \frac{t}{2(t+1)^5}.$$

On $\mathcal{P}$, every frozen pilot value is ε-accurate. On $\mathcal{G}_t$, every current snapshot obeys

$$\mu_e - 2a_e(t) \le v_e - a_e(t) \le \mu_e \le v_e + a_e(t) \le \mu_e + 2a_e(t), \tag{22}$$

and $U_e(t) \in [\mu_e, \mu_e + 2a_e(t)]$. On $\mathcal{H}_T$, every commitment interval obeys $L_e^* \le \mu_e \le U_e^*$; consequently every matching passing (21) is truly weakly stable, regardless of the adaptive time of the test.

Proof. For one endpoint of one edge, the product exponential process associated with its selected, centered observations is a nonnegative supermartingale. Stopping it at the mth observation, or at the finite horizon if that observation is not reached, gives

$$\Pr\Big(|\widehat{\mu}_m - \mu| > \sqrt{2x/m},\ m \text{ is reached}\Big) \le 2e^{-x} :$$

choosing $\lambda = \sqrt{2x/m}$ for the upper tail, the stopped exponential is at least e^x on the tail event, so Markov's inequality and optional stopping bound its probability by e^{-x}, and the same argument with $-\lambda$ handles the lower tail. No independence between the two endpoints, or across edges or rounds, is used.

The pilot has $2NK$ endpoint statistics at the deterministic count m_0, and a union bound with $x = \gamma$ gives $4NKe^{-\gamma} = p_T$ for the statistical part of $\mathcal{P}$; adding the quantization error η_T to the radius $\sqrt{2\gamma/m_0}$ gives ε. For $\mathcal{H}_T$, a union over $2NK$ statistics and the at most T realized positive counts gives $4NKTe^{-B_T} = T/[2(T+1)^5]$ because $e^{-B_T} = 1/[8NK(T+1)^5]$; for $\mathcal{G}_t$, a union over at most t counts gives $4NKte^{-b_t} = t/[2(t+1)^5]$.

Coverage of a raw mean by its clock radius, combined with $|Q_T(x) - x| \le \eta_T$, gives $|v_e - \mu_e| \le r_t(m_e) + \eta_T = a_e(t)$ on $\mathcal{G}_t$, which is equivalent to $v_e - a_e(t) \le \mu_e \le v_e + a_e(t)$ and implies the outer inequalities of (22). The $\mathcal{H}_T$ statement is identical with r_*, a_e^*. If a lower endpoint exceeds an upper endpoint, the corresponding true means are strictly ordered, so the pair is truly rejected by that side; a matching passing every test of (21) therefore has no blocking pair. The event $\mathcal{H}_T$ covers all possible invocation counts, including adaptive ones. □

Lemma 10 (Uniform pilot perturbation). *On $\mathcal{P}$,*

$$\sup_W |\bar{g}(W) - g(W)| \le 2\varepsilon, \qquad |\bar{\Delta} - \Delta| \le 2\varepsilon.$$

Here ε already includes the quantization error of the frozen pilot matrix.

Proof. Every real-partner reward difference changes by at most 2ε, because each endpoint value differs from its true mean by at most ε, and clipping cannot increase the error since all means lie in $[0, 1]$ both before and after clipping. The minimum and the maximum are 1-Lipschitz in the sup norm; applying this successively to the definitions of c_W, g, and Δ proves both bounds. □

F.2 Doubling snapshots

Lemma 11 (Snapshot invariant and effective radii). *At all times after the pilot, for every edge, $m_e \le n_e < 2m_e$, so*

$$a_e(t) \le 4\sqrt{b_t/n_e(t)}, \qquad a_e^* \le 4\sqrt{B_T/n_e}.$$

The snapshot counts follow $m_0, 2m_0, 4m_0, \ldots$; each edge has one initial snapshot and at most $\lfloor \log_2(T/m_0) \rfloor$ later snapshot updates.

Proof. At initialization $n_e = m_e = m_0$. The rule triggers exactly when $n_e = 2m_e$ and then sets $m_e = n_e$, so $m_e \le n_e$ is preserved, and between two updates $n_e < 2m_e$ holds because reaching $2m_e$ would have triggered an update; hence $m_e \le n_e < 2m_e$ at all times. Consequently $m_e > n_e/2$ and

$$a_e(t) = \sqrt{2b_t/m_e} + \eta_T < 2\sqrt{b_t/n_e} + \eta_T \le 4\sqrt{b_t/n_e},$$

where the last step uses $\eta_T \le 2\sqrt{b_t/n_e}$: since $n_e \le T$ and $b_t \ge \log(8NK \cdot 2^5) \ge \log 256$, we have $2\sqrt{b_t/n_e} \ge 2\sqrt{\log 256/T} \ge (T+1)^{-2} = \eta_T$ for every $T \ge 1$. The same argument gives $a_e^* \le 4\sqrt{B_T/n_e}$. Starting from m_0, each update doubles the snapshot count, so the snapshot values are $m_0 2^j$, with at most $\lfloor \log_2(T/m_0) \rfloor$ post-pilot updates. □

F.3 Preservation structure

The next two order-theoretic lemmas and the proof of Proposition 1 supply the partial-order tools used by the pointwise preservation lemma below.

Lemma 12 (Pivot extension). *Let $(X, \succ)$ be a finite strict partial order and fix $x \in X$. Some strict linear extension places x above every y that is not forced above x.*

Proof. Let $P_x = \{y : y \succ x\}$. Take a linear extension of the induced order on P_x, place x next, and append a linear extension of $X \setminus (P_x \cup \{x\})$. If some $z \notin P_x \cup \{x\}$ were required to precede an element of $P_x \cup \{x\}$, transitivity would give $z \succ x$, contradicting $z \notin P_x$. Hence this is a strict linear extension, and it places x above every element not forced above it. □

Lemma 13 (Extension with a prescribed induced order). *Let $(X, \succ)$ be a finite strict partial order, $S \subseteq X$, and let a strict total order on S extend the induced partial order. Some strict linear extension of $(X, \succ)$ restricts to that prescribed order on S.*

Proof. Add to the partial order every comparison between elements of S prescribed by the total order. A directed cycle of the enlarged relation would have to use a new comparison; contracting each original directed path between successive elements of S along the cycle produces steps that all move strictly forward in the prescribed order on S, because that order extends the induced order on S. Such steps cannot close a cycle. A topological ordering of the enlarged relation is therefore a strict linear extension of the original order and restricts to the prescribed order on S. □

Proof of Proposition 1. Suppose $g(W) \le -h$. Then some off-matching pair (i, a) has $c_W(i, a) \le -h$. If $a = W(j)$, both strict improvements $\mu^P_{i,a} - \mu^P_{i,W(i)} \ge h$ and $\mu^A_{a,i} - \mu^A_{a,j} \ge h$ belong to π_h and survive every strict linear extension, so (i, a) blocks W in every extension. If a is unmatched,

the player-side improvement is forced while the unmatched arm accepts every real player, with the same conclusion. Hence $W \notin \mathcal{W}_h$.

Conversely, suppose $g(W) > -h$. Apply Lemma 12 at every endpoint with the current W-partner as pivot, so each endpoint ranks its current partner above every challenger not forced above it by π_h. For a matched challenger $a = W(j)$, the inequality $g(W) > -h$ implies that at least one of the two strict improvements is below h; that endpoint is not forced to prefer the challenger and, by the pivot extension, strictly prefers its current partner. The pair cannot block. For an unmatched a, the player's improvement is below h, so the pivot extension ranks $W(i)$ above a and the pair does not block. Hence W is stable under the resulting strict completion, and $W \in \mathcal{W}_h$.

Every weakly stable W has $g(W) \geq 0$, and every unstable W has $g(W) \leq -\Delta_{\mathrm{pres}}$, so $\mathcal{W}_h = \mathcal{W}^\star$ for every $0 < h \leq \Delta_{\mathrm{pres}}$. In particular, at $h = \Delta_{\mathrm{pres}}$, every matching that is stable under some strict linear extension of $\pi_{\Delta_{\mathrm{pres}}}$ is truly weakly stable. For $h > \Delta_{\mathrm{pres}}$, a matching with $g(W) = -\Delta_{\mathrm{pres}} > -h$ belongs to $\mathcal{W}_h \setminus \mathcal{W}^\star$. This proves the characterization; if every matching is stable, both sides equal 1 by Definition 2. □

F.4 Resolved candidate classification

Lemma 14 (Empirical baseline)**.** *$\bar{g}(W_c) \geq 0$, where W_c is the tie-broken player-proposing GS output on the frozen pilot scores.*

Proof. The deterministic tie-breaking turns the empirical weak preferences into a strict profile that refines them. The GS output is stable under the strict refinement; a pair blocking under the empirical weak preferences improves strictly on both sides and therefore blocks under the refinement as well, a contradiction. Hence no pair blocks W_c under the empirical weak preferences, which is equivalent to $\bar{g}(W_c) \geq 0$. □

Lemma 15 (Exact candidate classification in the resolved regime)**.** *Assume $\mathcal{P}$ and $\varepsilon \leq h/16$. Every truly stable W has $\bar{g}(W) \geq -2\varepsilon > -4\varepsilon$, and every truly unstable W has $\bar{g}(W) \leq -h + 2\varepsilon \leq -\frac{7}{8}h < -4\varepsilon$. Hence $\bar{g}(W) > -4\varepsilon$ exactly when W is truly weakly stable; in particular, the baseline W_c is truly stable, every candidate with $\bar{g}(Z) > -4\varepsilon$ is truly stable, and every candidate with $\bar{g}(Z) \leq -4\varepsilon$ is truly unstable.*

Proof. Lemma 10 gives $\sup_W |\bar{g}(W) - g(W)| \leq 2\varepsilon$. A truly stable W has $g(W) \geq 0$ and a truly unstable W has $g(W) \leq -h$ by Definition 2, so the two displays follow from $2\varepsilon < 4\varepsilon$ and $\varepsilon \leq h/16 < 7h/32$. The equivalences are the contrapositives, and $\bar{g}(W_c) \geq 0$ by Lemma 14. □

F.5 Cumulative validation budget

Consider the validation engine and cumulative targets of Sections E.4–E.5, the candidate stream and its updates of E.6, and the commitment test and fallback-cache protocol of E.7, with $\bar{\Delta}$ always denoting the mathematical empirical optimum of Section E.1. The lemmas below analyze this policy directly; counts and the selector cursor are cumulative and persistent as specified there.

Lemma 16 (Safe early commitment)**.** *On $\mathcal{H}_T$, every early commitment has zero subsequent regret. Early checks preserve all nonnegative prefix-regret bounds for the same policy with early checks suppressed, with an additional expected cost of at most $NT\Pr(\mathcal{H}_T^c) = O(N)$ in aggregate and $O(1)$ per participant.*

Proof. Lemma 9 covers every endpoint, sample count, and adaptively chosen matching in (21), so a passing matching is truly weakly stable on $\mathcal{H}_T$. Before the first successful early check, reward actions, summaries, counts, and fallback caches coincide with those of the policy with early checks suppressed; the original test at completion of fast validation is retained. On $\mathcal{H}_T$, commitment therefore truncates such a trajectory and replaces its remaining nonnegative regret by zero. On $\mathcal{H}_T^c$, bound any additional loss by NT (or T for one participant) and apply Lemma 9. The bounds below count validation and fallback prefixes, so early commitment does not require the candidate stream to finish discovering its witnesses. □

Lemma 17 (Monotone, non-restarting targets)**.** *In the streaming branch, $\widehat{\beta}_i$ and $\widehat{h}$ are nonincreasing, b_t is increasing, and every target in (20) is nondecreasing in time. A target completed at some time is rechecked later against the larger clock, and no counter is reset.*

Proof. Every update of $\widehat{\beta}_i$ and $\widehat{h}$ is a minimum, and b_t is strictly increasing, so both the numerator $1024b_t$ of each target is nondecreasing and each denominator $\widehat{h}^2$ or $\max\{\widehat{h}, \widehat{\delta}_{i,a}\}^2$—a maximum of nonincreasing quantities—is nonincreasing. Their ratio is nondecreasing. Counts are cumulative by construction. □

Lemma 18 (Validation envelope)**.** *Let $a_\varepsilon = \min\{1, 4\varepsilon\}$, $q_{\max} = \lceil 1024 B_T / a_\varepsilon^2 \rceil$, and*

$$H_{\mathrm{val}} := N(K - N + 1)q_{\max} \le NKq_{\max}. \tag{23}$$

Throughout both branches, every target is nondecreasing in time and at most $q_{\max}$; on every observation history the total number of validation rounds is at most H_{val}, and

$$q_{\max} \le CB_T\ell, \qquad H_{\mathrm{val}} \le CNKB_T\ell.$$

Proof. A rejected candidate has $-\bar{g}(Z) \ge 4\varepsilon$, so the running margin never drops below a_ε: it starts at $1 \ge a_\varepsilon$, and if $4\varepsilon > 1$ no rejection is possible because $\bar{g}(Z) \ge -1$, so $-\bar{g}(Z) \le 1$. Hence $q_0(t), q_{i,a}(t) \le \lceil 1024 B_T / a_\varepsilon^2 \rceil = q_{\max}$, since $\max\{\widehat{h}, \widehat{\delta}_{i,a}\} \ge \widehat{h} \ge a_\varepsilon$. In the fast branch, $\bar{d}_{i,a} \ge \bar{\Delta} > 4\varepsilon$; this gate is impossible if $4\varepsilon \ge 1$, and otherwise both fast targets are at most $\lceil 1024 B_T / (4\varepsilon)^2 \rceil = q_{\max}$. Monotonicity follows from Lemma 17 in the streaming branch; the fast-branch targets are fixed.

For a fixed core matching C_s, let x_s count only the rounds on which the selector has played it as core validation. Every edge of C_s has accumulated at least x_s observations from those rounds alone. If C_s is selected again, some of its edges is still deficient, so before that play $x_s < q_0 \le q_{\max}$; hence $x_s \le q_{\max}$ at the end, even under arbitrary interleaving and increasing targets. There are N core matchings. Similarly, an outside edge is probed only while its count is below its current target, so the number of its designated outside probes is at most $q_{\max}$, and there are $N(K - N)$ outside edges. This accounting is independent of the cyclic order and of any early commitment. There is a single persistent validation selector with cumulative counts; fallback interleaving and later target increases never instantiate a fresh selector or reset previously collected samples. This proves the H_{val} bound. Finally,

$$a_\varepsilon^{-2} \le \max\left\{1, \frac{m_0}{32\gamma}\right\} \le 1 + \frac{m_0}{32\gamma} \le C\ell,$$

because $m_0 \le 2\ell$, $\gamma \ge 1$, and $\ell \ge 1$; this gives $q_{\max} \le CB_T\ell$ and $H_{\mathrm{val}} \le CNKB_T\ell$. □

F.6 One-sided accuracy and validation cost

Lemma 19 (One-sided accuracy before discovery)**.** *Assume $\mathcal{P}$ and $\varepsilon \le h/16$. At every time in the streaming branch,*

$$W_c \in \mathcal{W}^\star, \qquad \widehat{h} \ge h - 2\varepsilon \ge \tfrac{7}{8}h, \qquad \widehat{\beta}_i \ge \beta_i - \varepsilon.$$

For $s_{i,a} := (\beta_i - \mu_{i,a}^P)_+$ and every outside edge,

$$\max\{\widehat{h}, \widehat{\delta}_{i,a}\} \ge \tfrac{7}{8}\max\{h, s_{i,a}\}.$$

These conclusions require neither complete witness discovery nor an accurate upper estimate of any β_i.

Proof. By Lemma 15, every accepted candidate is truly stable, so its player payoffs are at least β_i and its clipped pilot payoff is at least $\beta_i - \varepsilon$; every rejected candidate satisfies $-\bar{g} \ge h - 2\varepsilon \ge 7h/8$, as does the initial cap 1; and the baseline is truly stable with $\bar{g}(W_c) \ge 0$. Taking running minima, including the baseline, preserves these bounds. Finally, $\widehat{\delta}_{i,a} \ge (s_{i,a} - 2\varepsilon)_+$. If $s_{i,a} \le h$, then $\max\{\widehat{h}, \widehat{\delta}_{i,a}\} \ge \widehat{h} \ge 7h/8 \ge 7s_{i,a}/8$. If $s_{i,a} > h$, then $2\varepsilon \le h/8 < s_{i,a}/8$ and $\widehat{\delta}_{i,a} \ge 7s_{i,a}/8$. □

Lemma 20 (Total streaming validation cost)**.** *Under the assumptions of Lemma 19, the aggregate regret of all streaming validation rounds up to T satisfies the pathwise bound*

$$R^{\Sigma}_{\mathrm{val}}(T) \le C\Psi(h)B_T,$$

and their total number is at most $CNKB_T/h^2$. Both statements hold for every prefix and before witness discovery.

Proof. By Lemma 19, $q_0(t) \le \lceil 1024 b_t \cdot (8/(7h))^2 \rceil \le Cb_T/h^2$. The selector accounting of Lemma 18 therefore bounds the number of core-validation rounds by $N\lceil Cb_T/h^2 \rceil$, and their aggregate cost by $CN^2 b_T/h^2$ since no round costs more than N in total. For an outside edge, at every time,

$$q_{i,a}(t) \le \left\lceil \frac{Cb_T}{\max\{h, s_{i,a}\}^2} \right\rceil.$$

The number of designated probes of that edge is at most this envelope, regardless of how often the estimates are updated or how much fallback sampling is interleaved; a probe played after further target increases is still counted against the final envelope, because targets only increase and counts are cumulative. Since the baseline is truly stable, only the moved player can have positive regret during such a probe, and its loss is $s_{i,a}$. Thus its total charge is at most

$$s_{i,a}\left\lceil \frac{Cb_T}{\max\{h, s_{i,a}\}^2} \right\rceil \le \frac{Cb_T}{h} + s_{i,a} \le \frac{C' b_T}{h},$$

using $s/\max\{h,s\}^2 \le 1/h$ for $s \ge 0$, $h > 0$, and $s \le 1$. Summing over the $N(K-N)$ outside edges proves the regret bound. Bounding every target by Cb_T/h^2 also gives the total-length bound; substituting $b_T = B_T$ gives the stated forms. The proof counts cumulative plays, never the number of target updates, so no restart multiplier appears. □

F.7 Deterministic witness discovery

Analytical witnesses. For every player fix, as a function of the true instance only,

$$W_i^- \in \arg\min_{W \in \mathcal{W}^\star} \mu^P_{i,W(i)}, \qquad \text{so that} \qquad \mu^P_{i,W_i^-(i)} = \beta_i,$$

and, if any unstable matching exists, also fix $V^- \in \arg\max_{W:\, g(W)<0} g(W)$, so that $g(V^-) = -\Delta_{\mathrm{pres}} = -h$. Remove duplicate matchings; there are at most $N+1$ of them. The algorithm does not know their identities; they are proof objects only.

Lemma 21 (Deterministic discovery of all witnesses)**.** *Let $S = |\mathcal{M}| = K!/(K-N)!$ and let τ be the index by which every analytical witness appears in the deterministic candidate stream of Section E.6. The stream visits every injection exactly once per period, so*

$$\tau \le S, \qquad \log(1+\tau) \le \log(1+S) \le CN\log(eK).$$

The algorithm never computes S and never stores the matching space; it executes only the stream prefix preceding commitment or the horizon.

Proof. The successor rule of Section E.6 advances through a fixed total order; after the last injection it wraps to the first, so the first S candidates are exactly the S distinct injections and every fixed witness appears at least once among them. Hence $\tau \le S$. Moreover $S = \prod_{j=0}^{N-1}(K-j) \le K^N$, so $\log(1+S) \le 1 + N\log K \le CN\log(eK)$. Thus the discovery bound is deterministic and needs no expectation over the stream. □

Lemma 22 (Two-sided accuracy after discovery)**.** *On $\mathcal{P}$, if $\varepsilon \le h/16$ and candidate τ is processed before commitment or the horizon, then from that update onward,*

$$|\widehat{\beta}_i - \beta_i| \le \varepsilon, \qquad |\widehat{h} - h| \le 2\varepsilon.$$

If every matching is stable, $h = 1$ and $\widehat{h} = 1$.

Proof. Lemma 19 already gives the lower bounds. Once W_i^- is emitted it is accepted, so $\widehat{\beta}_i \le \bar{\mu}^P_{i,W_i^-(i)} \le \beta_i + \varepsilon$; future minima cannot increase it. If unstable matchings exist, then $|\bar{g}(V^-) - g(V^-)| \le 2\varepsilon$ gives $h - 2\varepsilon \le -\bar{g}(V^-) \le h + 2\varepsilon$, so V^- is rejected and yields $\widehat{h} \le -\bar{g}(V^-) \le h + 2\varepsilon$, while future rejected candidates remain at least $h - 2\varepsilon$ by Lemma 15. If all matchings are stable, no candidate is rejected because $\bar{g}(W) \ge -2\varepsilon > -4\varepsilon$ for every W, so $\widehat{h}$ stays at its initial value $1 = h$ by the convention of Definition 2. □

F.8 Pointwise snapshot preservation

Lemma 23 (Pointwise snapshot preservation)**.** *Fix a reward round index t and suppose $\mathcal{P}$ and $\mathcal{G}_t$ hold, $\varepsilon \le h/16$, the witnesses of Lemma 21 have all been processed before selecting the action at t, and all current targets (20) are complete. Then a cache established at t from the current snapshot-UCB profile is truly weakly stable.*

Proof. Write $U = U(t)$ for the current snapshot-UCB scores and

$$E_0 = (\mathcal{I} \times A_c) \cup \{(i,a) : a \notin A_c,\ \widehat{\delta}_{i,a} \le \widehat{h}\}.$$

Step 1: effective radii. By Lemma 11, $a_e(t) \le 4\sqrt{b_t/n_e}$, and completeness of the targets gives $n_e \ge 1024 b_t/\widehat{h}^2$ on $\mathcal{I} \times A_c$ and on every outside edge with $\widehat{\delta}_{i,a} \le \widehat{h}$, hence $a_e(t) \le \widehat{h}/8$ there; for an outside edge not in E_0, $n_e \ge 1024 b_t/\widehat{\delta}^2_{i,a}$ and $a_e(t) \le \widehat{\delta}_{i,a}/8$.

Step 2: screening. Let $(i,a) \notin E_0$ with $a \notin A_c$. Then $\widehat{\delta}_{i,a} > \widehat{h}$, and Lemma 22 gives $2\varepsilon \le h/8 \le \widehat{h}/7 < \widehat{\delta}_{i,a}/7$ and $\beta_i - \mu^P_{i,a} \ge \widehat{\delta}_{i,a} - 2\varepsilon \ge \frac{6}{7}\widehat{\delta}_{i,a}$. Consequently, on $\mathcal{G}_t$,

$$U^P_{i,a} \le \mu^P_{i,a} + 2a_e \le \beta_i - \tfrac{6}{7}\widehat{\delta}_{i,a} + \tfrac{1}{4}\widehat{\delta}_{i,a} < \beta_i.$$

Conversely, if $\mu^P_{i,a} \ge \beta_i$ then $\widehat{\delta}_{i,a} \le 2\varepsilon < \widehat{h}$, so $(i,a) \in E_0$ unless $a \in A_c$, which is already in E_0. Thus every edge whose player mean is at least β_i lies in E_0.

Step 3: strong comparisons. If two E_0 edges incident to one endpoint have true means differing by at least h, then, using $a_e(t) \le \widehat{h}/8$ for both edges,

$$U_+ - U_- \ge h - 2a_+ - 2a_- \ge h - \tfrac{\widehat{h}}{2} \ge \tfrac{7}{16}h > 0,$$

because $\widehat{h} \le h + 2\varepsilon \le \frac{9}{8}h$. Thus the strict UCB order restricted to E_0 extends the induced order of the threshold profile π_h at that endpoint. By Lemma 13, extend these restricted orders to a complete strict profile σ that extends π_h, and let W be stable under σ.

Step 4: a reference matching stable for the full UCB profile. By Proposition 1 at threshold $h = \Delta_{\mathrm{pres}}$, W is truly weakly stable, so $\mu^P_{i,W(i)} \ge \beta_i$ and every edge of W lies in E_0 by the screening converse. An off-matching edge outside E_0 is rejected by its player under U, because $U^P_{i,a} < \beta_i \le \mu^P_{i,W(i)} \le U^P_{i,W(i)}$. For a challenger edge in E_0, every real edge appearing in its player-side or arm-side comparison also lies in E_0, and σ and U order E_0 identically, so stability under σ rules out that UCB blocker as well. Unmatched arms are last under both orders. Hence W is stable under the full current UCB profile.

Step 5: the player-proposing output is truly stable. Let M be the player-proposing UCB–GS output. Proposer optimality relative to W gives $U^P_{i,M(i)} \ge U^P_{i,W(i)} \ge \beta_i$; by the screening inequality of Step 2, no played edge can lie outside E_0 without violating this, so every played edge of M lies in E_0. If M were truly unstable, then $g(M) \le -h$, so some off-matching pair (i,a) has $c_M(i,a) \le -h$: for a matched arm both endpoints improve by at least h, and for an unmatched arm the player's improvement alone is at least h. Since the played edge is resolved and $U^P_{i,M(i)} \ge \beta_i$,

$$\mu^P_{i,M(i)} \ge U^P_{i,M(i)} - 2a_{i,M(i)} \ge \beta_i - \tfrac{\widehat{h}}{4}.$$

The blocking improvement is at least $h > \widehat{h}/4$, so $\mu^P_{i,a} > \beta_i$ and the challenger edge lies in E_0 by the screening converse. If a is matched, its incumbent edge is a played edge of M and hence lies

in E_0 as well; both strict improvements then survive in the UCB order by Step 3, producing a UCB blocker. If a is unmatched, the player's comparison survives while the arm accepts every real player. Either case contradicts UCB–GS stability.

This is a pointwise claim: it certifies each cache newly established at such a time, and makes no claim about caches established earlier. □

F.9 CACHED FALLBACK CHARGING

Lemma 24 (Cached prefix charging)**.** *Fix a deterministic cutoff u. Consider every cache that is established at a time $s \le \min\{u, T\}$, is confidence-good (that is, $\mathcal{G}_s$ holds), and is truly unstable. Count every replay round of those caches, including replays at times after u. If $\eta_T \le h/4$, these replay rounds number at most $64NKb_u/h^2$ and their aggregate regret is at most CN^2Kb_u/h^2. If $\eta_T > h/4$, the same bounds hold by the trivial horizon bound.*

Proof. If $\eta_T \le h/4$: take a charged cache established at s, computing the player-proposing UCB–GS matching M for the snapshot scores $U(s)$. Since M is truly unstable, $g(M) \le -h$, so some off-matching pair (i, a) has $c_M(i, a) \le -h$: for a matched arm both endpoints improve by at least h, and for an unmatched arm the player's improvement alone is at least h. Under the strict U-profile, M has no blocking pair, so at least one endpoint of (i, a) does not strictly prefer the challenger. If the player does not, then $U^P_{i,M(i)}(s) \ge U^P_{i,a}(s)$; on $\mathcal{G}_s$ the two-sided coverage gives $U^P_{i,M(i)} \le \mu^P_{i,M(i)} + 2a_e(s)$ and $U^P_{i,a} \ge \mu^P_{i,a}$ for the played incumbent edge $e = (i, M(i))$, so $2a_e(s) \ge \mu^P_{i,a} - \mu^P_{i,M(i)} \ge h$. If a is matched and the arm does not strictly prefer, the same argument at the played edge $e = (j, a)$, $j = M^{-1}(a)$, gives $2a_e(s) \ge h$. If a is unmatched, the arm ranks every real player above unmatched and cannot be the rejecting endpoint, so the player case applies. Charge the cache to the played edge e.

Since $2a_e(s) \ge h$ and $\eta_T \le h/4$, we get $\sqrt{2b_s/m_e(s)} \ge h/4$ and hence

$$m_e(s) \le \frac{32b_s}{h^2} \le \frac{32b_u}{h^2}.$$

While the cache remains valid and is replayed, no snapshot update has occurred since s: an update would change the public snapshot version, and the first subsequent round at which the selector returns COMPLETE refreshes the cache before any further play, so the stale cache is never replayed again. Hence $m_e(t) = m_e(s)$ at every replay time t of this cache, and by Lemma 11, $n_e(t) < 2m_e(t) \le 64b_u/h^2$.

Every replay of the cache samples the edge e. Over all charged caches, consider the charged replay times of a fixed edge. Each such time samples the edge, and all of them occur while $n_e < 64b_u/h^2$; two distinct charged times for the same edge have distinct values of n_e, because the earlier one increments n_e. There are fewer than $64b_u/h^2$ nonnegative integers below that bound, so one edge receives at most $64b_u/h^2$ charged replay rounds. Summing over the NK edges gives at most $64NKb_u/h^2$ rounds. Every charged replay plays a truly unstable matching, whose aggregate player regret is at most N, so their aggregate regret is at most CN^2Kb_u/h^2.

This count includes time-boundary refreshes, snapshot refreshes, validation pauses, multiple caches charged to the same edge, and replay after the cutoff u: refreshes and pauses create no additional charged rounds, and every cache sharing an edge is charged to that same edge.

If $\eta_T > h/4$, then $h < 4(T+1)^{-2}$, so $(T+1) < 2/\sqrt{h}$ and the trivial bound $NT \le 2N/\sqrt{h} \le CN^2K/h^2 \le CN^2Kb_u/h^2$ on all rounds of the horizon proves the claim. □

Lemma 25 (Confidence-failure caches)**.** *A cache established at time s is invalidated no later than the next dyadic time boundary, so its lifetime spans at most $2^{1+\lfloor \log_2 s \rfloor} - s \le s$ reward-round indices. The expected aggregate regret of all replay rounds of confidence-failure caches is $O(N)$, and their expected regret to any fixed participant is $O(1)$.*

Proof. The refresh rule compares $\lfloor \log_2 t \rfloor$ with $\kappa_s = \lfloor \log_2 s \rfloor$ at every round where the selector returns COMPLETE, so no replay occurs after the first boundary, which lies at index $2^{\kappa_s+1} \le 2s$. Bound the total regret by charging every round that replays a cache established at s with a failed event

$\mathcal{G}_s^c$ a cost of at most N in aggregate and 1 per participant. The number of rounds that can replay the cache of s is at most s, and the event that a cache is established at s together with $\mathcal{G}_s^c$ has probability at most $\Pr(\mathcal{G}_s^c) \le s/[2(s+1)^5]$, because both events are determined by the history before round s and their intersection is contained in $\mathcal{G}_s^c$. Summing the deterministic envelope,

$$\sum_{s \le T} N \cdot s \cdot \frac{s}{2(s+1)^5} \le CN \sum_{s \ge 1} \frac{1}{s^3} = O(N),$$

and the per-participant bound replaces N by 1. Confidence-failure damage is thus controlled by dyadic invalidation rather than by a per-round count. □

Lemma 26 (Witness waiting bound)**.** *Assume $\varepsilon \le h/16$. On histories where $\mathcal{P}$ holds and the streaming branch is selected, every confidence-good truly unstable fallback round belongs to a cache established at some time $s < u := t_0 + \tau$, where $t_0 = Km_0$ and τ is the deterministic discovery index. Consequently, pathwise,*

$$R^{\Sigma}_{\text{waiting}} \le \frac{CN^2 K b_u}{h^2}, \qquad \textit{and since} \qquad b_u \le C\big[N \log(eNK) + \log \ell\big], \tag{24}$$

also $R^{\Sigma}_{\text{waiting}} \le C\frac{N^2K}{h^2}[N \log(eNK) + \log \ell]$.

Proof. A fallback round occurs only when the selector returns COMPLETE, so the cache it replays was established at some time s at which all current targets were complete. If $s \ge u$, then candidate step $s - t_0 \ge \tau$ has already been processed, so $\mathcal{P}$, $\mathcal{G}_s$, $\varepsilon \le h/16$, and full witness discovery hold at s and Lemma 23 makes the cache truly stable. Hence every confidence-good truly unstable cache is established at some $s < u$, and all of their replays—including replays after u—are counted by Lemma 24 with cutoff u. This gives the first display. For the second, $u \le 2K\ell + K^N + 1$ by Lemma 21 and $t_0 = Km_0 \le 2K\ell$, so

$$\log(u+1) \le C + \log(2K\ell) + \log(1 + K^N) \le C\big[1 + \log \ell + N \log K\big],$$

and $b_u = \log(8NK) + 5\log(u+1) \le C[N \log(eNK) + \log \ell]$. □

F.10 FAST-PATH AND UNRESOLVED-RESOLUTION ESTIMATES

Lemma 27 (Fast validation on an accurate frozen pilot)**.** *On $\mathcal{P}$, if the fast gate is taken, then W_c is truly stable and*

$$g(W_c) > \tfrac{1}{2}\bar{\Delta}, \qquad \bar{\Delta} > \tfrac{2}{3}\Delta.$$

At completion of the targets, on $\mathcal{H}_T$ the matching W_c passes (21): the core snapshot radius is at most $\bar{\Delta}/8$ and every outside snapshot radius is at most $\bar{d}_{i,a}/8$. Every fast-validation prefix has aggregate regret at most $C\Psi(\Delta)B_T$.

Proof. Lemma 10 together with $\bar{\Delta} > 4\varepsilon$ gives $g(W_c) \ge \bar{\Delta} - 2\varepsilon > \frac{1}{2}\bar{\Delta} > 0$ and $\Delta \le \bar{\Delta} + 2\varepsilon < \frac{3}{2}\bar{\Delta}$. For an outside edge write $d_{i,a} = \mu^P_{i,W_c(i)} - \mu^P_{i,a}$. Since $\bar{d}_{i,a} \ge \bar{\Delta} > 4\varepsilon$,

$$\tfrac{1}{2}\bar{d}_{i,a} < d_{i,a} < \tfrac{3}{2}\bar{d}_{i,a}, \qquad \bar{d}_{i,a} \ge \bar{\Delta} > \tfrac{2}{3}\Delta.$$

After the targets complete, Lemma 11 and the target sizes give $a^*_e \le 4\sqrt{B_T/n_e} \le \bar{\Delta}/8$ on core edges and $a^*_e \le \bar{d}_{i,a}/8$ on outside edges. For a core off-matching pair, one side has true rejection strictly greater than $\bar{\Delta}/2$; on $\mathcal{H}_T$ its lower-minus-upper difference is at least that rejection minus twice the sum of the two radii, which is at most $\bar{\Delta}/2$, so the difference is strictly positive. For an outside pair the player's rejection satisfies $d_{i,a} > \bar{d}_{i,a}/2$, while twice the radii sum is at most $(\bar{\Delta} + \bar{d}_{i,a})/4 \le \bar{d}_{i,a}/2$, so the player test also passes. Thus W_c passes every comparison when the targets complete.

Core rounds number at most Nq_0^+ and cost at most N each, giving $CN^2B_T/\bar{\Delta}^2 \le CN^2B_T/\Delta^2$. In an outside probe only the moved player can lose, because every other player keeps a truly stable baseline partner; its shortfall $s_{i,a} = (\beta_i - \mu^P_{i,a})_+$ satisfies $s_{i,a} \le d_{i,a}$. Its full charge is at most

$$s_{i,a}\left(1024B_T/\bar{d}^2_{i,a} + 1\right) \le CB_T/\bar{d}_{i,a} + 1 \le C'B_T/\Delta + 1,$$

summing over the $N(K-N)$ outside edges gives at most $CB_T \cdot N(K-N)/\Delta + N(K-N) \le C\Psi(\Delta)B_T$ because $B_T \ge 1$, and every charge also bounds truncated prefixes. □

Lemma 28 (Resolution inversion with quantization)**.** *For $0 < u \le 1$, if $\varepsilon > u/16$, then*

$$\ell,\ B_T \le Cu^{-2}\log\frac{eNK}{u}, \qquad \textit{and} \qquad \min\{1, 4\varepsilon\} \ge u/4.$$

In that case $a_\varepsilon \ge u/4$, and all validation targets are at most CB_T/u^2.

Proof. The inequality $\min\{1, 4\varepsilon\} \ge u/4$ is immediate from $4\varepsilon > u/4$ and $u \le 1$. For the horizon bound, split into two cases. Case 1: $\eta_T \le u/32$. Then $\sqrt{2\gamma/m_0} = \varepsilon - \eta_T > u/32$, so $m_0 < 2048\gamma/u^2$. Put $A = 2048/u^2$, $c_0 = \log(4(NK)^3)$, and $y = \ell/A$. Since $\gamma = c_0 + 8\log\ell$ and $\ell \le m_0$,

$$y < \gamma = c_0 + 8\log A + 8\log y.$$

The elementary bound $\log y \le y/16 + \log 16 - 1$ implies $y \le 2c_0 + 16\log A + C$, which proves the bound on ℓ because $\log A \le C\log(e/u)$; since $B_T \le \log(8NK) + 5\ell$, the same bound holds for B_T, and $\log(8NK) \le C\log(eNK/u)$ is absorbed. Case 2: $\eta_T > u/32$. Then $u < 32(T+1)^{-2}$, so $T + 1 < \sqrt{32}\,u^{-1/2}$ and $T \le \sqrt{32}\,u^{-1/2} \le \sqrt{32}\,u^{-2}$; hence $\ell = \log(eT) \le eT \le Cu^{-2} \le Cu^{-2}\log(eNK/u)$ and similarly for B_T, so the displayed bound is trivial. Finally, $a_\varepsilon \ge u/4$, so Lemma 18 bounds every target by CB_T/u^2 since $B_T \ge 1$. □

Failure events are never charged an entire horizon. Let $F_T := CN^2KB_T/h^2$ denote the deterministic envelope of Lemma 24 at cutoff T. The validation envelope and the pilot-failure probability give

$$NH_{\mathrm{val}}p_T \le \frac{CB_T}{K\ell^7}, \qquad F_Tp_T \le \frac{CB_T}{Kh^2\ell^8}. \tag{25}$$

These inequalities require no independence: on any pilot-failure history, the corresponding costs obey the stated deterministic envelopes. Confidence-failure replays are charged by Lemma 25 and contribute $O(N)$ expected aggregate regret, and incorrect permanent commitment can occur only on $\mathcal{H}_T^c$, where it costs at most NT and has probability at most $T/[2(T+1)^5]$, so its expected contribution is $O(N)$. This is why no T-times-pilot-failure term appears.

F.11 Proof of Theorem 3

If the horizon ends during the pilot, aggregate regret is at most $NT \le NKm_0 \le 2NK\ell \le 2\Psi(u)\ell$ for every $0 < u \le 1$, and any participant loses at most $T \le 2K\ell$; all claims follow. Assume below that the pilot completes; its aggregate cost is at most $2NK\ell$. By Lemma 16, it suffices to analyze the same policy with early checks suppressed and the fast branch's final test retained: restoring the checks adds at most $O(N)$ expected aggregate regret and $O(1)$ per participant, absorbed in the bounds below.

Universal individual safety. Put $\delta = \delta_{\mathrm{safe}}$. First assume $\varepsilon \le \delta/16$. On $\mathcal{P}$, if $\Delta > 0$ then $\bar{\Delta} \ge \Delta - 2\varepsilon \ge 7\Delta/8 > 4\varepsilon$, so the fast branch is chosen and the gate shortcut cannot apply; its targets are $O(B_T/\Delta^2) \le O(B_T/\delta^2)$ and its validation length is $O(NKB_T/\delta^2)$ by the envelope. If $\Delta = 0$, then $\bar{\Delta} \le 2\varepsilon$ and the streaming branch is chosen; Lemma 19 bounds all of its targets by $CB_T/\Delta_{\mathrm{pres}}^2 = CB_T/\delta^2$, with the same length bound. On $\mathcal{P}^c$, any participant's validation cost is at most H_{val}, whose expectation is at most $H_{\mathrm{val}}p_T \le CB_T/(NK\ell^7)$. In either branch, the total confidence-good fallback regret to a fixed participant is at most $CNKB_T/\Delta_{\mathrm{pres}}^2 \le CNKB_T/\delta^2$: apply Lemma 24 with cutoff T in both branches to bound charged replays by $64NKB_T/h^2$, and use a loss of at most 1 per replay; confidence-failure replays contribute a further $O(1)$ by Lemma 25. Correct commitment costs nothing, incorrect commitment costs at most T and occurs with probability at most $T/[2(T+1)^5]$, well under 1 in expectation, and the pilot costs at most $2K\ell$. Combining gives (4). If $\varepsilon > \delta/16$, every validation target is at most CB_T/δ^2 by Lemma 28, deterministically; the same fallback, commitment, and pilot arguments apply. This proves (4) for all horizons and yields the coarse aggregate bound

$$\mathbb{E}R_P^{\Sigma}(T) \le \frac{CN^2KB_T}{\delta_{\mathrm{safe}}^2}. \tag{26}$$

Zero certificate margin. Assume $\Delta = 0$. If $\varepsilon > \Delta_{\mathrm{pres}}/16$, then (26) and Lemma 28 give $\mathbb{E}R_P^{\Sigma}(T) \le CN^2KB_T/\Delta_{\mathrm{pres}}^2 \le C\mathcal{L}_\star$, which proves the claim in the unresolved case. Suppose $\varepsilon \le \Delta_{\mathrm{pres}}/16$. On $\mathcal{P}$, $\bar{\Delta} \le 2\varepsilon$ and streaming starts after the pilot; Lemma 20 bounds every validation prefix by $C\Psi(\Delta_{\mathrm{pres}})B_T$, and Lemma 26 bounds the good fallback replays; on $\mathcal{P}^c$ use (25), and charge confidence-failure replays by Lemma 25. Altogether, with $h = \Delta_{\mathrm{pres}}$,

$$\mathbb{E}R_P^{\Sigma}(T) \le 2NK\ell + C\Psi(h)B_T + \frac{CN^2K}{h^2}\big[N\log(eNK) + \log\ell\big] + \frac{CB_T}{K\ell^7} + \frac{CB_T}{Kh^2\ell^8} + CN.$$

Now $B_T \le 5\ell + \log(8NK)$, and the absorption inequality $\log\ell \le \ell/K + \log(eK)$ gives

$$\frac{N^2K}{h^2}\log\ell \le \frac{N^2}{h^2}\ell + \frac{N^2K}{h^2}\log(eK).$$

Using $\Delta_{\mathrm{pres}} \le 1$ and $\Psi(h) \le 2N^2K/h^2$, every term that does not contain ℓ at the first power is bounded by

$$C\left(\frac{N^3K}{\Delta_{\mathrm{pres}}^2} + \frac{N^2K}{\Delta_{\mathrm{pres}}^4}\right)\log\frac{eNK}{\Delta_{\mathrm{pres}}} = C\mathcal{L}_\star,$$

while the remaining ℓ terms are at most $C\Psi(\Delta_{\mathrm{pres}})\ell$. This proves the $\Delta = 0$ bounds of (3); in particular, the deterministic candidate stream adds no extra factor of $\log T$ to the leading core–tail coefficient.

Positive certificate margin. Assume $\Delta > 0$. First suppose $\varepsilon \le \Delta/16$. On $\mathcal{P}$, the exact search selects the fast branch, and Lemma 27 bounds its validation prefixes by $C\Psi(\Delta)B_T$. On $\mathcal{H}_T$, a completed validation passes every test, so all remaining rounds are stable; a horizon-truncated validation is already covered by the prefix bound. On $\mathcal{P}^c$, use the deterministic validation envelope and the fallback bound; on $\mathcal{H}_T^c$, charge at most NT. Thus

$$\mathbb{E}R_P^{\Sigma}(T) \le 2NK\ell + C\Psi(\Delta)B_T + \frac{CB_T}{K\ell^7} + \frac{CB_T}{K\Delta_{\mathrm{pres}}^2\ell^8} + CN.$$

The $\log(8NK)$ part of B_T, both pilot-failure terms, and CN are absorbed by $C\mathcal{L}_\star$, using $\Delta, \Delta_{\mathrm{pres}} \le 1$ and $\Psi(\Delta) \le 2N^2K/\Delta^2$; the remaining terms are $C\Psi(\Delta)\ell$. This proves the $\Delta > 0$ bounds of (3) for $\varepsilon \le \Delta/16$. If $\varepsilon > \Delta/16$, the deterministic validation-target envelope CB_T/Δ^2 holds regardless of which branch the observed pilot selects, so validation costs at most CN^2KB_T/Δ^2 and good fallback rounds cost at most $CN^2KB_T/\Delta_{\mathrm{pres}}^2$. Adding the pilot and the $O(N)$ expected confidence-error costs gives

$$\mathbb{E}R_P^{\Sigma}(T) \le CN^2KB_T\left(\frac{1}{\Delta^2} + \frac{1}{\Delta_{\mathrm{pres}}^2}\right) + CN,$$

which Lemma 28 with $u = \Delta$ bounds by $C\mathcal{L}_\star$, because $1/\Delta_{\mathrm{pres}}^2 \le 1/\Delta^4 + 1/(\Delta_{\mathrm{pres}}^2\Delta^2)$ when $\Delta \le 1$. This proves the $\Delta > 0$ bounds of (3) in all cases. For every fixed market, Δ (when positive), Δ_{pres}, N, and K are constants, so all remainders are independent of T; dividing the finite-time bounds by $\log T$ and taking the limit superior proves the two asymptotic statements and completes the proof of Theorem 3.

G Paid Local Communication: Protocol and Proofs

This appendix gives the implementation used in Theorem 4. We call its communication-compatible statistical policy *Quantized ACTL* (ACTL^q): it retains ACTL's structural logic but changes quantization and validation constants, so it is not a physical-time copy of the original policy. The proof has three layers. Sections G.2–G.4 construct a shared virtual state with logarithmically many expected source bits; Section G.5 re-establishes its statistical regret bound; and Section G.6 charges communication slots to physical regret.

G.1 Physical time, local channels, and regret

The physical horizon is T slots; t denotes elapsed physical time and s counts only retained learning slots. A retained slot plays an injection in $\mathcal{M}$, and its endpoint observations enter ACTL^q. A

synchronization slot plays an injection or, for in-band signaling only, a partial assignment Q. A player may then be unmatched; set $\mu^P_{i,\perp} = 0$ for that slot's regret. Synchronization rewards are excluded from learning sums and counts, but their loss is charged to $R^{\Sigma}_P(T)$. All conditional sub-Gaussian assumptions in Section 3 hold given the complete physical history. Let

$$r_P(Q) := \sum_i (\beta_i - \mu^P_{i,Q(i)})_+, \qquad 0 \le r_P(Q) \le N.$$

The nodes are all N players and K arms. They have public labels, a common slot clock, reliable local controls, cooperative behavior, and deterministic code. We separate latency from disruption: λ is the number of physical slots needed to deliver one logical bit, while κ bounds the extra aggregate regret per signaling slot beyond playing a baseline matching. Formally, a source known from the common state sends one logical bit in exactly λ slots while each slot has $r_P(Q) \le r_P(W) + \kappa$ relative to its prescribed complete baseline W. The source, message order, and duration do not require a free side message. A dedicated all-node broadcast link, if present, is the paid special case $(\lambda, \kappa) = (1, 0)$.

A *record* encodes one statistical update as several logical source bits. A source bit can take multiple physical slots to disseminate; on a graph, each slot can carry many point-to-point link bits in parallel. These counts are kept separate below.

Connected graph realization. Let G be an undirected connected graph on all nodes, with one bit per directed neighbor link per slot and a known diameter upper bound D (the universal choice $N + K - 1$ suffices). Matching service runs concurrently. In each two-slot frame every node sends to every neighbor one of the symbols 00 (uninformed), 01 (known zero), or 10 (known one). At the end of a frame a previously uninformed node accepts an informed neighbor's bit. Induction on graph distance shows delivery within D frames. Padding makes $\lambda = 2D$ exact; every slot plays W, so $\kappa = 0$. Each source bit uses at most $4|E(G)|D$ directed link bits. Neither global topology nor a leader is required.

In-band realization. Here a player may propose to an arm or abstain, an arm observes incoming proposals and may accept one, and a player sees acceptance or rejection. These control observations are noiseless and distinct from rewards. To send a bit between player i and arm a, if $a = W(j)$ for $j \ne i$, park j at $W(i)$; all other players except i keep their partners. The parking choice is independent of the bit. For player-to-arm transmission, i proposes to a for one and abstains for zero. For arm-to-player transmission, i always proposes to a and a accepts for one or rejects for zero. There are no conflicting proposals, and at most two players differ from W, so $\kappa = 2$. A player source sends to all K arms, then the smallest-label informed arm relays to the other $N - 1$ players. An arm source uses the symmetric route. Both schedules last $\lambda = N + K - 1$ slots. Thus no auxiliary data channel or observation of another node's reward or winning player is assumed.

Theorem 5 (Finite-time LC-ACTL bound)**.** *Let $h_\star$ and $\mathcal{L}_\star$ be as in Theorem 3, $\ell = \log(eT)$, and $F(u) = u^{-4}\log^2(eNK/u)$. Under the communication interface above, LC-ACTL uses an expected $O(NK\ell)$ logical source bits and $O(\lambda NK\ell)$ physical communication slots, and satisfies*

$$\mathbb{E}R^{\Sigma}_P(T) \le C[\Psi(h_\star) + \kappa\lambda NK]\ell + C\{\mathcal{L}_\star + \lambda N^2 K[F(h_\star) + \log(e\lambda N)]\}. \tag{27}$$

For each fixed market and realization, the logarithmic coefficient is at most $C[\Psi(h_\star) + \kappa\lambda NK]$. The universal C and horizon-independent remainder have no $\Delta_{\min}$ dependence.

G.2 Quantized statistical core and pilot

Set

$$\begin{aligned} &\ell = \log(eT), \quad m_0 = \lceil \ell \rceil, \quad \gamma = \log(4(NK)^3\ell^8), \\ &\varepsilon = \sqrt{2\gamma/m_0} + m_0^{-1}, \quad b_s = \log(8NK(s+1)^5), \quad B_T = b_T. \end{aligned} \tag{28}$$

The pilot cycles through the K matchings $P_r(i) = 1 + ((i - 1 + r) \bmod K)$, taking m_0 retained samples per edge. If physical time runs out, the process stops. Otherwise each endpoint clips its pilot mean to $[0, 1]$ and sends the integer numerator of

$$\bar{\mu}_e = m_0^{-1} \left\lfloor m_0 \, \mathrm{clip}_{[0,1]}(\widehat{\mu}_{e,m_0}) \right\rfloor. \tag{29}$$

The numerator takes $L_0 = \lceil \log_2(m_0 + 1) \rceil$ bits, so initial synchronization uses exactly $2NKL_0$ logical bits. The initial parking injection is $W_{\mathrm{park}}(i) = i$; no stability property is asserted for it. After delivery, every node has the same frozen pilot matrix, retained counts $n_e = m_e = m_0$, snapshots $v_e = \bar{\mu}_e$, and error allowances $\zeta_e = m_0^{-1}$. Raw rewards never leave their endpoints.

ACTL^q uses the deterministic search, mode initialization, structural iterator, validation selector, certificate tests, and fallback cache of Appendix E, with four explicit changes: the pilot grid (29) and radius (28); the snapshot rule below; $A_0 = 4096$ in place of every target constant 1024 in (19) and (20); and, after commitment, repetition of the chosen matching with unused statistical state frozen. All decision clocks and dyadic cache keys use the virtual index s. In CERTIFY, W_c is the empirical maximum-certificate matching with $\bar{\Delta} > 4\varepsilon$ and the fixed targets use $A_0 B_T$. In PRESERVE, W_c is the label-tie-broken pilot GS matching, $\widehat{\beta}_i = \bar{\mu}^P_{i,W_c(i)}$, $\widehat{h} = 1$, and the adaptive targets use $A_0 b_s$. A candidate Z is processed once per uncommitted retained decision, never because a communication slot passes. This defines a centralized ACTL^q on any admissible retained reward stream.

G.3 Differential snapshots and exact decoding

At a retained count doubling $n_e = 2m_e =: m$ after virtual slot s, define public grid and endpoint integer

$$j(s,m) = \max\{0, \lfloor \tfrac{1}{2} \log_2(m/b_s) \rfloor\}, \quad \eta(s,m) = 2^{-j(s,m)}, \quad k_e = \left\lfloor \mathrm{clip}_{[0,1]}(\widehat{\mu}_{e,m})/\eta(s,m) \right\rfloor. \tag{30}$$

The new snapshot is $v_e = \eta k_e$, with $\zeta_e = \eta$. Given the old public value v_e^{old}, every node computes $k_e^{\mathrm{pred}} = \lfloor v_e^{\mathrm{old}}/\eta \rfloor$. The endpoint alone computes $d_e = k_e - k_e^{\mathrm{pred}}$. For $|d_e| \le 16$, it disseminates flag 0 followed by the six-bit integer $d_e + 16$; otherwise it sends flag 1 followed by the $(j+1)$-bit absolute integer $k_e \in \{0, \dots, 2^j\}$. Thus the ordinary record is seven bits and the escape record is exact. All nodes infer source order, edge identity, grid, and record length from their common state and the flag; there are no uncharged headers.

At a due update, endpoints are processed in player, arm, endpoint order. Each complete record is disseminated through the paid interface using W_c, with the virtual state frozen. After every due record is received, all nodes atomically install the staged snapshots, set $m_e = n_e$, and increment the common snapshot version once. If T interrupts a record, no further learning action begins. At virtual decision times use

$$\begin{aligned} a_e(s) &= \sqrt{2b_s/m_e} + \zeta_e, \qquad a_e^\star = \sqrt{2B_T/m_e} + \zeta_e, \\ U_e(s) &= v_e + a_e(s), \qquad [L_e^\star, U_e^\star] = [v_e - a_e^\star, v_e + a_e^\star]. \end{aligned} \tag{31}$$

The commitment test is exactly (21) with these intervals.

G.4 Concentration, bit complexity, and state synchronization

Let $\mathcal{P}$ mean that every pilot value differs from its true mean by at most ε; let $\mathcal{H}_T$ cover all reached retained endpoint means by $\sqrt{2B_T/m}$; let $\mathcal{G}_s$ similarly cover reached counts up to virtual time s by $\sqrt{2b_s/m}$. Conditional concentration at adaptively reached counts and a union bound give

$$\Pr(\mathcal{P}^c) \le p_T := ((NK)^2 \ell^8)^{-1}, \quad \Pr(\mathcal{H}_T^c) \le T/[2(T+1)^5], \quad \Pr(\mathcal{G}_s^c) \le s/[2(s+1)^5]. \tag{32}$$

The pilot grid contributes at most m_0^{-1} error. At every virtual decision, all due updates have completed and $m_e \le n_e < 2m_e$. For $z = \sqrt{b_s/m} < 1$, the new grid obeys $z \le \eta(s,m) < 2z$; otherwise $\eta = 1$. Hence $\zeta_e \le 2\sqrt{b_s/m_e}$ and

$$a_e(s) \le 6\sqrt{b_s/n_e}, \qquad a_e^\star \le 6\sqrt{B_T/n_e}. \tag{33}$$

On $\mathcal{G}_s$, $U_e(s) \ge \mu_e$ and its excess is at most $2a_e(s)$; on $\mathcal{H}_T$, every commitment interval covers its true mean.

The next lemma gives the two properties needed for decentralization: expected communication is logarithmic, and synchronization changes physical time without changing the retained statistical trajectory.

Algorithm 3 Local-Communication ACTL (LC-ACTL)

Require: N, K, T, public labels and clocks, and a baseline-respecting (λ, κ) realization

Set physical clock $t = 0$ and virtual clock $s = 0$; perform the cyclic pilot as retained slots, advancing both clocks; stop at $t = T$.
Using W_{park}, disseminate every pilot numerator (29) in fixed order; advance only t and discard these slots' rewards; stop at $t = T$.
Every node initializes the same ACTL^q mode, W_c, targets, candidate iterator, selector cursor, tests, and empty fallback cache.
while $t < T$ **do**
    if committed **then**
        Play the committed injection and advance t, s; leave the unused statistical state frozen.
    else
        Compute the next ACTL^q action from the common state and play it for one retained slot; advance t, s.
        Keep local endpoint rewards and update the common edge counts.
        For each due doubling, disseminate the differential record (30) with baseline W_c; advance only t, ignore these slots' rewards, and atomically install completed snapshots.
    end if
end while
Stop immediately at T, including inside pilot or synchronization; no partially decoded value is used.

Lemma 29 (Compression and retained-time synchronization)**.** *Let $B(T)$ count logical source bits, including a complete record if T interrupts its transmission. Then $\mathbb{E}B(T) \le CNK\ell$ and $B(T) \le CNK\ell^2$ on every history. After deleting synchronization slots, the retained action sequence is a prefix of ACTL^q on the same retained observations. This is a pathwise statement about the virtual sequence, not equality with the original ACTL's physical sequence.*

Proof. At a doubling, the new and old clipped empirical means differ by less than $7z$ on $\mathcal{G}_s$, using the two concentration radii and the old quantization error. The floor map then gives $|d_e| \le 8 < 16$; thus an escape at virtual time s implies $\mathcal{G}_s^c$. Each endpoint has at most $\lfloor \log_2(T/m_0) \rfloor$ updates, giving $2NKL_0 + 14NK \log_2(T/m_0) = O(NK\ell)$ ordinary bits. An escape adds $O(\log(es))$ bits. At most $2N$ endpoint updates occur after a retained slot, so (32) bounds expected extra bits by $CN \sum_{s \ge 1} s \log(es)/(s+1)^5 = O(N)$. Charging $O(\ell)$ bits to every update gives the deterministic bound. The convention about an interrupted record changes neither order.

For the coupling, pilot actions are label-determined and pilot integers are decoded exactly. Assume every node and ACTL^q agree before a retained slot. Deterministic candidate updates, certificate tests, selector, GS, cursor, and cache then choose the same injection without a globally observed realized assignment. Its endpoint observations and inferred count increments agree. Every due integer is decoded exactly, including escapes; the virtual state is frozen during transmission and all snapshots are installed together. This proves the induction. Commitment repeats a common injection without further communication. Truncation simply shortens the prefix. □

The expected physical synchronization time is $\lambda \mathbb{E}B(T) \le C\lambda NK\ell$. The logical-bit count is distinct from the graph's expected $O(|E(G)|DNK\ell)$ total directed link-bit traffic.

G.5 Statistical robustness of ACTL^q

Because ACTL^q changes quantization and validation constants, its statistical guarantee does not follow from the centralized pathwise argument. We verify directly that the concentration, validation, preservation, and fallback arguments of Appendix F remain valid up to universal constants. Put $h = \Delta_{\mathrm{pres}}$ and δ_{safe} as in (4).

Lemma 30 (Quantized-core guarantee)**.** *For every admissible retained reward process, ACTL^q satisfies*

$$\mathbb{E}R^{\Sigma}_{P,\mathrm{ACTL}^q}(T) \le C\Psi(h_\star)\ell + C\mathcal{L}_\star, \qquad \max\{\max_i \mathbb{E}R_i(T), \max_a \mathbb{E}R^A_a(T)\} \le CNKB_T/\delta^2_{\mathrm{safe}}.$$

On $\mathcal{P}$, if $\varepsilon \le h_\star/16$, its fixed W_c is truly weakly stable.

Proof. By (33), every proof in Appendix F that requires $a_e = O(\sqrt{b_s/n_e})$ remains valid after replacing its constant by 6 and its target constant by 4096. For clarity, we spell out the margin-dependent steps and failure charges.

Certificate values change by at most 2ε under the pilot perturbation. If $\Delta > 0$, $\varepsilon \le \Delta/16$, and $\mathcal{P}$ holds, then $\bar{\Delta} > 4\varepsilon$, $g(W_c) \ge \bar{\Delta} - 2\varepsilon > 0$, and the CERTIFY targets make all commitment comparisons strict on $\mathcal{H}_T$. The core rotations cost $O(N^2 B_T/\Delta^2)$; each outside probe costs its baseline shortfall and yields the $O(N(K-N)B_T/\Delta)$ charge.

If $\Delta = 0$, $\varepsilon \le h/16$, and $\mathcal{P}$ holds, then $\bar{\Delta} \le 2\varepsilon$, so PRESERVE is selected. The empirical GS baseline is stable: an unstable matching has $\bar{g} \le -h + 2\varepsilon < -4\varepsilon$, whereas empirical GS has $\bar{g}(W_c) \ge 0$. The structural stream therefore classifies every candidate correctly. Before its witnesses appear, $\widehat{h} \ge 7h/8$, $\widehat{\beta}_i \ge \beta_i - \varepsilon$, and each target denominator is at least a constant multiple of $\max\{h, (\beta_i - \mu^P_{i,a})_+\}$. Cumulative validation costs $O(\Psi(h)B_T)$. At most $N+1$ analytical witnesses are visited by index $\tau \le K!/(K-N)!$; thereafter $|\widehat{\beta}_i - \beta_i| \le \varepsilon$ and $|\widehat{h} - h| \le 2\varepsilon$.

The order-theoretic step is unchanged: a matching is stable in some strict extension of the comparisons of true size at least h exactly when $g(W) > -h$ (Proposition 1). At $h = \Delta_{\mathrm{pres}}$, these are exactly the truly stable matchings. Completed targets and (33) make snapshot radii at most $\widehat{h}/8$ on core and unresolved outside edges, and at most the estimated shortfall divided by 8 on screened outside edges. The latter have $U^P_{i,a} < \beta_i$; all comparisons of size at least h on the former survive the UCB order. Extending that order while preserving the threshold comparisons gives a truly stable reference matching. Player-proposing UCB–GS gives every player a UCB payoff at least that reference payoff, so its played edges cannot be screened. A true blocker would then remain a UCB blocker, a contradiction. Thus every newly established confidence-good fallback cache after discovery is stable; no earlier cache is asserted to become stable retroactively.

Before discovery, an unstable cache established at virtual time s has a played edge with effective radius at least $h/2$, so its snapshot count is $O(b_s/h^2)$. Replaying it samples that edge until doubling or the dyadic cache refresh. Distinct charged replays on an edge have distinct retained counts, yielding $O(N^2 K b_u/h^2)$ aggregate loss for caches established by any cutoff u. Confidence-failure caches cost $O(N)$ in expectation because one established at s can be replayed at most s times and $\sum_s s \Pr(\mathcal{G}^c_s) < \infty$; incorrect commitment costs at most $NT\Pr(\mathcal{H}^c_T) = O(N)$.

On every history, each cumulative target is at most $O(B_T\ell)$: rejected candidate margins are at least 4ε, the CERTIFY gate also exceeds 4ε, and $\min\{1, 4\varepsilon\}^{-2} = O(\ell)$. Hence validation uses at most $CNKB_T\ell$ slots even on $\mathcal{P}^c$. Multiplying this finite envelope, rather than T, by p_T controls pilot failures. If $\varepsilon > u/16$, solving (28) gives $\ell, B_T \le Cu^{-2}\log(eNK/u)$; this resolution inversion absorbs the unresolved branch into $\mathcal{L}_\star$ with $u = h_\star$. Combining these charges exactly as in the two regimes of Theorem 3 gives the displayed aggregate and individual bounds. A pilot truncated at T costs at most $NKm_0 \le 2NK\ell$. □

G.6 PHYSICAL-TIME REGRET DECOMPOSITION AND THEOREM

We separate statistical learning from communication: retained slots inherit ACTL^q's bound, initial dissemination is charged directly, and later communication repeats the baseline with at most κ extra loss per slot. Let C_{post} be post-pilot communication slots and R_{init} the pilot dissemination loss. On every history,

$$R^{\Sigma}_P(T) \le R_{\mathrm{active}}(T) + R_{\mathrm{init}}(T) + C_{\mathrm{post}} r_P(W_c) + \kappa C_{\mathrm{post}}. \tag{34}$$

Here active regret counts retained slots only. The initial phase uses $2NKL_0$ source bits and has

$$R_{\mathrm{init}}(T) \le 2\lambda N^2 K L_0 \le C\lambda N^2 K \log(e\ell). \tag{35}$$

No stability of W_{park} is presumed. The last two terms follow slot by slot from the communication interface. The virtual coupling and nonnegative losses let us extend the retained prefix to T admissible virtual slots on an enlarged probability space; Lemma 30 then bounds $\mathbb{E}R_{\mathrm{active}}$ by $C\Psi(h_\star)\ell + C\mathcal{L}_\star$ without substituting a random stopping time into the theorem.

For the baseline term, the deterministic bit bound gives $C_{\mathrm{post}} \le C\lambda NK\ell^2$. If $\varepsilon \le h_\star/16$, Lemma 30 gives $r_P(W_c) = 0$ on $\mathcal{P}$; without any independence assumption,

$$\mathbb{E}[C_{\mathrm{post}} r_P(W_c)] \le C\lambda N^2 K \ell^2 p_T.$$

If $\varepsilon > h_\star/16$, resolution inversion yields $\ell \le Ch_\star^{-2}\log(eNK/h_\star)$ and hence $\mathbb{E}[C_{\mathrm{post}} r_P(W_c)] \le C\lambda N^2 K F(h_\star)$. Together the cases give this last bound, since $F(h_\star) \ge 1$; critically there is no Tp_T term. Also $\mathbb{E}[\kappa C_{\mathrm{post}}] \le C\kappa\lambda NK\ell$. Adding (35) proves the intermediate form with an extra $C\lambda N^2 K\log(e\ell)$. Since $\Psi(h_\star) \ge NK$, the inequality $A\log(ex) \le Px + CA\log(e + A/P)$, with $A = \lambda N^2 K$, $P = \Psi(h_\star)$, $x = \ell$, absorbs that term into the leading statistical term and $C\lambda N^2 K\log(e\lambda N)$. This proves (27). Dividing by $\log T$ proves the fixed-market asymptotic statement. Each participant's active regret is bounded by Lemma 30 and its control-slot loss by at most $\lambda B(T)$, giving $CNK[B_T/\delta_{\mathrm{safe}}^2 + \lambda\ell]$ for players and arms.

The two cases of Theorem 4 follow by substituting their (λ, κ) and, for graph traffic, multiplying the source-bit count by $4|E(G)|D$. The lower bound of Section 5.3 remains statistical: conditional messages and routing are functions of prior observations and shared protocol randomness, so they add no reward KL divergence. The optimal in-band signaling overhead remains an open communication question.

H Policy Classes and Matched Coefficients

The policy class $\Pi_{\log}$ of Section 5.3 allows coefficients depending on N, K, ν, but requires them to be finite and independent of T. Policies know N, K, T, not the true gaps or tie pattern. By (4), ACTL belongs to this class with a coefficient at most $CNK\delta_{\mathrm{safe}}^{-2}\log(eNK)$ for each participant, where δ_{safe} is the safety scale of (4). Lemma 30 and the physical-time analysis in Appendix G.6 give membership for LC-ACTL on every fixed realization with finite λ, κ. In particular, each fixed-market regret is $o(T^a)$ for every $a > 0$.

Corollary 1 (Certificate-margin minimax coefficient)**.** *For $s \in \{\mathrm{str}, \mathrm{tie}\}$, $N \ge 3$, $K \ge N$, and $0 < h \le 1/128$, the first identity of (5) holds. It also holds when policies are restricted to the cooperative connected-local-network class of Theorem 4, or when $\Pi_{\log}$ is enlarged to all policies with subpolynomial weak-player regret and pointwise sublinear arm regret on every fixed market.*

Proof. For the upper bound use ACTL and (3); for $\Delta \ge h$, the coefficient is at most $C\Psi(h)$. For the lower bound set $d = 2h$ in Theorem 1. The finite family has $\Delta \ge h$ and its rate is a constant multiple of $\Psi(h)$. For any fixed policy, finiteness gives

$$\max_{\nu\in\mathcal{H}^s} \limsup_{T\to\infty} \frac{\mathbb{E}_\nu R_P^\Sigma(T)}{\log T} = \limsup_{T\to\infty} \max_{\nu\in\mathcal{H}^s} \frac{\mathbb{E}_\nu R_P^\Sigma(T)}{\log T}.$$

The lower theorem holds for every policy, so also for either restricted or enlarged class. Theorem 4 gives the local-network upper bound. Finally use $\Psi(h) \le N^2/h^2 + NK/h \le 2\Psi(h)$. □

Corollary 2 (Zero-certificate optimality)**.** *Let $\mathcal{C}_{\mathrm{pres}}(h)$ contain the size-N, K markets with $\Delta = 0$ and $\Delta_{\mathrm{pres}} \ge h$. For $N \ge 5$, $K \ge N$, and $0 < h \le 1/64$,*

$$\inf_{\pi\in\Pi_{\log}} \limsup_{T\to\infty} \frac{\sup_{\nu\in\mathcal{C}_{\mathrm{pres}}(h)} \mathbb{E}_\nu R_P^{\Sigma,\pi}(T)}{\log T} = \Theta\left(\frac{N^2}{h^2} + \frac{N(K-N)}{h}\right).$$

The same order holds with the supremum and the horizon limit exchanged, since the finite-time remainder $\mathcal{L}_\star$ of (3) is uniformly bounded over $\Delta_{\mathrm{pres}} \ge h$ and the lower bound is witnessed by a fixed finite family.

Proof. For the upper bound, every member has $\Delta_{\mathrm{pres}} \ge h$, so (3) gives $\mathbb{E}_\nu R_P^\Sigma(T) \le C\Psi(h)\log(eT) + C\mathcal{L}_\star$ because Ψ and $\mathcal{L}_\star$ are decreasing; the remainder is uniformly bounded over the class, and the supremum can be taken inside the horizon limit. For the lower bound use the enlarged families of Theorem 2 with $d = h$. The same two facts give the exchanged ordering. □

The certificate-margin result fixes the market before taking the limit, and its remainder may grow as $\Delta_{\mathrm{pres}}^{-1}$; the zero-certificate class bounds Δ_{pres} away from zero, so there both orderings of supremum and limit coincide. On each finite lower-bound family this distinction disappears as well: all instance-dependent remainders are bounded by a finite maximum.

I AN UNKNOWN-GAP BARRIER FOR SELECTION-SPECIFIC BENCHMARKS

This appendix explains why the benchmark of Section 3 is natural for a single policy spanning strict and exact-tie markets. Without a known lower bound on the strict comparison gap, logarithmic regret under exact ties is incompatible with simultaneously requiring logarithmic player-optimal player regret and arm-pessimal arm regret on all nearby strict instances.

Let M^{po} and M^{ap} be the player-optimal and arm-pessimal stable matchings of a strict instance, and let $\beta_i^{\mathrm{ind}} := \min_{M \in \mathcal{W}^\star} \mu^P_{i,M(i)}$ collect the coordinate-wise lowest weakly stable player payoffs of an exact-tie instance. The three regret notions aggregate positive-part losses against these benchmarks, over players, arms, and players, respectively:

$$R_P^{\mathrm{po}}(T) := \sum_i \sum_{t=1}^{T} \big(\mu^P_{i,M^{\mathrm{po}}(i)} - \mu^P_{i,M_t(i)}\big)_+ ,$$

$$R_A^{\mathrm{ap}}(T) := \sum_a \sum_{t=1}^{T} \big(\mu^A_{a,M^{\mathrm{ap},-1}(a)} - \mu^A_{a,M_t^{-1}(a)}\big)_+ ,$$

$$R_P^{\mathrm{ind}}(T) := \sum_i \sum_{t=1}^{T} \big(\beta_i^{\mathrm{ind}} - \mu^P_{i,M_t(i)}\big)_+ .$$

The first two are selection-specific strict-market objectives; the third uses the coordinate-wise lowest weakly stable payoff baseline of Section 3, which coincides with player-pessimal rewards in strict markets.

The hard family has $N = K = 3$, with players $1, 2, 3$ and arms A, B, C. The player-side means $\mu^P_{1,\cdot} = (1, 3/4, 1/8)$, $\mu^P_{2,\cdot} = (1, 3/4, 1/2)$, $\mu^P_{3,\cdot} = (3/4, 1, 1/8)$ are fixed across instances, giving $1 : A \succ B \succ C$, $2 : A \succ B \succ C$, $3 : B \succ A \succ C$. Arms A and C are also fixed, with $\mu^A_{A,\cdot} = (3/4, 1/8, 1)$ and $\mu^A_{C,\cdot} = (1, 3/4, 1/8)$, i.e. $A : 3 \succ 1 \succ 2$ and $C : 1 \succ 2 \succ 3$. Arm B always ranks player 1 first, with $\mu^A_{B,1} = 1$ and $\mu^A_{B,3} = 1/2$; only its mean for player 2 varies. With $M^+ := \{(1, A), (2, C), (3, B)\}$ and $M^- := \{(1, B), (2, C), (3, A)\}$, for $\varepsilon \in (0, 1/8]$ the three instances are

Instance	$\mu^A_{B,2}$	B's order	stable outcome(s)
$\mathcal{I}_0$	$1/2$	$1 \succ \{2, 3\}$	$\mathcal{W}^\star = \{M^+, M^-\}$
$\mathcal{I}_+$	$1/2 - \varepsilon$	$1 \succ 3 \succ 2$	M^+, M^- stable; $M^+ = M^{\mathrm{po}} = M^{\mathrm{ap}}$
$\mathcal{I}_-$	$1/2 + \varepsilon$	$1 \succ 2 \succ 3$	M^- uniquely stable; $M^- = M^{\mathrm{po}} = M^{\mathrm{ap}}$

All rewards are Bernoulli with the displayed means. Direct blocking-pair inspection gives the stable sets listed: throughout, the four matchings other than $M^\pm$ are blocked by $(3, A)$ or $(1, B)$, and $(2, B)$ blocks M^+ exactly under $\mathcal{I}_-$.

Proposition 4 (Unknown-gap benchmark barrier)**.** *Fix any learning algorithm and a constant $C_0 > 0$. Suppose that on the exact-tie instance $\mathcal{I}_0$, $\mathbb{E}_{\mathcal{I}_0} R_P^{\mathrm{ind}}(T) \le C_0 \log T$ for all sufficiently large T. Then there is a universal constant $c > 0$ such that, for every sufficiently large T, there are strict perturbations $\mathcal{I}_+, \mathcal{I}_-$ with $\Delta_{\min} = \Theta(1/\sqrt{\log T})$ satisfying*

$$\max\big\{\mathbb{E}_{\mathcal{I}_+} R_P^{\mathrm{po}}(T),\ \mathbb{E}_{\mathcal{I}_-} R_A^{\mathrm{ap}}(T)\big\} \ge cT.$$

Consequently, logarithmic indifferent regret on exact-tie instances is incompatible with simultaneous $O(\log T/\Delta^2_{\min})$ guarantees for player-optimal player regret and arm-pessimal arm regret uniformly over all strict instances without a known gap lower bound.

Proof. Hidden edge. The table gives $\mathcal{W}^\star(\mathcal{I}_0) = \{M^+, M^-\}$, hence $\beta_1^{\mathrm{ind}} = \min\{1, 3/4\} = 3/4$, $\beta_2^{\mathrm{ind}} = 1/2$, and $\beta_3^{\mathrm{ind}} = \min\{1, 3/4\} = 3/4$. If player 2 receives arm B, then arms A and C are split between players 1 and 3, and whichever of the two receives C loses $3/4 - 1/8 = 5/8$. Thus every feasible matching containing edge $(2, B)$ incurs at least $5/8$ aggregate indifferent regret under $\mathcal{I}_0$, so with $N_{2B}(T)$ the number of plays of $(2, B)$ up to time T we have $\mathbb{E}_{\mathcal{I}_0} N_{2B}(T) \le \frac{8}{5} \mathbb{E}_{\mathcal{I}_0} R_P^{\mathrm{ind}}(T) \le C_1 \log T$, where $C_1 := \frac{8}{5} C_0$.

Indistinguishability. Let $\mathbb{P}_0, \mathbb{P}_+, \mathbb{P}_-$ be the laws of the interaction history through T under $\mathcal{I}_0, \mathcal{I}_+, \mathcal{I}_-$, and write TV for total variation distance. These laws differ only in the arm-side reward distribution

of edge $(2, B)$: $\mathrm{Bernoulli}(1/2)$ under $\mathcal{I}_0$ versus $\mathrm{Bernoulli}(1/2 \mp \varepsilon)$ under $\mathcal{I}_\pm$. For $\varepsilon \le 1/8$ the Bernoulli divergence satisfies $\mathrm{kl}(\mathrm{Bernoulli}(1/2) \,\|\, \mathrm{Bernoulli}(1/2 - \varepsilon)) \le 8\varepsilon^2$, so the chain rule for KL divergence under adaptive sampling gives

$$\max\{\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_+),\ \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_-)\} \le 8\varepsilon^2\, \mathbb{E}_{\mathcal{I}_0} N_{2B}(T) \le 8C_1\varepsilon^2 \log T\,.$$

Choosing $\varepsilon = \varepsilon_T := (16\sqrt{C_1 \log T})^{-1}$ makes both divergences at most $1/32$ and, by Pinsker's inequality, $\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_\pm) \le 1/8$; the triangle inequality then gives $\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) \le 1/4$.

Strict-side separation. Let $N_+(T) := \sum_{t=1}^{T} \mathbf{1}\{M_t = M^+\}$. Under $\mathcal{I}_+$, every matching other than M^+ loses the players at least $1/2$ in total relative to the player-optimal benchmark $\mu^P_{1,A} = 1$, $\mu^P_{2,C} = 1/2$, $\mu^P_{3,B} = 1$, so $R^{\mathrm{po}}_P(T) \ge \frac{1}{2}(T - N_+(T))$. Under $\mathcal{I}_-$, the arm-pessimal benchmark is M^-, and playing M^+ loses arms A and B at least $(1 - 3/4) + (1 - 1/2) = 3/4$ in total, so $R^{\mathrm{ap}}_A(T) \ge \frac{3}{4}N_+(T)$. For every t, $\mathbb{P}_+(M_t \ne M^+) + \mathbb{P}_-(M_t = M^+) \ge 1 - \mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) \ge 3/4$, hence $\mathbb{E}_{\mathcal{I}_+} R^{\mathrm{po}}_P(T) + \mathbb{E}_{\mathcal{I}_-} R^{\mathrm{ap}}_A(T) \ge \frac{1}{2} \cdot \frac{3}{4}\, T = \frac{3}{8}\, T$, and at least one of the two terms is at least $\frac{3}{16}T$: the proposition holds with $c = 3/16$.

Finally, $\Delta_{\min} = \varepsilon_T = \Theta(1/\sqrt{\log T})$, so $O(\log T/\Delta^2_{\min}) = O((\log T)^2) \ll T$ for sufficiently large T: the linear lower bound above rules out any simultaneous strict-side guarantee of this order for both player-optimal player regret and arm-pessimal arm regret. □

The same impossibility applies to uniform individual guarantees, since summing such bounds over the finitely many participants would imply the aggregate guarantees ruled out above. Together with the structural nonexistence of a player-optimal weakly stable matching under ties, this barrier motivates the selection-rule-free coordinate-wise stable benchmark of Section 3 for a single policy spanning strict and exact-tie markets.

J Additional Literature Review on Matching Markets with Ties

The classical *Stable Marriage* (SM) problem, introduced by Gale and Shapley, studies one-to-one matchings in a bipartite market where each agent provides a strict ranking over all partners; the deferred acceptance (DA) algorithm always produces a stable matching and underlies a large body of work in matching theory and market design Gale & Shapley (1962); Roth & Sotomayor (1992). Two practically motivated extensions relax these assumptions: (i) *incomplete lists* (unacceptable partners), and (ii) *ties* (indifferences). When ties are allowed, several notions of stability coexist, including *weak*, *strong*, and *super* stability, depending on whether blocking pairs require strict improvement on both sides or allow weak improvement on one/both sides (Irving, 1994; Manlove, 1999). In this paper we focus on weak stability, as it is the most common relaxation in applications with indifferences and is preserved by arbitrary tie-breaking (breaking ties and running DA returns a matching that is weakly stable with respect to the original indifferences).

Existence and the role of optimization under ties. While weakly stable matchings are guaranteed to exist and can be found efficiently via tie-breaking and DA, optimizing over weakly stable matchings can be difficult. In the presence of ties and/or incompleteness, different stable matchings may have different cardinalities and welfare properties (e.g., egalitarian cost, regret, sex-equality), and the problem often becomes computationally hard once one seeks *optimal* stable solutions rather than *some* stable solution (Iwama et al., 1999; Manlove et al., 2002). In particular, allowing both ties and incomplete lists yields NP-hardness phenomena and strong inapproximability for certain optimality objectives, even though each relaxation alone remains tractable (Halldórsson et al., 2003; 2007). Further work gives approximation algorithms and hardness results for maximum-cardinality and welfare-optimal variants of SMTI (Irving & Manlove, 2007; Király, 2011), as well as parameterized and restricted-case analyses (e.g., bounded list lengths, restricted tie structures, or additional constraints on edges) (Marx & Schlotter, 2010).

Pareto-optimality among stable matchings and improvement cycles. In school choice and matching with weak priorities, another question concerns the relation between *stability* and *Pareto optimality* for one side of the market. In school choice, it is common for schools to have weak priority orders (ties), and DA with an arbitrary tie-breaking rule produces a stable allocation but may introduce

"artificial" constraints that make the outcome Pareto dominated for students (Abdulkadiroğlu & Sönmez, 2003). Erdil & Ergin (2008) show that if a stable matching is Pareto dominated by another stable matching, then it admits a *stable improvement cycle*. In their school-choice setting, repeatedly implementing such cycles yields a student-optimal stable matching that Pareto improves upon DA outcomes under tie-breaking. This *stable-improvement-cycles* viewpoint provides both a structural characterization and a polynomial-time method to reach Pareto-undominated stable outcomes under indifferences. A parallel economics literature studies the incentive consequences: while Pareto improvements over DA are possible under weak priorities, such improvements typically come at the cost of strategy-proofness, motivating an active line of work on trade-offs between stability, efficiency, and incentives in markets with ties (Abdulkadiroğlu et al., 2009; Erdil & Ergin, 2017).

Computational approaches and extensions. On the algorithmic side, SMTI has also been studied via complete-search and constraint-based methods (constraint programming, SAT/ASP encodings), local search heuristics, and hybrid approaches, motivated by the empirical difficulty of large SMTI instances and the need to optimize secondary criteria (Gent et al., 2001; Gelain et al., 2010; Gent et al., 2002). Moreover, many-to-one generalizations such as Hospitals/Residents with Ties (HRT) inherit similar phenomena: weak stability is easy to obtain, but optimization (maximum size, egalitarian, minimum regret, etc.) can be computationally challenging (Manlove et al., 2002; Irving et al., 2003). Monographs and surveys collect these results and their algorithmic methods on matching under preferences (Manlove, 2013; Iwama & Miyazaki, 2008).

Learning-theoretic variants. Within the learning literature, further lines study decentralized matching bandits and conflict resolution (Sankararaman et al., 2021), protocols that avoid communication entirely (Maheshwari et al., 2022), many-to-one markets with quotas or incentive constraints (Wang et al., 2022; Kong & Li, 2024; Zhang & Fang, 2024a), Thompson sampling (Kong et al., 2022), collision and ranking design (Zhang et al., 2022; Su et al., 2022), contextual or non-stationary preferences (Li et al., 2022; Muthirayan et al., 2022; Ghosh et al., 2022), and multi-stage or uncertain-preference markets (Dai & Jordan, 2021b;a; Li et al., 2024). Adjacent objectives include fairness and stability guarantees (Cen & Shah, 2022), equilibrium learning (Jagadeesan et al., 2021), no-regret reinforcement learning (Min et al., 2022), and stability through sampling (Hosseini et al., 2024).

Decentralized information and communication models. Prior decentralized matching-bandit work uses different ways to coordinate local learners: observed assignments or outcomes (Liu et al., 2021), shared boards or public flags (Pagare & Ghosh, 2024; Basu, 2025), and signaling through market interactions (Sankararaman et al., 2021; Basu et al., 2021). Communication- and coordination-free learning is another model (Maheshwari et al., 2022). Two-sided decentralized learning has also been studied directly (Zhang & Fang, 2024b; Pagare & Ghosh, 2024); it should not be conflated with one-sided learning where one side's preferences are known. These works differ in feedback, objectives, preference assumptions, and whether signaling consumes reward time, so their guarantees are not interchangeable with ours. Under two-sided unknown rewards and exact ties, LC-ACTL preserves the leading certificate-based core–tail rate for the coordinate-wise lowest weakly stable payoff baseline over connected local networks while communication consumes physical time. In-band signaling also achieves logarithmic regret without an auxiliary channel, with an explicit signaling term.

K Broader Impact

Exploration can benefit one participant while harming another. Our results quantify these losses but do not model strategic reporting or fairness constraints. Applications would need to assess those objectives separately. No personal data or human-subject experiments are used.