EX-ANTE ADMISSION WITH STOCHASTIC PRIORITY COMPETITION: LEARNING WHAT MATTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study an online ex-ante admission problem in which multiple devices compete for a unit-capacity resource through an unknown priority order and stochastic gating mechanism. At each round, the controller observes demand probabilities and selects an admission set before the actual demands are realized. The active admitted devices then compete according to the latent priority order, while the served device generates a stochastic reward with an unknown mean. This setting creates a coupled learning problem: the controller must simultaneously learn reward parameters and acquire priority information from noisy and randomly occurring competition events. We first characterize the full-information admission oracle through a dynamic program and develop an Exact-Priority Two-Phase algorithm based on a sequential noisy priority comparator. We then show that exact recovery of the entire priority order can lead to substantial over-exploration, since many pairwise ordering errors have only limited influence on the admission value. Motivated by a sensitivity analysis of priority misordering, we propose Relevance-Aware Priority-Upper Confidence Bound (UCB), which probes unresolved priority relations only when their potential decision impact remains significant. We establish regret guarantees that capture both statistical informativeness and decision relevance, and derive a complementary decision-theoretic lower bound showing that both decision sensitivity and statistical distinguishability are intrinsic. Experiments confirm that the proposed strategy substantially reduces unnecessary priority exploration, particularly in regimes with nearly homogeneous rewards, low effective service probabilities, or rare contextual overlap.

1 INTRODUCTION

Multi-armed bandits (MABs) provide a canonical framework for sequential decision making under uncertainty, capturing the fundamental tradeoff between exploration and exploitation (Auer et al., 2002; Bubeck & Cesa-Bianchi, 2012; Chen et al., 2013; Anantharam et al., 2003). Many modern computing and communication systems involve multiple units simultaneously requesting service from a common pool of limited resources. Such settings have motivated a broad literature on multi-player bandits, collision-based learning, and bandits with shareable or capacity-constrained resources (Lagrée et al., 2016; Liu & Zhao, 2010; Wang et al., 2022b;a).

Resource competition, however, cannot always be eliminated by coordinating the participating units. A number of operational systems implement an explicit *priority-based contention mechanism*. When several eligible units request the same resource, a higher-priority request is considered first, while lower-priority requests receive service only when earlier contenders do not consume the resource. Moreover, access itself can be stochastic—for instance because of channel availability, eligibility checks, transient failures, or randomized gating—so that a high-priority unit may fail to acquire the resource and the opportunity is passed to the next contender. Competition is therefore an intrinsic part of the service mechanism rather than an event that the learning algorithm can simply eliminate.

A second complication is that admission decisions often have to be made *ex-ante*. A controller may have access to historical statistics, contextual information, or predictive models that estimate the probability that each device will become active, but it does not know the exact demand realization when deciding which devices should be admitted. This type of ex-ante control is not merely a modeling convenience. It is embedded in operational standards and systems. For example, in cellular networks, Access Class Barring (ACB) allows the network to broadcast barring parameters (e.g.,

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.

barring rates and durations) so that devices probabilistically defer access attempts, protecting the access channel under overload (Myung, 2008). Consequently, even after the controller selects an admission set, it does not know which admitted devices will actually contend for the resource.

In this paper, we study an online resource-allocation problem that combines these two mechanisms: *ex-ante admission and stochastic priority-based competition*. In each round, the controller observes demand probabilities for a collection of devices and chooses an admission set before their demands are realized. The admitted devices that become active subsequently compete for a unit-capacity resource according to an unknown but fixed priority order. When an active device is reached in this order, it successfully passes a stochastic gate with a device-dependent probability; otherwise the resource opportunity continues to lower-priority contenders. A successfully served device generates a stochastic reward with an unknown mean. The controller therefore has to learn both reward parameters and the latent priority structure while repeatedly choosing admission sets.

This combination creates a learning problem whose information structure differs from several closely related bandit and assortment formulations. The first difficulty is *latent competition*. Admitting two devices does not imply that they will actually compete in that round. Thus, priority learning must be performed through noisy and randomly occurring comparison opportunities. The second difficulty is that information acquisition is *endogenous to the admission policy*. The admission set simultaneously determines the reward collected in the current round and which priority relations may become observable. This creates a coupling between structural learning and reward maximization that is absent from settings in which the contestants in a comparison are directly observable or freely selected. Most importantly, there is a fundamental *mismatch between exact priority identification and optimal decision making*. Recovering the complete priority permutation may require resolving many statistically difficult pairwise relations. Yet not every such relation has the same consequence for the admission value. An algorithm that insists on learning all of these relations exactly can therefore spend substantial exploration effort on priority distinctions that are statistically nontrivial but decision-theoretically negligible. This raises the central question of this work: *How much of the latent priority order must be learned in order to make near-optimal ex-ante admission decisions?*

At a one-round level, our service mechanism is equivalent to a random consideration set (RCS) choice rule (Manzini & Mariotti, 2014; Gallego & Li, 2024): admitted devices form a random effective consideration set, from which the highest-priority device is served. This connection also relates our problem to online RCS assortment optimization (Lu, 2023), preselection bandits (Bengs & Hüllermeier, 2020), and cascading bandits (Kveton et al., 2015). Our learning problem, however, differs in the interaction of the unknown quantities and the decision objective: the effective participation probabilities are observed and time varying, the priority order is latent, and the service rewards are stochastic and separately unknown. Consequently, the value of learning a priority relation varies across contexts and with reward uncertainty. Our central objective is therefore not full recovery of the latent order, but determining *which parts of that order are worth learning for near-optimal admission.* This leads to a decision-sensitive exploration problem that is absent when the goal is parameter identification, preference learning, or identifying the most attractive items. We provide detailed comparison with existing works in Appendix A.

To address the challenges, we characterize the admission problem when the priority order and reward parameters are known. The resulting oracle admits a dynamic-programming representation. This structure provides a natural benchmark for learning and also reveals how local changes in the priority order propagate to the final admission value. Building on this oracle, we develop Exact-Priority Two-Phase, a conceptually clean baseline that first learns the complete priority order and then performs reward learning under the estimated order. We then move beyond exact identification by studying the *decision sensitivity of priority misordering*. Starting from a local adjacent inversion, we characterize how an incorrect pairwise order changes the service value through several interpretable factors. These sensitivity bounds allow us to quantify the aggregate decision impact of the currently unresolved portion of the priority order. This analysis leads to our main algorithm, Relevance-Aware Priority-UCB. Rather than asking whether every pair has been statistically resolved, the algorithm asks whether the remaining priority uncertainty can still change the admission value by a meaningful amount. It maintains a certificate for the possible loss induced by unresolved order relations and continues priority probing only when this loss exceeds a shrinking tolerance. Once the residual order uncertainty becomes decision-benign, the algorithm can exploit without re-

covering the complete permutation. In this sense, the learning objective changes from *identifying the latent priority structure* to *acquiring only the priority information needed for near-optimal control*.

Our main contributions are summarized as follows. *1)* We formulate the online ex-ante admission problem under a stochastic priority mechanism, and characterize the oracle required for subsequent learning analysis. *2)* We analyze the sensitivity of the oracle value to local priority inversions and use this structure to design Relevance-Aware Priority-UCB, which shifts priority learning from full structural identification to decision-directed information acquisition. *3)* We establish regret bounds for the proposed algorithms. We also complement the upper bounds with a decision-theoretic lower bound and identify favorable regimes, in which relevance-aware learning can substantially reduce unnecessary exploration. Simulations validate these mechanisms and our theoretical analysis.

2 System Model and Problem Formulation

We consider a system consisting of M devices, indexed by $[M] = \{1, \dots, M\}$, that share a single-unit service resource over a horizon of T time slots. In each slot, the resource can serve at most one device. At the beginning of slot t, the controller observes a vector of demand probabilities $p_t = (p_{1,t}, \dots, p_{M,t}) \in [0,1]^M$, and device i will have a conditionally independent random demand $D_{i,t} \sim \text{Bernoulli}(p_{i,t})$, where $D_{i,t} = 1$ indicates that device i has an active request in slot t. After observing p_t, but before observing $D_t = (D_{1,t}, \dots, D_{M,t})$, the controller needs to select an admission set $A_t \subseteq [M]$. Only devices in A_t are allowed to compete for the resource during slot t.

The devices are associated with an unknown but fixed priority order. We represent this order by a permutation $\pi = (\pi(1), \dots, \pi(M))$, where $\pi(1)$ is the highest-priority device and $\pi(M)$ is the lowest-priority device. Equivalently, let $r_\pi(i) \in [M]$ denote the rank of device i, with a smaller value corresponding to a higher priority. We write $i \succ j$ when $r_\pi(i) < r_\pi(j)$. The devices compete for the resource under a stochastic priority-gated mechanism. Each device i is associated with a gate success probability $\gamma_i \in (0,1)$, which is assumed to be known to the controller. In slot t, device i generates a gate outcome $G_{i,t} \sim \text{Bernoulli}(\gamma_i)$, which is not revealed to the controller. Then the effective activation indicator is $Z_{i,t} = D_{i,t}G_{i,t}$. Among all admitted devices satisfying $Z_{i,t} = 1$, the resource is allocated to the one with the highest priority. We use I_t to denote this device. If every admitted device either has no demand or fails its gate, no device is served, and we write $I_t = 0$.

When device i is served in slot t, the controller observes a random reward $X_{i,t} = \theta_i + \eta_{i,t}$, where $\theta_i \in [0,1]$ is an unknown mean service reward and $\eta_{i,t}$ is conditionally σ-sub-Gaussian. At the end of slot t, the controller observes the demand realizations of admitted devices, the served device identity I_t and the realized reward $X_{I_t,t}$. The individual gate outcomes $G_{i,t}$ are not directly observed. In particular, when two admitted devices both have demand and the lower-priority device is eventually served, the controller cannot determine from a single observation whether that device actually has higher priority or whether the higher-priority device merely failed its gate.

Regret Benchmark. For a given admission set A, priority order π, and reward vector θ, define the conditional expected reward in slot t as $F_t(A; \pi, \theta) = \mathbb{E}[X_{I_t,t} \mid \mathcal{F}_{t-1}, p_t, A_t = A] = \sum_{i=1}^{M} \theta_i \Pr(I_t = i \mid \mathcal{F}_{t-1}, p_t, A_t = A)$. The service probabilities depend jointly on the admission set, demand probabilities, gate probabilities, and latent priority order. In particular, admitting an additional device may increase the probability that some request is successfully served, but it may also allow that device to preempt lower-priority devices with larger rewards. Consequently, the expected reward is generally not monotone in the admission set. We benchmark the learning algorithm against a dynamic oracle that knows the true priority order π and the reward vector θ. The oracle observes the probability vector p_t but does not observe the actual demand or gate realizations. Its optimal action is $A_t^\star \in \arg\max_{A \subseteq [M]} F_t(A; \pi, \theta)$. Thus, the benchmark is an ex-ante oracle rather than a clairvoyant scheduler. For a sequence of admission decisions $A_1, \dots, A_T$, we define the cumulative pseudo-regret as $R_T = \sum_{t=1}^{T} \left[F_t(A_t^\star; \pi, \theta) - F_t(A_t; \pi, \theta)\right]$. When the algorithm is randomized, R_T is itself random through the selected admission sets and observed history, and we evaluate either its expectation $\mathbb{E}[R_T]$ or a high-probability upper bound.

3 Oracle Policy under Known Priority and Reward Parameters

Fix a slot t and device i, the effective activation indicator $Z_{i,t} = D_{i,t}G_{i,t} \sim \text{Bernoulli}(a_{i,t})$, where $a_{i,t} = p_{i,t}\gamma_i$. We refer to $a_{i,t}$ as the effective activation probability. For notational convenience, we relabel the devices according to decreasing priority, so that $1 \succ 2 \succ \cdots \succ M$. For any admitted

subset $A \subseteq [M]$, the expected reward in round t becomes

$$F_t(A; \pi, \theta) = \sum_{k \in A} \theta_k a_{k,t} \prod_{j \in A: j<k} (1 - a_{j,t}). \tag{1}$$

This expression suggests a recursive structure: whether it's beneficial to include a device depends on its own reward contribution and on the continuation value obtainable from lower-priority devices when the current device happens to be inactive. It can be solved using dynamic programming. To characterize the oracle, define $V_{k,t}$ as the maximum expected reward obtainable in round t when only devices in $\{k, k+1, \cdots, M\}$ are available for admission. The terminal value is $V_{M+1,t} = 0$. Now consider device k. If device k is not selected, then the best achievable expected reward is simply $H_{k,t} := V_{k+1,t}$. If device k is selected, then two cases arise. With probability $a_{k,t}$, device k is immediately served, yielding expected reward θ_k; with probability $1 - a_{k,t}$, device k is inactive, and the system falls through to the lower-priority devices, which contribute the continuation value $V_{k+1,t}$. Hence the value of selecting device k is $G_{k,t} := a_{k,t}\theta_k + (1 - a_{k,t})V_{k+1,t}$. This gives the Bellman recursion $V_{k,t} = \max\{V_{k+1,t}, a_{k,t}\theta_k + (1 - a_{k,t})V_{k+1,t}\}$. Equivalently, $V_{k,t} = V_{k+1,t} + a_{k,t}\big(\theta_k - V_{k+1,t}\big)_+$, where $(x)_+ = \max\{x, 0\}$. The oracle value in round t is then $V_{1,t}$. And the corresponding optimal admission set is $A_t^* = \{k \in [M] : G_{k,t} \geq H_{k,t}\}$.

For later analysis, it is convenient to express the dynamic program through device-specific Bellman operators. For device i in slot t, define

$$\mathcal{T}_{i,t}(v) = \max\{v, a_{i,t}\theta_i + (1 - a_{i,t})v\} = v + a_{i,t}(\theta_i - v)_+. \tag{2}$$

The quantity $\mathcal{T}_{i,t}(v)$ is the optimal expected value obtained when device i is considered immediately before a lower-priority subsystem whose continuation value is v. The oracle value can then be written as the composition $V_{1,t} = \mathcal{T}_{\pi(1),t} \circ \mathcal{T}_{\pi(2),t} \circ \cdots \circ \mathcal{T}_{\pi(M),t}(0)$. This representation makes explicit that priority determines the order in which the Bellman operators are composed. In general, $\mathcal{T}_{i,t} \circ \mathcal{T}_{j,t} \neq \mathcal{T}_{j,t} \circ \mathcal{T}_{i,t}$, so an incorrect priority order may change both the computed value and the resulting admission set. The Bellman operators satisfy several basic properties, which will be used in the sensitivity analysis and regret proof. We provide detailed statement and proof for these properties in Lemma 1, which is postponed to Appendix C.

4 Exact-Priority Learning: A Two-Phase Baseline

We introduce a natural two-phase learning strategy, which first identifies the priority order using sequential noisy comparators, and then applies an optimistic version of the oracle dynamic program to learn the service rewards. Detailed description and proofs are provided in Appendix B.

Log-Likelihood Ratio Based Sequential Noisy Comparator. Consider two devices i and j whose relative priority is unknown. By admitting them in a probing round, the controller may obtain information about whether $i \succ j$ or $j \succ i$. Such information is intrinsically noisy for two reasons. First, the pair becomes informative only when the relevant joint-demand event occurs. Second, even conditional on both devices being active, stochastic gating prevents a single service outcome from deterministically revealing their order.

Let q_{ij} denote a lower bound on the probability that a probing slot produces an informative joint-activity event, and let I_{ij} denote the statistical information separating the service-outcome distributions under the two hypotheses $H_{ij} : i \succ j$, and $H_{ji} : j \succ i$. We use a sequential likelihood-ratio comparator that accumulates evidence until one of the two orders can be certified with the prescribed confidence level. Theorem 9 in Appendix B.1 shows that for any pair (i, j), the comparator identifies its true relative priority with probability at least $1 - \delta_{ij}$, using $\widetilde{O}(1/q_{ij}I_{ij})$ expected probing slots, where the logarithmic factors depend on the desired confidence level. Thus, priority relations that are rarely jointly active or statistically difficult to distinguish require substantially more effort.

A Two-Phase Baseline. To recover the complete priority order, the comparator can be embedded into any comparison-based sorting procedure. We use a merge-sort implementation, which requires $O(M \log M)$ pairwise comparisons. The precise comparison schedule and its interaction with the sequential comparator are given in Appendix B.2. During Phase I, the controller performs priority probes according to the sorting procedure. Each required pairwise relation is repeatedly probed until the sequential comparator resolves it with sufficiently high confidence. Once the sorting procedure terminates, the resulting permutation $\widehat{\pi}$ is taken as the estimated global priority order. During Phase II, priority exploration is stopped. For each device i, the controller maintains an empirical reward

estimate $\widehat{\theta}_i(t)$ and a corresponding upper confidence bound $U_i(t)$. At every round, it substitutes the optimistic reward vector $U(t)$ and the learned order $\widehat{\pi}$ into the oracle dynamic program developed in Section 3, and admits the resulting set.

Let K_M be a deterministic upper bound on the number of pairwise comparator calls made by the sorting procedure. Assume that every pair (i, j) that may be probed satisfies $p_{i,t}p_{j,t} \geq \underline{q} > 0$, and define $I_{\min} = \min_{i \neq j} I_{ij} > 0$. Then we have the following result.

Theorem 1 (Regret of exact-priority two-phase learning)**.** *Let $\delta_\pi, \delta_\theta \in (0, 1)$. Under the stochastic priority-gating model, the exact-priority two-phase algorithm satisfies*

$$\mathbb{E}[R_T] \leq C_1 \min\{T, K_M \log(2K_M/\delta_\pi)/\underline{q}I_{\min}\} + C_2[M + \sigma\sqrt{MT \log(4MT/\delta_\theta)}] + T(\delta_\pi + \delta_\theta), \quad (3)$$

where $C_1, C_2 > 0$ are universal constants. In particular, since $K_M = O(M \log M)$, choosing $\delta_\pi = \delta_\theta = 1/T$ gives $\mathbb{E}[R_T] = \widetilde{O}(\min\{T, M \log M/\underline{q}I_{\min}\} + \sqrt{MT})$.

5 Decision Sensitivity to Priority Misordering

Exact priority recovery is sufficient for optimal admission, but it may be stronger than necessary. Two devices may be difficult to compare statistically while their relative order has little effect on the achievable reward. In this section, we study how priority misordering propagates through the Bellman recursion. Throughout this section, we fix a slot t and write $a_i = a_{i,t} = p_{i,t}\gamma_i$ when the slot index is clear. Detailed proofs for this section are provided in Appendix C.

Recall the device-specific Bellman operator $\mathcal{T}_i(v) = v + a_i(\theta_i - v)_+$. For a priority order $\pi(1) \succ \cdots \succ \pi(M)$, the oracle value is $V_t^\pi = \mathcal{T}_{\pi(1)} \circ \cdots \circ \mathcal{T}_{\pi(M)}(0)$. Priority matters because the Bellman operators do not generally commute. For two devices i and j, define their local adjacent-inversion sensitivity at continuation value v as

$$\kappa_{ij,t}(v) = |\mathcal{T}_i(\mathcal{T}_j(v)) - \mathcal{T}_j(\mathcal{T}_i(v))|. \quad (4)$$

Hence, $\kappa_{ij,t}(v)$ measures the exact value distortion caused by reversing the priorities of two adjacent devices while leaving the lower-priority suffix unchanged. Suppose that i and j are both admitted and adjacent in the priority order. Let h and ℓ denote the higher- and lower-reward devices in the pair, respectively, i.e. $\theta_h \geq \theta_\ell$. The following closed-form expression holds.

Theorem 2. *For any continuation value $v \in [0, 1]$, $\mathcal{T}_h\big(\mathcal{T}_\ell(v)\big) - \mathcal{T}_\ell\big(\mathcal{T}_h(v)\big) = a_\ell \cdot \mathbf{1}\{v < \theta_\ell\} \cdot \min\{a_h(\theta_h - \theta_\ell), (1 - a_h)(\theta_\ell - v)\} \geq 0$. Taking absolute value gives $\kappa_{ij,t}(v) \leq a_{i,t}a_{j,t}|\theta_i - \theta_j|$.*

An adjacent inversion may occur below a prefix of higher-priority devices. Such a prefix cannot amplify the local distortion. To be specific, let $\mathcal{P} = \mathcal{T}_{k_1} \circ \cdots \circ \mathcal{T}_{k_r}$ be the composition of the Bellman operators corresponding to devices ranked above the swapped pair. Since every Bellman operator is nonexpansive, $|\mathcal{P}(x) - \mathcal{P}(y)| \leq |x - y|$. Consequently, $|\mathcal{P} \circ \mathcal{T}_i \circ \mathcal{T}_j(v) - \mathcal{P} \circ \mathcal{T}_j \circ \mathcal{T}_i(v)| \leq \kappa_{ij,t}(v) \leq a_{i,t}a_{j,t}|\theta_i - \theta_j|$. The actual effect of a priority pair inversion can only be smaller because higher-priority devices can screen the inverted pair from being reached. The bound in Theorem 2 ignores this additional attenuation and provides a simple certificate that does not require knowing the exact continuation and reach probabilities.

We now extend the local sensitivity result to two arbitrary priority permutations. For permutations π and $\widehat{\pi}$, define their inversion set as $\mathrm{Inv}(\pi, \widehat{\pi}) = \{\{i, j\} : i \succ_\pi j \text{ and } j \succ_{\widehat{\pi}} i\}$. Define the slot-dependent weighted inversion distance as

$$d_t(\pi, \widehat{\pi}; \theta) = \textstyle\sum_{i,j \in \mathrm{Inv}(\pi,\widehat{\pi})} a_{i,t}a_{j,t}|\theta_i - \theta_j|. \quad (5)$$

Using this inversion distance, we can bound the value distortion under π and $\widehat{\pi}$ as follows.

Theorem 3. *For any two priority permutations π and $\widehat{\pi}$, we have $|V_t^\pi - V_t^{\widehat{\pi}}| \leq d_t(\pi, \widehat{\pi}; \theta)$.*

Let $A_t^\star(\pi) \in \arg\max_{A \subseteq [M]} F_t(A; \pi, \theta)$ denote the true oracle action, and let $A_t^\star(\widehat{\pi}) \in \arg\max_{A \subseteq [M]} F_t(A; \hat{\pi}, \theta)$ be the admission set obtained by solving the dynamic program under $\widehat{\pi}$. For any fixed admission set A, the expected-reward distortion caused by replacing π with $\widehat{\pi}$ satisfies $|F_t(A; \pi, \theta) - F_t(A; \widehat{\pi}, \theta)| \leq d_t(\pi, \widehat{\pi}; \theta)$. This gives the following decision guarantee.

Theorem 4 (Policy loss under priority misordering)**.** *The admission set computed under $\widehat{\pi}$ satisfies $F_t(A_t^\star(\pi); \pi, \theta) - F_t(A_t^\star(\widehat{\pi}); \pi, \theta) \leq 2d_t(\pi, \widehat{\pi}; \theta)$.*

Theorem 4 provides the key connection between priority estimation and online decision performance. Exact permutation recovery requires $\widehat{\pi} = \pi$, whereas near-optimal admission only requires the weighted inversion error to be small. In particular, if $d_t(\pi, \widehat{\pi}; \theta) \leq \varepsilon_t$, then the per-slot regret attributable to priority misordering is at most $2\varepsilon_t$. Therefore, pairwise priority relations need not be learned with uniform accuracy. Theorem 4 suggests a different learning objective: rather than identifying every pairwise relation, the controller may maintain a partial order and refine only those unresolved pairs whose potential contribution to $d_t(\pi, \widehat{\pi}; \theta)$ is significant.

6 Relevance-Aware Priority and Reward Learning

This section develops a relevance-aware algorithm that learns only the priority information required for near-optimal admission. The algorithm maintains a partial order, estimates the rewards concurrently, and probes an unresolved pair only when its possible contribution to the current decision loss is sufficiently large. We provide details of the relevance-aware algorithm in Appendix D.

The algorithm maintains three coupled sources of information: *1)* a directed acyclic graph representing high-confidence priority relations; *2)* a sequential likelihood statistic for each unresolved pair; and *3)* a confidence interval for the mean reward of each device. These components interact in two directions. First, priority information reduces the set of permutations that remain statistically plausible. Second, reward information determines whether the remaining priority ambiguity matters for admission. A pair can therefore become harmless either because its order has been identified or because the reward estimates show that reversing the pair has little effect. In each slot, the algorithm computes an upper bound on the value distortion caused by all unresolved priority relations. If this bound is below a prescribed tolerance, the algorithm exploits by running an optimistic dynamic program under a priority order consistent with the current partial order. Otherwise, it probes a decision-relevant unresolved pair. Priority refinement and reward learning are interleaved, and the algorithm may return to priority exploration if a previously negligible relation becomes important under a later demand context.

Partial-Order Representation. Let $\mathcal{P}_t = ([M], E_t)$ denote the directed graph of priority relations established before slot t. An edge $(i, j) \in E_t$ means that the algorithm has concluded, with high confidence, that $i \succ j$. The graph is maintained transitively: whenever the current relations imply $i \succ j$, the pair is treated as resolved even if it has not been directly probed. Let $\mathcal{L}(\mathcal{P}_t)$ denote the set of linear extensions of $\mathcal{P}_t$, namely, all complete permutations consistent with the established relations. On the event that all accepted pairwise comparisons are correct, the true priority order satisfies $\pi \in \mathcal{L}(\mathcal{P}_t)$ for every slot t. Define the set of unresolved pairs as $\mathcal{U}_t$. For each unordered pair $\{i, j\}$, the algorithm maintains a log-likelihood ratio $L_{ij}(t)$, defined with the orientation H_{ij}^{+} : $i \succ j, H_{ij}^{-} : j \succ i$. The statistic is updated only in slots in which the pair is deliberately isolated for comparison and both devices realize demand. When $L_{ij}(t)$ crosses a prescribed positive or negative boundary, the corresponding directed edge is added to $\mathcal{P}_t$.

Reward Confidence Intervals. Let $N_i(t-1)$ be the number of observed rewards from device i before slot t, and let $\hat{\theta}_i(t)$ be the empirical mean of these observations. For confidence radius $\beta_i(t)$, the clipped interval is $\mathcal{C}_i(t) = [L_i(t), U_i(t)]$, where $L_i(t) = \max\{0, \hat{\theta}_i(t) - \beta_i(t)\}$ and $U_i(t) = \min\{1, \hat{\theta}_i(t) + \beta_i(t)\}$. For a pair $\{i, j\}$, define the largest reward difference as $G_{ij}(t) = \max\{U_i(t) - L_j(t), U_j(t) - L_i(t)\}$. On the event that all reward confidence intervals are valid, $|\theta_i - \theta_j| \leq G_{ij}(t)$. Thus, even if the relation between i and j remains unknown, its potential value impact decreases as the reward confidence intervals become more informative.

An Order-Uncertainty Certificate. For the current demand context, the pairwise relevance upper bound is $C_{ij,t} = a_{i,t} a_{j,t} G_{ij}(t)$. We aggregate the contributions of all unresolved pairs as

$$B_t^{\mathrm{ord}} = \sum_{\{i,j\} \in \mathcal{U}_t} C_{ij,t} = \sum_{\{i,j\} \in \mathcal{U}_t} a_{i,t} a_{j,t} G_{ij}(t). \tag{6}$$

This certificate is deliberately conservative. Its advantage is that it is directly computable from the current partial order, demand probabilities, and reward confidence intervals. Section 7 shows that using an arbitrary priority order consistent with $\mathcal{P}_t$ instead of the true order can induce at most $2B_t^{\mathrm{ord}}$ policy loss in slot t. The algorithm compares this certificate with a tolerance sequence $\{\varepsilon_t\}_{t \geq 1}$. The tolerance controls the tradeoff between priority exploration and residual decision error. A smaller ε_t requires a more informative partial order, while a larger ε_t permits more unresolved relations but increases the possible approximation loss. A typical choice is $\varepsilon_t = c_\varepsilon \sqrt{M \log T / t}$.

If $B_t^{\mathrm{ord}} \le \varepsilon_t$, the current priority uncertainty is sufficiently small for slot t, and the algorithm enters an exploitation step. It selects a linear extension $\hat{\pi}_t \in \mathcal{L}(\mathcal{P}_t)$. and perform optimistic dynamic programming just like in Algorithm 1. If $B_t^{\mathrm{ord}} > \varepsilon_t$, the current partial order is not sufficiently informative for near-optimal admission, and the algorithm performs a priority-probing step. A natural probing pair selection criterion is the current decision relevance $C_{ij,t} = a_{i,t}a_{j,t}G_{ij}(t)$. However, relevance alone does not capture how much information can be obtained from probing the pair in the current slot. For pair (i, j), a co-demand observation occurs with probability $q_{ij,t} = p_{i,t}p_{j,t}$, and each such observation provides a KL separation of order I_{ij}. We therefore define the instantaneous information rate $J_{ij,t} = p_{i,t}p_{j,t}I_{ij}$. The default pair-selection score is $S_{ij,t} = C_{ij,t}J_{ij,t}$. The algorithm selects $\{i_t, j_t\} \in \arg\max_{i,j\in\mathcal{U}_t} S_{ij,t}$. When the probed pair is active simultaneously, we update the pairwise log-likelihood ratio. If the ratio crosses a boundary, the corresponding edge is added and we update the transitive closure of $\mathcal{P}_t$. Whenever a device is served, the observed reward is used to update its reward estimate and confidence interval. Appendix D provides the complete procedure in Algorithm 2.

7 Performance Analysis of Relevance-Aware Learning

Throughout this section, let $\mathcal{E}_\pi$ denote the event that every priority relation accepted by the algorithm is correct, and let $\mathcal{E}_\theta$ denote the event that all reward confidence intervals simultaneously contain the true mean rewards. With the confidence allocations described in Appendix D, $\Pr(\mathcal{E}_\pi \cap \mathcal{E}_\theta) \ge 1 - \delta_\pi - \delta_\theta$. The following result formalizes the interpretation of the certificate in equation 6.

Theorem 5. *Suppose that $\mathcal{E}_\pi \cap \mathcal{E}_\theta$ holds. For any slot t, any two linear extensions $\pi_1, \pi_2 \in \mathcal{L}(\mathcal{P}_t)$, and any reward vector $x_i \in [L_i(t), U_i(t)]$, we have $\sup_A |F_t(A; \pi_1, x) - F_t(A; \pi_2, x)| \le B_t^{\mathrm{ord}}$. In particular, since the true priority order satisfies $\pi \in \mathcal{L}(\mathcal{P}_t)$, every candidate extension $\hat{\pi}_t \in \mathcal{L}(\mathcal{P}_t)$ satisfies $\sup_A |F_t(A; \pi, \theta) - F_t(A; \hat{\pi}_t, \theta)| \le B_t^{\mathrm{ord}}$.*

The certificate is therefore a decision-oriented uncertainty measure. It does not quantify whether the complete priority permutation has been recovered; instead, it bounds how much any remaining ambiguity can affect the current admission problem. Let N_T^{probe} be the total number of probing time slots, We have the following central performance guarantee for Algorithm 2.

Theorem 6. *For any predictable tolerance sequence $\{\varepsilon_t\}_{t=1}^T$, Algorithm 2 satisfies*

$$\mathbb{E}[R_T] \le \mathbb{E}[N_T^{\mathrm{probe}}] + 2\sum_{t=1}^T \varepsilon_t + C[M + \sigma\sqrt{MT\log(4MT/\delta_\theta)}] + T(\delta_\pi + \delta_\theta), \qquad (7)$$

for a universal constant $C > 0$. In particular, choosing $\varepsilon_t = c_\varepsilon/\sqrt{t}$ gives $\mathbb{E}[R_T] = \widetilde{O}(\mathbb{E}[N_T^{\mathrm{probe}}] + \sqrt{MT})$. Alternatively, choosing $\varepsilon_t = c_\varepsilon\sqrt{M/t}$ allows an order-approximation budget of the same scale as the reward-learning term.

This theorem isolates the benefit of relevance-aware learning into a single quantity N_T^{probe}, which can be substantially smaller in instances containing sufficiently screenable priority relations. We next derive an instance-dependent bound for N_T^{probe}. For an unresolved pair $e = (i, j)$, let $c_e(t)$ denote its contribution to the order-sensitivity certificate, so that $B_t^{\mathrm{ord}} = \sum_{e\in\mathcal{U}_t} c_e(t)$. Recall that $J_e(t) = q_e(t)I_e$ denote the information acquired from one probing slot, where $q_e(t)$ is the probability of an informative joint-activity event and I_e is the conditional information for distinguishing the two possible orders. Algorithm 2 probes only when $B_t^{\mathrm{ord}} > \varepsilon_t$, and selects the unresolved pair maximizing $s_e(t) := c_e(t)J_e(t)$. Thus $s_e(t)$ is a relevance score weighted by the rate at which priority information can be acquired in the current context.

Theorem 7 (Relevance-aware probe complexity)**.** *Let $\mathcal{E}$ denote the set of candidate device pairs. Fix a time t_0, and consider a good event $\mathcal{G}$, with $\Pr(\mathcal{G}) \ge 1 - \delta$, on which, for every $t \ge t_0$ and $e \in \mathcal{E}$, $c_e(t) \le \bar{c}_e, s_e(t) \le \bar{s}_e, J_e(t) \ge \underline{J}_e > 0$. For a set $D \subseteq \mathcal{E}$, define*

$$\Phi_T(D) := \textstyle\sum_{e\in D} \bar{c}_e + \big(\max_{e\in D} \bar{s}_e\big) \sum_{e\notin D} \frac{1}{\underline{J}_e},$$

with $\max_{e\in\emptyset} \bar{s}_e = 0$. We call D screenable if $\Phi_T(D) \le \varepsilon_T$. Let $L_e(\delta) = O(\log(|\mathcal{E}|T/\delta)/\underline{J}_e)$ be the expected number of pair-specific probing slots required by the sequential comparator to resolve pair e. Then we have $\mathbb{E}[N_T^{probe}] \le t_0 + \mathcal{C}_{\mathrm{RA}}(T, t_0, \delta) + \delta T$, where

$$\mathcal{C}_{\mathrm{RA}}(T, t_0, \delta) := \min_{D:\,\Phi_T(D)\le\varepsilon_T} \textstyle\sum_{e\notin D} L_e(\delta).$$

In particular, after t_0, Algorithm 2 never explicitly probes a pair belonging to any screenable set D.

Table 1: Favorable-regime scaling of relevance, information, RA score, and exact-identification cost. Logarithmic confidence factors and constants independent of the regime parameter are suppressed.

Regime	**Relevance** c_{ij}	**Information** J_{ij}	**RA score** $c_{ij}J_{ij}$	**Exact identification cost**
Reward gap Δ	$O(\Delta)$	$\Theta(1)$	$O(\Delta)$	$\widetilde{\Theta}(1)$
Core–low gate g	$O(g)$	$\Theta(g)$	$O(g^2)$	$\widetilde{\Theta}(g^{-1})$
Low–low gate g	$O(g^2)$	$\Theta(g^3)$	$O(g^5)$	$\widetilde{\Theta}(g^{-3})$
Rare overlap ρ	$O(\rho)$	$\Theta(\rho)$	$O(\rho^2)$	$\widetilde{\Theta}(\rho^{-1})$

Combining this theorem with Theorem 6 immediately gives $\mathbb{E}[R_T] \le t_0 + \mathcal{C}_{\mathrm{RA}}(T, t_0, \delta) + 2\sum_{t=1}^{T}\varepsilon_t + \widetilde{O}(\sqrt{MT}) + O(\delta T)$. Hence the priority-learning cost is controlled by the information complexity of only the pairs that cannot be screened by their downstream decision relevance.

Favorable Regimes. Theorem 7 also reveals when relevance-aware learning can be substantially cheaper than exact priority recovery. Table 1 summarizes three representative regimes. Here, Δ denotes a small reward contrast between a pair of devices, g denotes the gate probability of a low-impact device, and ρ denotes the joint-activity probability of a rarely co-active pair. We report the scaling of the pairwise decision relevance c_{ij}, the per-slot priority information J_{ij}, the relevance-aware probing score $c_{ij}J_{ij}$, and the cost required for exact pair identification. The central phenomenon is that a priority relation can become harder to identify while simultaneously becoming less useful to identify. Small reward gaps reduce the loss of a misordering without reducing its statistical identification cost; low gate probabilities and rare contextual overlap additionally suppress the information available for resolving the pair. Accordingly, the RA score decays as $O(g^2)$, $O(g^5)$, and $O(\rho^2)$ for core–low, low–low, and rare-overlap pairs, respectively, while exact identification requires $\widetilde{\Theta}(g^{-1})$, $\widetilde{\Theta}(g^{-3})$, and $\widetilde{\Theta}(\rho^{-1})$ probes. Appendix E.7 gives the full corollaries.

Decision-Theoretic Lower Bound. Our relevance-aware upper bound depends jointly on pairwise decision sensitivity and statistical distinguishability. We show that the same two quantities also determine an intrinsic difficulty of the problem. Specifically, we construct two instances differing only in the order of one pair (i, j), but having different optimal admission decisions. Distinguishing the two instances requires repeatedly taking an informative action that is suboptimal in one of them.

Theorem 8 (Informal)**.** *Let $\Delta_{ij}^{\mathrm{dec}}$ be the per-round loss of this informative action and $J_{ij} = q_{ij}I_{ij}$ the information obtained from one such round. For any uniformly efficient policy, there exists one of the two constructed instances such that*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R_T]}{\log T} \ge \frac{\Delta_{ij}^{\mathrm{dec}}}{J_{ij}}.$$

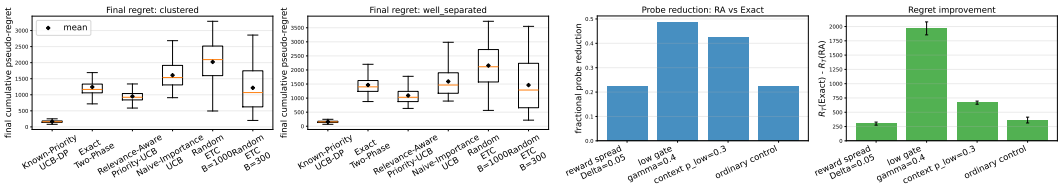
On a suitable family of instances, $\Delta_{ij}^{\mathrm{dec}} = \Theta(c_{ij})$, where c_{ij} is the local priority-misordering sensitivity. Thus, $R_T = \Omega\Big(\frac{c_{ij}}{q_{ij}I_{ij}}\log T\Big)$, showing that decision relevance relative to available priority information is an intrinsic component of the learning complexity.

The full construction and change-of-measure argument are deferred to Appendix E.8.

8 Experiments

We evaluate the proposed algorithms through simulation experiments. Unless otherwise specified, we consider $M = 16$ devices and a horizon of $T = 10^4$. The gate probabilities are randomly generated in the range $[0.5, 0.9]$, while the demand probabilities lie in $[0.3, 0.7]$. Rewards are Bernoulli with device-dependent means. Appendix F provides more experiment results, implementation details, parameter-generation procedures, intermediate diagnostic data, and detailed analysis.

We first compare the proposed Exact-Priority Two-Phase and Relevance-Aware (RA) algorithms with three baselines. Known-Priority UCB-DP is given the true priority order and only needs to learn the unknown reward means. Since it eliminates priority-learning uncertainty entirely, it serves as an oracle benchmark for the regret that remains when only reward estimation is required. Naive-Importance UCB is a oracle-assisted coarse-importance heuristic that may be used when approximate device importance is available. Random Explore-Then-Commit (Random ETC) allocates the first B rounds to random priority probing. For the reward parameters, we consider two configurations. In the "clustered" setting, reward means are generated around several cluster centers, approximately $0.30, 0.45, 0.60$, and 0.75. In the "well-separated" setting, reward means are evenly spread between 0.15 and 0.85 and then randomly assigned to devices.



(a) Final cumulative pseudo-regret under clustered and well-separated reward means.

(b) Empirical benefits of relevance-aware priority learning in favorable regimes.

Figure 1: Comparison of the proposed algorithms and baseline methods.

Figure 1a reports the final cumulative pseudo-regret across repeated runs. Among all algorithms without prior knowledge of the order, RA achieves the lowest regret in both settings. Compared with Exact-Priority Two-Phase, RA consistently reduces regret by avoiding the requirement of identifying the entire priority permutation before exploiting the learned structure. This difference is particularly relevant when several devices have similar rewards: a pairwise order can be statistically uncertain while having only a small effect on the value of the resulting admission policy. Exact-Priority continues to spend samples resolving such relations, whereas RA can tolerate them once their impact becomes sufficiently small.

Section 7 predicts three favorable regimes where precise priority identification becomes less valuable. Now we directly compare RA with Exact-Priority learning in these regimes. The first regime has small reward differences, where neighboring device reward means differ by only $\Delta = 0.05$. Misordering two devices with similar rewards then causes only a limited change in the resulting policy value. The second regime creates low-gate, low-impact devices. Devices are divided into a core set with relatively high gate probabilities and a second group whose gate probability is reduced to $\gamma = 0.4$. Since these latter devices are less likely to successfully terminate the competition, many priority relations involving them have reduced influence on the expected reward. The third regime introduces rare cross-group competition. Devices are partitioned into four groups. Devices within the same group are likely to be active together, whereas the cross-group simultaneous-activity probability is reduced to 0.3. Consequently, many cross-group priority relations are encountered only rarely in decision-relevant contexts. For comparison, we also include an ordinary control setting without deliberately imposed low-relevance structure.

Figure 1b summarizes both the reduction in priority probes, and the corresponding regret improvement. The results support the central motivation of relevance-aware learning. RA uses fewer priority probes than Exact-Priority in all tested regimes, but the reduction becomes particularly pronounced when the underlying competition structure makes many relations low-impact. The strongest effect is observed in the low-gate regime, where RA eliminates nearly half of the priority probing effort and yields a substantially larger regret reduction than in the ordinary setting. A similarly clear effect appears under rare cross-group activity: because many pairs seldom become jointly relevant, RA can safely leave these relations unresolved and devote exploration to the smaller subset that affects actual admission decisions. The small-reward-gap experiment shows a more moderate improvement. Although small reward differences directly reduce the loss caused by a local priority inversion, reward values must themselves be learned online, and the algorithm cannot immediately certify that two devices are nearly interchangeable. Consequently, the reduction in probing is less dramatic than in the low-gate and low-overlap regimes. Nevertheless, RA still obtains lower regret without requiring exact recovery of the full order.

9 Conclusion

We investigated ex-ante admission under stochastic priority-based resource competition. Our main finding is that exact priority identification is generally stronger than what near-optimal admission requires. By analyzing the sensitivity of the admission value to local priority misordering, we showed that different unresolved pairwise relations can have highly heterogeneous decision consequences. This led to the Relevance-Aware Priority-UCB algorithm, which adaptively suppresses priority exploration once the residual ordering uncertainty becomes sufficiently benign for the downstream decision problem. The resulting analysis connects regret to both the information available for distinguishing a priority relation and the loss incurred when that relation remains unresolved. Theoretical guarantees and simulations consistently support this perspective.

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks, design or provide feedback on research methodology or experiments, implement methods, and assist with translation. We have not used generative AI tools to generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, support qualitative and thematic data analysis, and interpret results, and clean and reformat dataset task is not applicable to this work. Additionally, we used generative AI tools to create or edit software code, draft parts of the research paper, summarize or analyse existing literature, sourcing/searching for information, edit the research paper to improve readability, and identify relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation was involved. All datasets used were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

All theoretical results are supported by complete proofs provided in the appendices. Algorithms are fully specified in the main text and appendices, including pseudocode and complexity analyses. The assumptions required for our results are explicitly stated and discussed in the problem definition and algorithmic sections. Our experimental evaluation is based on synthetic data generated according to a well-defined random procedure, with details given in Appendix F. We will release the source code of all algorithms and experiments upon publication. The experiments rely only on standard Python libraries and require modest computational resources.

REFERENCES

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in neural information processing systems*, 29, 2016.

Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, 2003.

Ali Aouad, Vivek Farias, and Retsef Levi. Assortment optimization under consider-then-choose choice models. *Management Science*, 67(6):3368–3386, June 2021. doi: 10.1287/mnsc.2020.3681. URL https://ideas.repec.org/a/inm/ormnsc/v67y2021i6p3368-3386.html.

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

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

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.

Viktor Bengs and Eyke Hüllermeier. Preselection bandits. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 778–787. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/bengs20a.html.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7): 1–108, 2021.

Lilian Besson and Emilie Kaufmann. Multi-Player Bandits Revisited. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pp. 56–92. PMLR, 07–09 Apr 2018.

Etienne Boursier and Vianney Perchet. Sic-mmab: Synchronisation involves communication in multiplayer multi-armed bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Sébastien Bubeck, Yuanzhi Li, Yuval Peres, and Mark Sellke. Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without. In *Conference on Learning Theory*, pp. 961–987. PMLR, 2020.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pp. 151–159. PMLR, 2013.

Guillermo Gallego and Anran Li. A random consideration set model for demand estimation, assortment optimization, and pricing. *Operations Research*, 72(6):2358–2374, November 2024. ISSN 0030-364X. doi: 10.1287/opre.2019.0333. Publisher Copyright: © 2024 INFORMS.

Xinyi Hu and Aldo Pacchiano. Meet me at the arm: The cooperative multi-armed bandits problem with shareable arms. *arXiv preprint arXiv:2506.10127*, 2025.

Wei Huang, Richard Combes, and Cindy Trinh. Towards optimal algorithms for multi-player bandits without collision sensing information. In *Conference on Learning Theory*, pp. 1990–2012. PMLR, 2022.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. *Advances in neural information processing systems*, 30, 2017.

Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 767–776, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/kveton15.html.

Paul Lagrée, Claire Vernade, and Olivier Cappe. Multiple-play bandits in the position-based model. *Advances in Neural Information Processing Systems*, 29, 2016.

Keqin Liu and Qing Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, 58(11):5667–5681, 2010.

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.

Wentao Lu. Online assortment optimization for random consideration set choice model. *submitted to POMS*, 2023.

Wentao Lu. Bounds on revenue for the random consideration set choice model. *Operations Research Letters*, 53:107070, 2024. ISSN 0167-6377. doi: https://doi.org/10.1016/j.orl.2024.107070. URL https://www.sciencedirect.com/science/article/pii/S0167637724000063.

Gábor Lugosi and Abbas Mehrabian. Multiplayer bandits without observing collision information. *Mathematics of Operations Research*, 47(2):1247–1265, 2022.

Paola Manzini and Marco Mariotti. Stochastic choice and consideration sets. *Econometrica*, 82(3):1153–1176, 2014. doi: https://doi.org/10.3982/ECTA10575. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA10575.

Hyung G Myung. Technical overview of 3gpp lte. *Polytechnic University of New York*, 2008.

Jonathan Rosenski, Ohad Shamir, and Liran Szlak. Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*, pp. 155–163. PMLR, 2016.

Aadirupa Saha and Aditya Gopalan. Combinatorial bandits with relative feedback. *Advances in Neural Information Processing Systems*, 32, 2019.

Xuchuang Wang, Hong Xie, and John C. S. Lui. Multi-player multi-armed bandits with finite shareable resources arms: Learning algorithms & applications. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 3537–3543, 7 2022a.

Xuchuang Wang, Hong Xie, and John CS Lui. Multiple-play stochastic bandits with shareable finite-capacity arms. In *International Conference on Machine Learning*, pp. 23181–23212. PMLR, 2022b.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pp. 1254–1262. PMLR, 2016.

Hong Xie, Haoran Gu, Yanying Huang, Tao Tan, and Defu Lian. Multiple-play stochastic bandits with prioritized arm capacity sharing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36766–36774, 2026.

Renzhe Xu, Haotian Wang, Xingxuan Zhang, Bo Li, and Peng Cui. Competing for shareable arms in multi-player multi-armed bandits. In *International Conference on Machine Learning*, pp. 38674–38706. PMLR, 2023.

Daoyuan Zhou, Xuchuang Wang, Lin Yang, and Yang Gao. Distributed algorithms for multi-agent multi-armed bandits with collision. *arXiv preprint arXiv:2510.06683*, 2025.

A Related Work

A.1 Random Consideration, Assortment, and Preselection Bandits

Random consideration and assortment optimization. Our model is closely related to the random consideration set (RCS) literature. Manzini & Mariotti (2014) introduced a stochastic choice model in which each available alternative independently enters a consideration set according to an attention probability, after which the decision maker chooses the most preferred considered alternative according to a fixed preference order. Consequently, the probability of choosing item i from an assortment S takes the product form $a_i \Pi_{j \in S: j \succ i}(1 - a_j)$. Gallego & Li (2024) operationalized this model by studying identification of the preference order and attention probabilities from transaction data, together with revenue-maximizing assortment optimization and cardinality-constrained variants. More broadly, consider-then-choose models have been extensively studied in assortment optimization as a mechanism for separating random or limited consideration from subsequent choice (Aouad et al., 2021). Most closely related to our online setting is Lu (2023), who studies online assortment optimization under the RCS model when the preference order, the attention probabilities, or both are unknown. The proposed learning schemes exploit binary relations rather than full parameter estimation within an explore-then-commit framework, and the work also considers nonstationary preference and attention parameters. Lu (2024) further studies revenue perturbation bounds under changes of RCS parameters, whereas our sensitivity bounds are context-dependent pairwise certificates constructed specifically to control online priority acquisition under jointly unknown rewards. These results establish an important connection between structural learning and assortment revenue under RCS.

Preselection and cascade bandits. A related subset-selection perspective appears in the Preselection Bandit problem of Bengs & Hüllermeier (2020), where the learner first presents a subset and a user subsequently selects an arm according to a Plackett–Luce model; regret upper and lower bounds are developed for learning subsets that lead to high-utility choices. In contrast to RCS, randomness there acts directly through the stochastic preference model rather than through independent consideration followed by a deterministic order. Cascading bandits Kveton et al. (2015) exhibit another related product-form structure: a user scans a learner-chosen list sequentially and stops at the first attractive item, yielding censored feedback about subsequent items. The classical cascading-bandit objective is to maximize the probability of at least one attractive item, so the selected set, rather than its internal order, determines the expected payoff.

Positioning of our work. At the single-round choice level, our stochastic-priority mechanism admits a direct RCS interpretation: an admitted device i effectively enters the service consideration set with probability $a_{i,t} = p_{i,t}\gamma_i$, and the served device is the highest-priority member of this random set. We therefore do not view the product-form choice mechanism itself as the source of novelty. Our focus is instead on the online learning problem created by three additional features. First, the effective consideration probabilities $a_{i,t}$ are observed before admission and may vary from round to round, making the relevance of a priority relation inherently context dependent. Second, the service reward of each device is itself stochastic and unknown, and is conceptually separate from the latent priority order; hence both reward uncertainty and structural uncertainty jointly determine the optimal admission decision. Third, rather than treating recovery of the complete preference/priority structure as the learning objective, we explicitly quantify the downstream value loss caused by unresolved pairwise relations and allocate priority exploration accordingly. The resulting relevance-aware algorithm may intentionally leave statistically unresolved relations unidentified when their effect on admission value is small. Our regret analysis and decision-theoretic lower bound characterize this interaction between statistical distinguishability and decision relevance, which is the main focus distinguishing our work from parameter-identification and explore-then-commit approaches to RCS assortment learning.

A.2 Other Related Literature

Collision-based multi-player bandits. A classical line of work studies multi-player multi-armed bandits (MMAB) where multiple players simultaneously pull from the same set of arms and collisions occur when two or more players choose the same arm (Liu & Zhao, 2010). A common collision model assumes colliding players receive zero reward, and the goal is to jointly learn arm qualities while avoiding conflicts, typically by achieving an implicit orthogonalization of players over the best

arms (Zhou et al., 2025; Liu & Zhao, 2010; Rosenski et al., 2016; Besson & Kaufmann, 2018; Boursier & Perchet, 2019). Related progress includes logarithmic-regret, gap-free guarantees (Lugosi & Mehrabian, 2022), no-sensing algorithms(Huang et al., 2022), and adversarial multi-player bandits (Bubeck et al., 2020). Our problem differs from collision-based MMAB in both the contention resolution mechanism and the induced learning objective. In collision models, contention is typically treated as a pure inefficiency that can be mitigated by coordination, and the core algorithmic task is to learn while avoiding conflicts. In our setting, contention is resolved by a priority mechanism under a finite service capacity, which makes "squeezing out" (i.e., some demand being dropped due to insufficient priority) structurally unavoidable under ex-ante admission. Consequently, the focus shifts from eliminating collisions to learning and optimizing under inevitable priority-induced censoring and cross-unit coupling.

Shareable arms with finite capacities or prioritized capacity allocation. Several recent models allow multiple players to share the same arm while accounting for limited and shareable resources. One representative formulation assumes each arm has an unknown finite shareable capacity and an unknown per-load reward distribution (Wang et al., 2022b;a). The realized reward is load-dependent and saturates when demand exceeds capacity. Further works focus on the Nash equilibrium-guided learning and logarithmic regret (Xu et al., 2023; Hu & Pacchiano, 2025). The most closely related work to ours is the recent multiple-play model with prioritized capacity allocation (Xie et al., 2026). It associates each play with a priority weight and allocates each arm's stochastic capacity to competing plays in decreasing priority order. Despite this surface similarity, our setting differs in key modeling and technical aspects. First, existing capacity-sharing model attaches capacities to arms, while our focus is on competition for a shared resource with unit capacity. Second, in the above works priority is typically an exogenous attribute known a priori, while our formulation treats the priority mechanism as unknown and explicitly leverages priority probing as a learning primitive. Third, our ex-ante admission constraint restricts the learner to decisions based on statistical context rather than per-slot realized demands, making priority-induced drop events unavoidable and introducing a distinct censored feedback structure.

Ranking feedback and competing bandits in matching markets. These literature studies learning with relative or ordinal feedback rather than absolute rewards. The dueling-bandit replaces numeric rewards with pairwise comparisons or more general preference signals (Bengs et al., 2021; Saha & Gopalan, 2019). Competing bandits in matching markets incorporate priority-like structure via two-sided preferences (Liu et al., 2020; Basu et al., 2021; Xie et al., 2026). Our work differs from ranking-feedback and matching-market bandits in both the source of relative outcomes and the optimization objective. In ranking bandits, relative feedback is typically directly elicited, and the goal is to identify and optimize with respect to an underlying preference model. In matching markets, the benchmark and dynamics are tied to stable matchings. In contrast, our model is challenged by unavoidable demand squeezing and censored rewards. The objective is system utility maximization under this competition mechanism rather than preference elicitation or stability.

Knapsack, constrained, and conservative bandits. Bandits with knapsacks and constrained bandits study online learning under resource limitations, where each action consumes one or multiple resources, and the process is constrained by cumulative budgets (Badanidiyuru et al., 2018; Agrawal & Devanur, 2016). Conservative bandits enforce constraints such as maintaining performance above a baseline or avoiding unsafe arms defined by risk thresholds (Wu et al., 2016; Kazerouni et al., 2017). These constrained-learning frameworks are conceptually related in that they impose additional feasibility or safety requirements. However, they do not capture the per-round, competition-driven service mechanism central to our model. The key challenge in our work is not a cumulative budget tradeoff, but learning and optimizing under unavoidable priority-induced squeezing in a multi-unit competitive environment.

B PROOFS FOR THE TWO-PHASE BASELINE

B.1 LOG-LIKELIHOOD RATIO BASED SEQUENTIAL NOISY COMPARATOR.

Consider devices i and j, to compare their priorities, the controller admits only these two devices, and we focus on slots in which both devices realize demand: $D_{i,t} = D_{j,t} = 1$. Conditional on this co-demand event, the service outcome O_t is i, j or 0, where $O_t = 0$ means that both gates fail.

We need to test the two hypotheses $H_{ij}^+ : i \succ j$ and $H_{ij}^- : j \succ i$. Under H_{ij}^+, the outcome distribution is

$$P_{ij}^+ = (\gamma_i, (1-\gamma_i)\gamma_j, (1-\gamma_i)(1-\gamma_j)),$$

where the three coordinates correspond to $O_t = i, j, 0$, respectively. Under H_{ij}^-,

$$P_{ij}^- = ((1-\gamma_j)\gamma_i, \gamma_j, (1-\gamma_i)(1-\gamma_j)).$$

Therefore, the log-likelihood ratio associated with one co-demand observation is

$$\ell_{ij}(O_t) = \log \frac{P_{ij}^+(O_t)}{P_{ij}^-(O_t)} = \begin{cases} -\log(1-\gamma_j), & O_t = i, \\ \log(1-\gamma_i), & O_t = j, \\ 0, & O_t = 0. \end{cases}$$

Let

$$L_{ij}(n) = \textstyle\sum_{s=1}^{n} \ell_{ij}(O_s)$$

denote the cumulative log-likelihood ratio after n co-demand observations. For a target comparison error probability δ_{ij}, define the threshold $b_{ij} = \log(2/\delta_{ij})$. Then the pairwise comparator stops at

$$N_{ij} = \inf\{n \geq 1 : L_{ij}(n) \notin (-b_{ij}, b_{ij})\}.$$

It returns $i \succ j$ if $L_{ij}(N_{ij}) \geq b_{ij}$, and $j \succ i$ if $L_{ij}(N_{ij}) \leq -b_{ij}$. Define the directional information quantities

$$d_{ij}^+ = D_{\mathrm{KL}}\left(P_{ij}^+ \| P_{ij}^-\right)$$

and

$$d_{ij}^- = D_{\mathrm{KL}}\left(P_{ij}^- \| P_{ij}^+\right),$$

and let

$$I_{ij} = \min\{d_{ij}^+, d_{ij}^-\}.$$

We have the following theorem on the correctness and stopping time of the comparator.

Theorem 9 (Pairwise comparison)**.** *For any pair (i, j), the sequential comparator satisfies* $\Pr(\textit{incorrect comparison}) \leq \delta_{ij}$. *Moreover, the expected number of co-demand observations required to terminate satisfies* $\mathbb{E}[N_{ij}] = O\left(\log(1/\delta_{ij})/I_{ij}\right)$. *Suppose that, whenever pair (i, j) is being probed, $p_{i,t}p_{j,t} \geq q_{ij} > 0$. Then the expected number of actual slots required by the comparator is* $\mathbb{E}[\tau_{ij}] = O\left(\log(1/\delta_{ij})/q_{ij}I_{ij}\right)$.

When all gate probabilities are bounded away from zero and one, I_{ij} is bounded below by a positive constant. In contrast, if $\gamma_i = \gamma_j = \gamma$ and γ is small, then $I_{ij} = \gamma^2[-\log(1-\gamma)] = \Theta(\gamma^3)$. Priority identification may therefore become statistically expensive when successful gates are rare.

B.2 EXACT-PRIORITY TWO-PHASE BASELINE

We embed the sequential comparator into a comparison-based sorting procedure. In particular, merge sort recovers a permutation using at most $K_M = O(M \log M)$ comparisons. Let $\mathcal{C}$ denote the set of pairs compared by the sorting procedure. To ensure the entire permutation is recovered with probability at least $1 - \delta_\pi$, we can assign $\delta_{ij} = \delta_\pi / K_M$.

Corollary 1 (Exact priority recovery)**.** *Let $\hat{\pi}$ be the permutation returned by merge sort equipped with the sequential comparator. Then* $\Pr(\hat{\pi} = \pi) \geq 1 - \delta_\pi$. *The expected duration of the priority-learning phase satisfies* $\mathbb{E}[\tau_\pi] = O(\sum_{(i,j)\in\mathcal{C}} \log(K_M/\delta_\pi)/q_{ij}I_{ij})$. *In particular, if $q_{ij} \geq \underline{q}$ and $I_{ij} \geq I_{\min}$ for every compared pair, then* $\mathbb{E}[\tau_\pi] = \widetilde{O}(M \log M / \underline{q} I_{\min})$.

During priority exploration, the controller observes a reward whenever one of the two probed devices is served. These observations are retained and used in the subsequent reward-learning phase. After obtaining $\hat{\pi}$, the algorithm learns the unknown mean rewards using upper confidence bounds. Let

$$N_i(t-1) = \sum_{s<t} \mathbf{1}\{I_s = i\}$$

be the number of observed rewards from device i before slot t, including observations collected during the priority-learning phase. Let $\hat{\theta}_i(t)$ denote the corresponding empirical mean. Define

$$\beta_i(t) = \sigma\sqrt{2\log(4MT/\delta_\theta)/N_i(t-1)},$$

and $U_i(t) = \min\{1, \hat{\theta}_i(t) + \beta_i(t)\}$. In slot t, the algorithm computes $a_{i,t} = p_{i,t}\gamma_i$ and runs the oracle dynamic program under the estimated priority order $\hat{\pi}$ and the optimistic reward vector $U(t)$. Specifically, it initializes

$$\overline{V}_{M+1,t} = 0$$

and, for $k = M, M-1, \ldots, 1$, with $i = \hat{\pi}(k)$, computes

$$\overline{V}_{k,t} = \max\{\overline{V}_{k+1,t}, a_{i,t}U_i(t) + (1-a_{i,t})\overline{V}_{k+1,t}\}.$$

The corresponding maximizing decisions form the admission set A_t. We refer to this baseline two-phase method as Exact-Priority Two-Phase Learning. Alg. 1 summarizes the implementation details.

Algorithm 1 Exact-Priority Two-Phase Learning (EP)

Require: Horizon T, number of devices M, gating probabilities γ_i

Phase I: exact priority learning.
Run merge sort on the devices.
When the sorting procedure requests a comparison between i and j, repeatedly admit only $\{i, j\}$.
Update the pairwise log-likelihood ratio on co-demand slots.
Stop the comparison when its likelihood ratio crosses a boundary.
Continue until a complete permutation $\hat{\pi}$ is obtained.
Use all observed service rewards to update the reward estimates.
Phase II: optimistic reward learning.
Construct reward upper confidence bounds $U_i(t)$.
Run the oracle dynamic program using $\hat{\pi}$, $U(t)$, and $a_{i,t} = p_{i,t}\gamma_i$.
Execute the resulting admission set and update the reward estimate of the served device.

B.3 Proof of Theorem 9

We provide a complete analysis of the sequential likelihood-ratio comparator introduced in Section 4. Throughout the proof, we assume $0 < \gamma_i, \gamma_j < 1$. The boundary cases in which one of the gate probabilities equals one are discussed at the end. Consider an isolated comparison between devices i and j. The two hypotheses are $H_{ij}^+ : i \succ_\pi j, H_{ij}^- : j \succ_\pi i$. The comparator updates its statistic only on slots in which both devices have demand. Conditional on such a co-demand event, the service outcome $O \in \{i, j, 0\}$ has distribution

$$P_{ij}^+ = (\gamma_i,\ (1-\gamma_i)\gamma_j,\ (1-\gamma_i)(1-\gamma_j))$$

under H_{ij}^+, and

$$P_{ij}^- = ((1-\gamma_j)\gamma_i,\ \gamma_j,\ (1-\gamma_i)(1-\gamma_j))$$

under H_{ij}^-. For brevity, write $P^+ = P_{ij}^+, P^- = P_{ij}^-$. For the n-th co-demand observation O_n, define $X_n = \log(P^+(O_n)/P^-(O_n))$. The cumulative log-likelihood ratio is $L_n = \sum_{s=1}^n X_s, L_0 = 0$. Its one-step increments are

$$X_n = \begin{cases} u_{ij} := -\log(1-\gamma_j) > 0, & O_n = i, \\ -v_{ij} := \log(1-\gamma_i) < 0, & O_n = j, \\ 0, & O_n = 0, \end{cases}$$

where $v_{ij} = -\log(1-\gamma_i)$. Let $b = \log(2/\delta)$ be the symmetric likelihood boundary, and define the stopping time, measured in the number of co-demand observations, as $N_{ij} =$

$\inf\{n \geq 1 : L_n \notin (-b, b)\}$. The comparator declares $i \succ j$ if $L_{N_{ij}} \geq b$, and declares $j \succ i$ if $L_{N_{ij}} \leq -b$. We prove correctness and the stopping-time bound separately.

Let $\mathcal{F}_n = \sigma(O_1, \ldots, O_n)$ be the filtration generated by the co-demand observations. Under H_{ij}^-, define

$$M_n^+ = \exp(L_n) = \prod_{s=1}^n \frac{P^+(O_s)}{P^-(O_s)}.$$

Then $M_0^+ = 1$. Moreover,

$$\mathbb{E}_-\left[M_{n+1}^+ \mid \mathcal{F}_n\right] = M_n^+ \, \mathbb{E}_-\left[\frac{P^+(O_{n+1})}{P^-(O_{n+1})}\right] = M_n^+ \sum_{o \in \{i,j,0\}} P^-(o) \frac{P^+(o)}{P^-(o)} = M_n^+ \sum_o P^+(o) = M_n^+.$$

Hence, $\{\exp(L_n)\}_{n \geq 0}$ is a nonnegative martingale under H_{ij}^-. Similarly, under H_{ij}^+,

$$M_n^- = \exp(-L_n) = \prod_{s=1}^n \frac{P^-(O_s)}{P^+(O_s)}$$

is a nonnegative martingale with initial value one. These are the standard likelihood-ratio martingales associated with the two simple hypotheses.

Suppose first that the true relation is $H_{ij}^+ : i \succ j$. An incorrect declaration occurs only if the lower boundary is crossed, i.e. $L_{N_{ij}} \leq -b$. Therefore,

$$\Pr_+(\text{incorrect comparison}) \leq \Pr_+\left(\inf_{n \geq 0} L_n \leq -b\right) = \Pr_+\left(\sup_{n \geq 0} e^{-L_n} \geq e^b\right).$$

Since e^{-L_n} is a nonnegative martingale under H_{ij}^+, Ville's inequality gives

$$\Pr_+\left(\sup_{n \geq 0} e^{-L_n} \geq e^b\right) \leq e^{-b}.$$

Consequently,

$$\Pr_+(\text{incorrect comparison}) \leq e^{-b}.$$

The same argument under H_{ij}^- uses the martingale e^{L_n}. An incorrect comparison now requires the upper boundary to be crossed, and therefore

$$\Pr_-(\text{incorrect comparison}) \leq \Pr_-\left(\sup_{n \geq 0} e^{L_n} \geq e^b\right) \leq e^{-b}.$$

With $b = \log(2/\delta)$, we obtain $e^{-b} = \delta/2$. Thus, under either priority relation,

$$\Pr(\text{incorrect comparison}) \leq \tfrac{\delta}{2} \leq \delta.$$

The extra factor of two in the chosen threshold is not essential; a threshold $b = \log(1/\delta)$ would already yield an error probability at most δ for each fixed true hypothesis.

Under H_{ij}^+,

$$\mathbb{E}_+[X_n] = D_{\mathrm{KL}}(P^+ \| P^-) = d_{ij}^+ > 0.$$

Using the explicit outcome probabilities,

$$d_{ij}^+ = \gamma_i\left[-\log(1-\gamma_j)\right] + (1-\gamma_i)\gamma_j \log(1-\gamma_i) = \gamma_i \log \tfrac{1}{1-\gamma_j} + (1-\gamma_i)\gamma_j \log(1-\gamma_i).$$

Similarly, under H_{ij}^-,

$$\mathbb{E}_-[X_n] = -D_{\mathrm{KL}}(P^- \| P^+) = -d_{ij}^-,$$

where

$$d_{ij}^- = \gamma_j \log \tfrac{1}{1-\gamma_i} + (1-\gamma_j)\gamma_i \log(1-\gamma_j).$$

Because $P^+ \neq P^-$ whenever $\gamma_i, \gamma_j > 0$, both KL divergences are strictly positive. Define $I_{ij} = \min\{d_{ij}^+, d_{ij}^-\}$. Under H_{ij}^+, L_n therefore has positive drift d_{ij}^+, whereas under H_{ij}^- it has negative drift $-d_{ij}^-$. In particular, by the strong law of large numbers,

$$\tfrac{L_n}{n} \longrightarrow d_{ij}^+ > 0 \quad \text{a.s. under } H_{ij}^+,$$

and

$$\tfrac{L_n}{n} \longrightarrow -d_{ij}^- < 0 \quad \text{a.s. under } H_{ij}^-.$$

Hence,

$$N_{ij} < \infty \qquad \text{almost surely}$$

under either hypothesis.

We now derive a finite-sample upper bound on $\mathbb{E}[N_{ij}]$. We first consider H_{ij}^{+}. For an integer $m \geq 1$, define the truncated stopping time $N_m = N_{ij} \wedge m$. Since

$$L_{N_m} = \sum_{n=1}^{m} X_n \, \mathbf{1}\{N_{ij} \geq n\},$$

and the event $\{N_{ij} \geq n\}$ depends only on $O_1, \ldots, O_{n-1}$, it is $\mathcal{F}_{n-1}$-measurable. Therefore,

$$\begin{aligned}\mathbb{E}_{+}[L_{N_m}] &= \sum_{n=1}^{m} \mathbb{E}_{+}\left[X_n \mathbf{1}\{N_{ij} \geq n\}\right] \\ &= \sum_{n=1}^{m} \mathbb{E}_{+}\left[\mathbf{1}\{N_{ij} \geq n\} \mathbb{E}_{+}[X_n \mid \mathcal{F}_{n-1}]\right] \\ &= d_{ij}^{+} \sum_{n=1}^{m} \Pr_{+}(N_{ij} \geq n) = d_{ij}^{+} \mathbb{E}_{+}[N_m].\end{aligned}$$

This is the stopped-sum identity normally obtained from Wald's equation, written here directly so that no integrability assumption on the untruncated stopping time is required.

We next bound L_{N_m} pathwise. If $N_{ij} > m$, then the test has not yet stopped and therefore $L_m < b$. If $N_{ij} \leq m$ and the upper boundary is crossed, immediately before the final observation we have $L_{N_{ij}-1} < b$. The largest possible positive LLR increment is $u_{ij} = -\log(1-\gamma_j)$. Hence, $L_{N_{ij}} < b + u_{ij}$. If instead the lower boundary is crossed, $L_{N_{ij}} \leq -b < b + u_{ij}$. Thus, in every case, $L_{N_m} \leq b + u_{ij}$. Combining this with the stopped-sum identity gives $d_{ij}^{+}\mathbb{E}_{+}[N_m] \leq b + u_{ij}$. Therefore,

$$\mathbb{E}_{+}[N_m] \leq \tfrac{b - \log(1-\gamma_j)}{d_{ij}^{+}}.$$

Since $N_m \uparrow N_{ij}$, the monotone convergence theorem yields

$$\mathbb{E}_{+}[N_{ij}] \leq \tfrac{b - \log(1-\gamma_j)}{d_{ij}^{+}}.$$

The argument under H_{ij}^{-} is symmetric. Consider $-L_n$, which has positive mean increment d_{ij}^{-}. The largest positive increment of $-L_n$ is $v_{ij} = -\log(1-\gamma_i)$. Repeating the preceding argument gives

$$\mathbb{E}_{-}[N_{ij}] \leq \tfrac{b - \log(1-\gamma_i)}{d_{ij}^{-}}.$$

Consequently,

$$\sup_{h \in \{+,-\}} \mathbb{E}_h[N_{ij}] \leq \max\left\{ \tfrac{b - \log(1-\gamma_j)}{d_{ij}^{+}}, \; \tfrac{b - \log(1-\gamma_i)}{d_{ij}^{-}} \right\}.$$

This expression explicitly accounts for the finite overshoot of the likelihood ratio beyond the stopping boundary. If the gate probabilities are uniformly bounded away from one, i.e., $\gamma_i, \gamma_j \leq \bar{\gamma} < 1$, then $-\log(1-\gamma_i), -\log(1-\gamma_j) \leq -\log(1-\bar{\gamma}) =: c_{\bar{\gamma}}$. Since $b = \Theta(\log(1/\delta))$ and $b \geq \log 2$, the preceding finite-sample bound implies $\mathbb{E}_h[N_{ij}] \leq C_{\bar{\gamma}} \frac{\log(1/\delta)}{I_{ij}}$ for a constant $C_{\bar{\gamma}}$ depending only on the upper gate bound. Thus,

$$\mathbb{E}[N_{ij}] = O\left(\tfrac{\log(1/\delta)}{I_{ij}}\right).$$

For fixed γ_i, γ_j, the same expression also holds asymptotically as $\delta \to 0$, since the overshoot term is independent of δ.

We finally convert the number of informative co-demand observations into the number of physical probing slots. Let τ_{ij} denote the number of actual slots spent probing pair (i,j) until the SPRT terminates. For the s-th probing slot, define $C_s = \mathbf{1}\{D_{i,s} = D_{j,s} = 1\}$. Then, pathwise, $N_{ij} = \sum_{s=1}^{\tau_{ij}} C_s$. Let $\mathcal{G}_{s-1}$ denote the information available immediately before the demand realizations in the s-th probing slot. The event $\{\tau_{ij} \geq s\}$ is $\mathcal{G}_{s-1}$-measurable. Suppose that throughout the probing process, $p_{i,s}p_{j,s} \geq q_{ij} > 0$. By conditional independence of the demands,

$$\mathbb{E}[C_s \mid \mathcal{G}_{s-1}] = p_{i,s}p_{j,s} \geq q_{ij}.$$

Using Tonelli's theorem,

$$\begin{aligned}\mathbb{E}[N_{ij}] &= \mathbb{E}\left[\textstyle\sum_{s=1}^{\infty} \mathbf{1}\{\tau_{ij} \geq s\} C_s\right] = \sum_{s=1}^{\infty} \mathbb{E}\left[\mathbf{1}\{\tau_{ij} \geq s\} C_s\right] \\ &= \textstyle\sum_{s=1}^{\infty} \mathbb{E}\left[\mathbf{1}\{\tau_{ij} \geq s\} \mathbb{E}[C_s \mid \mathcal{G}_{s-1}]\right] \geq q_{ij} \sum_{s=1}^{\infty} \Pr(\tau_{ij} \geq s) = q_{ij}\, \mathbb{E}[\tau_{ij}].\end{aligned}$$

Hence,

$$\mathbb{E}[\tau_{ij}] \leq \tfrac{\mathbb{E}[N_{ij}]}{q_{ij}}.$$

Combining this with the directional bounds above gives, under H_{ij}^{+},

$$\mathbb{E}_{+}[\tau_{ij}] \le \frac{b-\log(1-\gamma_j)}{q_{ij}d_{ij}^{+}},$$

and under H_{ij}^{-},

$$\mathbb{E}_{-}[\tau_{ij}] \le \frac{b-\log(1-\gamma_i)}{q_{ij}d_{ij}^{-}}.$$

Therefore, when the gate probabilities are bounded away from one,

$$\mathbb{E}[\tau_{ij}] = O\left(\frac{\log(1/\delta)}{q_{ij}I_{ij}}\right).$$

If $p_{i,t} \ge \underline{p}$ and $p_{j,t} \ge \underline{p}$ during all probing slots, we may set $q_{ij} = \underline{p}^2$, yielding $\mathbb{E}[\tau_{ij}] = O\left(\frac{\log(1/\delta)}{\underline{p}^2 I_{ij}}\right)$. This completes the proof. □

B.4 Proof of Theorem 1

We prove the regret guarantee by separately controlling the priority-learning phase, the reward-learning phase, and the low-probability failure events.

Priority-Recovery Event and Stopping Time. Let $\mathcal{E}_\pi = \{\widehat{\pi} = \pi\}$ denote the event that the first phase returns the correct priority permutation. The sorting algorithm invokes at most K_M pairwise comparators. Each comparator is assigned error probability $\delta_{ij} = \frac{\delta_\pi}{K_M}$. By Corollary 1, every comparator returns an incorrect relation with probability at most δ_π/K_M. Therefore, a union bound gives $\Pr(\mathcal{E}_\pi^c) \le \delta_\pi$.

We next bound the duration of the priority-learning phase. Consider one comparator between devices i and j. Conditional on a co-demand observation, the log-likelihood ratio has expected increment $d_{ij}^{+} = D_{\mathrm{KL}}(P_{ij}^{+}\|P_{ij}^{-})$ under $i \succ j$, and expected decrement of magnitude $d_{ij}^{-} = D_{\mathrm{KL}}(P_{ij}^{-}\|P_{ij}^{+})$ under $j \succ i$. Hence, under either hypothesis, the drift toward the correct boundary has magnitude at least $I_{ij} = \min\{d_{ij}^{+}, d_{ij}^{-}\}$. The comparator uses threshold

$$b = \log\frac{2K_M}{\delta_\pi}.$$

For $0 < \gamma_i, \gamma_j < 1$, the log-likelihood increments are bounded. Standard stopping-time bounds for the sequential probability ratio test therefore imply

$$\mathbb{E}[N_{ij}] \le C\frac{b}{I_{ij}},$$

where N_{ij} is the number of co-demand observations and $C > 0$ is a universal constant that absorbs the bounded overshoot. If a gate probability equals one, the comparison produces deterministic or one-sided evidence and satisfies the same upper bound after an appropriate limiting interpretation.

Let τ_{ij} be the number of actual probing slots required by this comparator. During every such slot, $\Pr(D_{i,t} = D_{j,t} = 1 \mid \mathcal{F}_{t-1}) = p_{i,t}p_{j,t} \ge \underline{q}$. If $C_{ij}(s)$ denotes the indicator that the s-th probing slot produces a co-demand observation, then $\mathbb{E}[C_{ij}(s) \mid \mathcal{F}_{s-1}] \ge \underline{q}$. Applying the tower property up to the stopping time gives

$$\mathbb{E}[N_{ij}] = \mathbb{E}\left[\sum_{s=1}^{\tau_{ij}} C_{ij}(s)\right] \ge \underline{q}\mathbb{E}[\tau_{ij}].$$

Consequently,

$$\mathbb{E}[\tau_{ij}] \le C\frac{\log(2K_M/\delta_\pi)}{\underline{q}I_{ij}} \le C\frac{\log(2K_M/\delta_\pi)}{\underline{q}I_{\min}}.$$

Since the sorting procedure invokes at most K_M comparators, the stopping time τ_π of the entire priority-learning phase satisfies

$$\mathbb{E}[\tau_\pi] \le C\frac{K_M\log(2K_M/\delta_\pi)}{\underline{q}I_{\min}}.$$

The expected regret incurred in the first phase is bounded by

$$\mathbb{E}[R_T^{\mathrm{pri}}] \le \mathbb{E}[\tau_\pi \wedge T] \le \min\left\{T, C\frac{K_M\log(2K_M/\delta_\pi)}{\underline{q}I_{\min}}\right\}, \tag{8}$$

because the per-slot pseudo-regret is at most one.

Uniform Reward Confidence Event. For each device i, index its observed rewards according to the number of times it has been served. Let $X_{i,1}, X_{i,2}, \dots$ be the resulting reward sequence. Even though the service times are adaptive, the noise in each observed reward remains conditionally zero mean and σ-sub-Gaussian. Let $\widehat{\theta}_{i,n} = \frac{1}{n}\sum_{s=1}^{n} X_{i,s}$. Define $L_\theta = \log \frac{4MT}{\delta_\theta}$. For every $i \in [M]$ and $n \in [T]$, conditional sub-Gaussian concentration gives

$$\Pr\left(|\widehat{\theta}_{i,n} - \theta_i| > \sigma\sqrt{2L_\theta/n}\right) \le 2e^{-L_\theta}.$$

Taking a union bound over all M devices and all $n \le T$, we obtain the event

$$\mathcal{E}_\theta = \left\{|\widehat{\theta}_{i,n} - \theta_i| \le \sigma\sqrt{2L_\theta/n}, \quad \forall i \in [M], n \in [T]\right\}$$

with $\Pr(\mathcal{E}_\theta^c) \le \delta_\theta$. On $\mathcal{E}_\theta$, the upper confidence bound satisfies $U_i(t) \ge \theta_i$ for every device and slot. Moreover, if $N_i(t-1) \ge 1$, then

$$U_i(t) - \theta_i \le 2\sigma\sqrt{2L_\theta/N_i(t-1)}.$$

If $N_i(t-1) = 0$, the algorithm sets $U_i(t) = 1$, and therefore $U_i(t) - \theta_i \le 1$.

Optimistic Instantaneous Regret Bound. Consider a slot t after the priority-learning phase, and suppose that $\mathcal{E}_\pi \cap \mathcal{E}_\theta$ holds. Since $\widehat{\pi} = \pi$, the algorithm runs the dynamic program under the correct priority order. For a fixed admission set A, write

$$w_{i,t}(A;\pi) = \Pr(I_t = i \mid \mathcal{F}_{t-1}, p_t, A_t = A).$$

The expected reward is linear in the reward vector:

$$F_t(A;\pi,\theta) = \textstyle\sum_{i=1}^{M} w_{i,t}(A;\pi)\theta_i.$$

Let $A_t^\star$ be the admission set selected by the true oracle, and let A_t be the admission set selected by the UCB dynamic program. Since

$$U_i(t) \ge \theta_i, F_t(A_t^\star;\pi,\theta) \le F_t(A_t^\star;\pi,U(t)),$$

and A_t maximizes the optimistic objective, $F_t(A_t^\star;\pi,U(t)) \le F_t(A_t;\pi,U(t))$. Combining the two inequalities yields

$$r_t = F_t(A_t^\star;\pi,\theta) - F_t(A_t;\pi,\theta) \le F_t(A_t;\pi,U(t)) - F_t(A_t;\pi,\theta) = \textstyle\sum_{i=1}^{M} w_{i,t}(A_t;\pi)\big(U_i(t) - \theta_i\big).$$

Define

$$b_i(n) = \begin{cases} 1, & n = 0, \\ 2\sigma\sqrt{2L_\theta/n}, & n \ge 1. \end{cases}$$

Then, on $\mathcal{E}_\theta$,

$$r_t \le \textstyle\sum_{i=1}^{M} w_{i,t}(A_t;\pi) b_i\big(N_i(t-1)\big).$$

Summing the Reward-Learning Regret. Let $Y_{i,t} = \mathbf{1}\{I_t = i\}$ be the service indicator. Conditional on the history before slot t, $\mathbb{E}[Y_{i,t} \mid \mathcal{F}_{t-1}] = w_{i,t}(A_t;\pi)$. The quantity $b_i\big(N_i(t-1)\big)$ and the indicator that slot t belongs to the reward-learning phase are both measurable with respect to the information available before the service outcome is observed. Hence,

$$\mathbb{E}\left[\textstyle\sum_{t>\tau_\pi} w_{i,t}(A_t;\pi) b_i\big(N_i(t-1)\big)\right] = \mathbb{E}\left[\textstyle\sum_{t>\tau_\pi} Y_{i,t} b_i\big(N_i(t-1)\big)\right].$$

The right-hand side can be bounded pathwise. The first observed reward of device i contributes at most one. Every subsequent observation contributes $2\sigma\sqrt{2L_\theta/n}$ when the device has previously been observed n times. Therefore,

$$\begin{aligned} \textstyle\sum_{t=1}^{T} Y_{i,t} b_i\big(N_i(t-1)\big) &\le 1 + 2\sigma\sqrt{2L_\theta}\textstyle\sum_{n=1}^{N_i(T)-1} \frac{1}{\sqrt{n}} \\ &\le 1 + 4\sigma\sqrt{2L_\theta N_i(T)}. \end{aligned}$$

Summing over devices gives

$$\textstyle\sum_{i=1}^{M}\sum_{t=1}^{T} Y_{i,t} b_i\big(N_i(t-1)\big) \le M + 4\sigma\sqrt{2L_\theta}\sum_{i=1}^{M}\sqrt{N_i(T)}.$$

Since at most one device is served per slot, $\sum_{i=1}^{M} N_i(T) \le T$. By the Cauchy–Schwarz inequality,

$$\sum_{i=1}^{M} \sqrt{N_i(T)} \le \sqrt{M \sum_{i=1}^{M} N_i(T)} \le \sqrt{MT}.$$

It follows that, on the joint good event,

$$\mathbb{E}\left[R_T^{\mathrm{rew}} \mathbf{1}\{\mathcal{E}_\pi \cap \mathcal{E}_\theta\}\right] \le M + 4\sigma\sqrt{2MT \log \tfrac{4MT}{\delta_\theta}}.$$

Reward observations collected during the priority-learning phase are included in $N_i(t)$. Ignoring them in the preceding counting argument could only increase the upper bound, so the result remains valid.

Completing the Proof. On the event $\mathcal{E}_\pi \cap \mathcal{E}_\theta$, the total regret is bounded by the sum of the priority-phase regret and the reward-learning regret. On the complement of this event, the cumulative pseudo-regret is at most T. Therefore,

$$\mathbb{E}[R_T] \le \mathbb{E}[\tau_\pi \wedge T] + M + 4\sigma\sqrt{2MT \log \tfrac{4MT}{\delta_\theta}} + T \Pr(\mathcal{E}_\pi^c \cup \mathcal{E}_\theta^c).$$

Using $\Pr(\mathcal{E}_\pi^c) \le \delta_\pi, \Pr(\mathcal{E}_\theta^c) \le \delta_\theta$, and the stopping-time bound from equation 8, we obtain

$$\mathbb{E}[R_T] \le C_1 \min\left\{T, \tfrac{K_M \log(2K_M/\delta_\pi)}{\underline{q} I_{\min}}\right\} + C_2\left[M + \sigma\sqrt{MT \log \tfrac{4MT}{\delta_\theta}}\right] + T(\delta_\pi + \delta_\theta),$$

for universal constants $C_1, C_2 > 0$. This proves Theorem 1. □

C Proofs for Priority-Misordering Sensitivity

Lemma 1 (Basic properties of the Bellman operator)**.** *For every device i, slot t, and $v, v' \in [0,1]$, the Bellman operator in equation 2 satisfies:*

1. *Monotonicity in the continuation value: $v \le v' \Longrightarrow \mathcal{T}_{i,t}(v) \le \mathcal{T}_{i,t}(v')$;*
2. *Value improvement: $\mathcal{T}_{i,t}(v) \ge v$;*
3. *Boundedness: $0 \le \mathcal{T}_{i,t}(v) \le 1$;*
4. *Nonexpansiveness: $|\mathcal{T}_{i,t}(v) - \mathcal{T}_{i,t}(v')| \le |v - v'|$;*
5. *Monotonicity in the reward parameter: if $\theta_i \le \theta_i'$, then $\mathcal{T}_{i,t}(v;\theta_i) \le \mathcal{T}_{i,t}(v;\theta_i')$;*
6. *Lipschitz continuity in the reward parameter: $|\mathcal{T}_{i,t}(v;\theta_i) - \mathcal{T}_{i,t}(v;\theta_i')| \le a_{i,t}|\theta_i - \theta_i'|$.*

Proof. The first three properties follow directly from $\mathcal{T}_{i,t}(v) = \max\{v, a_{i,t}\theta_i + (1-a_{i,t})v\}$ and the assumptions $a_{i,t}, \theta_i, v \in [0,1]$.

For nonexpansiveness, both functions inside the maximum are nondecreasing and Lipschitz in v, with slopes 1 and $1 - a_{i,t}$, respectively. Their pointwise maximum is therefore 1-Lipschitz.

Monotonicity in θ_i is immediate because the admission value is increasing in θ_i. Finally, the positive-part representation gives $|\mathcal{T}_{i,t}(v;\theta_i) - \mathcal{T}_{i,t}(v;\theta_i')| = a_{i,t}\,|(\theta_i - v)_+ - (\theta_i' - v)_+| \le a_{i,t}|\theta_i - \theta_i'|$, where the inequality follows from the 1-Lipschitz continuity of the positive-part function. □

C.1 Fixed-Action Effect of an Adjacent Swap

We first establish the source of the basic pairwise sensitivity term. Fix an admission set A, and suppose that devices $i, j \in A$ are adjacent under a priority order. Let

$$H = \{k \in A : k \text{ is ranked above both } i \text{ and } j\}.$$

Let L denote the devices in A ranked below both i and j. Consider two priority orders that differ only in whether i or j is ranked first. Define

$$\rho_H = \textstyle\prod_{k \in H}(1 - a_k),$$

as the probability that every higher-priority admitted device is effectively inactive.

Conditional on reaching the pair, when $i \succ j$, the expected contribution of i, j, and the lower suffix is

$$a_i\theta_i + (1-a_i)a_j\theta_j + (1-a_i)(1-a_j)W_L,$$

where W_L is the conditional expected reward of the fixed lower-priority suffix. When $j \succ i$, the corresponding value is

$$a_j\theta_j + (1-a_j)a_i\theta_i + (1-a_i)(1-a_j)W_L.$$

Subtracting the two expressions gives $a_i a_j(\theta_i - \theta_j)$. Multiplying by the probability ρ_H that the pair is reached yields

$$F_t(A; i \succ j) - F_t(A; j \succ i) = \rho_H a_i a_j (\theta_i - \theta_j).$$

Therefore,

$$|F_t(A; i \succ j) - F_t(A; j \succ i)| \le a_i a_j |\theta_i - \theta_j|.$$

If either i or j is not admitted, exchanging their positions does not change the value of the fixed admission set, and the same inequality holds trivially.

C.2 Proof of Theorem 2

Let h and ℓ satisfy $\theta_h \geq \theta_\ell$. For brevity, write $x = \theta_h, y = \theta_\ell, a = a_h, b = a_\ell$. The Bellman operators are $\mathcal{T}_h(v) = v + a(x - v)_+$ and $\mathcal{T}_\ell(v) = v + b(y - v)_+$. We compare $\mathcal{T}_h\big(\mathcal{T}_\ell(v)\big)$ and $\mathcal{T}_\ell\big(\mathcal{T}_h(v)\big)$.

Case 1: $v \geq y$. Since $x \geq y$, the lower-reward device does not improve the continuation value:

$$\mathcal{T}_\ell(v) = v.$$

Moreover, $\mathcal{T}_h(v) \geq v \geq y$, so applying $\mathcal{T}_\ell$ after $\mathcal{T}_h$ also has no effect. Hence,

$$\mathcal{T}_h\big(\mathcal{T}_\ell(v)\big) = \mathcal{T}_h(v) = \mathcal{T}_\ell\big(\mathcal{T}_h(v)\big),$$

and the difference is zero.

Case 2: $v < y$ and $\mathcal{T}_h(v) < y$. Since $v < y \leq x$, both devices improve the original continuation value. The condition $\mathcal{T}_h(v) < y$ also implies that device ℓ improves the continuation value after device h has been considered.

We have $\mathcal{T}_\ell(v) = v + b(y - v)$, and, since $\mathcal{T}_\ell(v) \leq y \leq x$,

$$\mathcal{T}_h\big(\mathcal{T}_\ell(v)\big) = \mathcal{T}_\ell(v) + a\left(x - \mathcal{T}_\ell(v)\right) = ax + (1 - a)\big[v + b(y - v)\big].$$

Similarly, $\mathcal{T}_h(v) = v + a(x - v)$, and the case condition gives

$$\mathcal{T}_\ell\big(\mathcal{T}_h(v)\big) = \mathcal{T}_h(v) + b\left(y - \mathcal{T}_h(v)\right) = by + (1 - b)\big[v + a(x - v)\big].$$

Subtracting and simplifying, $\mathcal{T}_h\big(\mathcal{T}_\ell(v)\big) - \mathcal{T}_\ell\big(\mathcal{T}_h(v)\big) = ab(x - y)$. Returning to the original notation gives $a_h a_\ell(\theta_h - \theta_\ell)$.

Case 3: $v < y$ and $\mathcal{T}_h(v) \geq y$. As before, $\mathcal{T}_\ell(v) = v + b(y - v)$, and $\mathcal{T}_h\big(\mathcal{T}_\ell(v)\big) = ax + (1 - a)\big[v + b(y - v)\big]$. However, since $\mathcal{T}_h(v) \geq y$, the lower-reward device does not improve the continuation value after h has been considered. Therefore, $\mathcal{T}_\ell\big(\mathcal{T}_h(v)\big) = \mathcal{T}_h(v) = v + a(x - v)$. Subtracting gives

$$\mathcal{T}_h\big(\mathcal{T}_\ell(v)\big) - \mathcal{T}_\ell\big(\mathcal{T}_h(v)\big) = b(1 - a)(y - v).$$

In the original notation, this is $(1 - a_h)a_\ell(\theta_\ell - v)$.

The three cases establish the following piecewise expression:

$$\mathcal{T}_h\big(\mathcal{T}_\ell(v)\big) - \mathcal{T}_\ell\big(\mathcal{T}_h(v)\big) = \begin{cases} 0, & v \geq \theta_\ell, \\ a_h a_\ell(\theta_h - \theta_\ell), & v < \theta_\ell \text{ and } \mathcal{T}_h(v) < \theta_\ell, \\ (1 - a_h)a_\ell(\theta_\ell - v), & v < \theta_\ell \text{ and } \mathcal{T}_h(v) \geq \theta_\ell. \end{cases}$$

Each expression is nonnegative, proving that placing the higher-reward device first weakly increases the local oracle value. To obtain the compact representation in Theorem 2, note that when $v < y$, $\mathcal{T}_h(v) < y$ is equivalent to $a(x - y) < (1 - a)(y - v)$. Thus, in Case 2, $ab(x - y) = b \min\left\{a(x - y), (1 - a)(y - v)\right\}$, whereas in Case 3 the minimum is $b(1 - a)(y - v)$. This proves $\mathcal{T}_h\big(\mathcal{T}_\ell(v)\big) - \mathcal{T}_\ell\big(\mathcal{T}_h(v)\big) = b\mathbf{1}\{v < y\} \min\left\{a(x - y), (1 - a)(y - v)\right\}$.

Next we prove the continuation value independent form. Assume without loss of generality that $\theta_h \geq \theta_\ell$. By Theorem 2, the local difference is either zero, $a_h a_\ell(\theta_h - \theta_\ell)$, or $(1 - a_h)a_\ell(\theta_\ell - v)$. The second expression already equals the claimed upper bound. In the third regime, $\mathcal{T}_h(v) \geq \theta_\ell$, which is equivalent to $v + a_h(\theta_h - v) \geq \theta_\ell$. Rearranging gives $(1 - a_h)(\theta_\ell - v) \leq a_h(\theta_h - \theta_\ell)$. Multiplying by a_ℓ yields $(1 - a_h)a_\ell(\theta_\ell - v) \leq a_h a_\ell(\theta_h - \theta_\ell)$. Therefore, $\kappa_{ij,t}(v) \leq a_i a_j |\theta_i - \theta_j|$.

□

C.3 Proof of Theorem 3

Let $\pi^0 = \pi, \cdots, \pi^L = \widehat{\pi}$ be a sequence of permutations such that each π^s is obtained from π^{s-1} by swapping one adjacent inverted pair. Such a sequence can be constructed by bubble sort, and each pair in $\mathrm{Inv}(\pi, \widehat{\pi})$ is swapped exactly once. Therefore, $L = \left|\mathrm{Inv}(\pi, \widehat{\pi})\right|$.

Suppose that the s-th swap exchanges devices i_s and j_s. The two permutations π^{s-1} and π^s have the same prefix above the swapped pair and the same suffix below it. Let v_s be the continuation value generated by the common suffix. By prefix nonexpansiveness and Theorem 2,

$$|V_t^{\pi^{s-1}} - V_t^{\pi^s}| \le a_{i_s,t} a_{j_s,t} |\theta_{i_s} - \theta_{j_s}|.$$

Using the triangle inequality,

$$|V_t^{\pi} - V_t^{\widehat{\pi}}| \le \sum_{s=1}^{L} |V_t^{\pi^{s-1}} - V_t^{\pi^s}| \le \sum_{s=1}^{L} a_{i_s,t} a_{j_s,t} |\theta_{i_s} - \theta_{j_s}|.$$

Since each inversion pair is swapped exactly once, the final sum is $d_t(\pi, \widehat{\pi}; \theta)$. Hence,

$$|V_t^{\pi} - V_t^{\widehat{\pi}}| \le d_t(\pi, \widehat{\pi}; \theta).$$

□

C.4 Fixed-Action Distortion under Two Permutations

We next prove that, for any fixed admission set A,

$$|F_t(A; \pi, \theta) - F_t(A; \widehat{\pi}, \theta)| \le d_t(\pi, \widehat{\pi}; \theta).$$

Use the same adjacent-swap sequence as in Appendix C.3. For one adjacent swap of devices i_s, j_s, Appendix C.1 gives

$$\left|F_t(A; \pi^{s-1}, \theta) - F_t(A; \pi^s, \theta)\right| \le a_{i_s,t} a_{j_s,t} |\theta_{i_s} - \theta_{j_s}|.$$

This inequality is also valid when one or both devices are absent from A, in which case the left-hand side is zero. Summing over the adjacent-swap path gives

$$|F_t(A; \pi, \theta) - F_t(A; \widehat{\pi}, \theta)| \le \sum_{i,j \in \mathrm{Inv}(\pi, \widehat{\pi})} a_{i,t} a_{j,t} |\theta_i - \theta_j| = d_t(\pi, \widehat{\pi}; \theta).$$

C.5 Proof of Theorem 4

Let $A^\star = A_t^\star(\pi)$ and $\widehat{A} = A_t^\star(\widehat{\pi})$. By adding and subtracting the expected rewards evaluated under $\widehat{\pi}$,

$$\begin{aligned} F_t(A^\star; \pi, \theta) - F_t(\widehat{A}; \pi, \theta) =& [F_t(A^\star; \pi, \theta) - F_t(A^\star; \widehat{\pi}, \theta)] + [F_t(A^\star; \widehat{\pi}, \theta) - F_t(\widehat{A}; \widehat{\pi}, \theta)] \\ &+ [F_t(\widehat{A}; \widehat{\pi}, \theta) - F_t(\widehat{A}; \pi, \theta)]. \end{aligned}$$

Since $\widehat{A}$ maximizes the expected reward under $\widehat{\pi}$,

$$F_t(A^\star; \widehat{\pi}, \theta) - F_t(\widehat{A}; \widehat{\pi}, \theta) \le 0.$$

The fixed-action distortion bound from Appendix C.4 gives

$$|F_t(A^\star; \pi, \theta) - F_t(A^\star; \widehat{\pi}, \theta)| \le d_t(\pi, \widehat{\pi}; \theta)$$

and

$$|F_t(\widehat{A}; \widehat{\pi}, \theta) - F_t(\widehat{A}; \pi, \theta)| \le d_t(\pi, \widehat{\pi}; \theta).$$

Combining these inequalities yields

$$F_t(A^\star; \pi, \theta) - F_t(\widehat{A}; \pi, \theta) \le 2 d_t(\pi, \widehat{\pi}; \theta).$$

□

D Implementation Details for Relevance-Aware Learning

We summarize Relevance-Aware Priority-UCB in Algorithm 2.

Algorithm 2 Relevance-Aware Priority-UCB

Require: confidence levels $\delta_\pi, \delta_\theta$; tolerance sequence ε_t; known gate probabilities γ_i.
Initialize $\mathcal{P}_1 = ([M], \emptyset)$, $L_{ij}(1) = 0, \forall i < j$, initialize all reward counts and estimates to zero.
for $t = 1, \ldots, T$ **do**
  Observe p_t and compute $a_{i,t} = p_{i,t}\gamma_i$.
  Compute $L_i(t)$, $U_i(t)$, and $G_{ij}(t) = \max\{U_i(t) - L_j(t), U_j(t) - L_i(t)\}$.
  Compute the incomparable-pair set $\mathcal{U}_t$ induced by the transitive closure of $\mathcal{P}_t$.
  Evaluate $B_t^{\mathrm{ord}} = \sum_{i,j\in\mathcal{U}_t} a_{i,t}a_{j,t}G_{ij}(t)$.
  if $B_t^{\mathrm{ord}} \le \varepsilon_t$, **then**
    Select a linear extension $\widehat{\pi}_t \in \mathcal{L}(\mathcal{P}_t)$, run optimistic DP under $\widehat{\pi}_t$ and $U(t)$ to get A_t.
  else
    Choose $\{i_t, j_t\} \in \arg\max_{i,j\in\mathcal{U}_t} S_{ij,t}$, and set $A_t = \{i_t, j_t\}$.
  Observe demand realization $D_{i,t}$, served device I_t, and its reward $X_{I_t,t}$.
  Update the empirical mean and confidence interval of device I_t.
  if slot t is a probing slot, and both probed devices have demand **then**
    Update the log-likelihood ratio of the probed pair.
    Add a directed edge when a boundary is crossed and update the transitive closure.

D.1 Confidence Parameters

For the reward estimates, one possible horizon-dependent confidence radius is

$$\beta_i(t) = \sigma\sqrt{2\log(4MT/\delta_\theta)/N_i(t-1)}, \quad N_i(t-1) \ge 1.$$

For an unobserved device, the confidence interval is set to $[0, 1]$. To control all accepted pairwise relations, allocate the priority error budget uniformly over all unordered pairs:

$$\delta_{ij} = 2\delta_\pi/M(M-1).$$

A convenient likelihood boundary is

$$b_{ij} = \log(2/\delta_{ij}) = \log(M(M-1)/\delta_\pi).$$

The corresponding pairwise likelihood-ratio test has error probability at most δ_{ij}. A union bound then implies that every accepted directed edge is correct with probability at least $1 - \delta_\pi$.

Uniform confidence allocation is not essential. Pair-specific budgets can be used when prior information or a deterministic upper bound on pair relevance is available. The theoretical analysis only requires that $\sum_{i<j} \delta_{ij} \le \delta_\pi$.

D.2 Precise Pairwise Update

For each unordered pair i, j, fix an arbitrary canonical orientation, for example $i < j$, and interpret $L_{ij} > 0$ as evidence for $i \succ j$. When pair i, j is isolated and both devices have demand, update

$$L_{ij} \leftarrow L_{ij} + \ell_{ij}(O_t),$$

where

$$\ell_{ij}(O_t) = \begin{cases} -\log(1-\gamma_j), & O_t = i, \\ \log(1-\gamma_i), & O_t = j, \\ 0, & O_t = 0. \end{cases}$$

No priority update is performed when only one device has demand, because such an outcome does not depend on their relative priority. Once a pair is implied by the transitive closure, its likelihood statistic no longer needs to be updated. Any partially accumulated statistic may be retained for diagnostic purposes but has no effect on the algorithm.

D.3 Maintaining the Partial Order

A simple implementation stores a reachability matrix $R_t \in \{0,1\}^{M\times M}$, where $R_t(i,j) = 1$ means that the current partial order implies $i \succ j$. Initially, $R_1(i,j) = 0$ for all $i \neq j$. When a new edge $i \to j$ is accepted, the transitive closure can be updated by setting $R_t(u,v) = 1$ for every u satisfying $u = i$ or $R_t(u,i) = 1$, and every v satisfying $v = j$ or $R_t(j,v) = 1$. The update requires $O(M^2)$ Boolean operations in a direct implementation. The unresolved-pair set is $\mathcal{U}_t = \{i,j : R_t(i,j) = R_t(j,i) = 0\}$.

On the joint high-probability event that all accepted comparisons are correct, no accepted edge creates a directed cycle. In an implementation, a cycle safeguard may nevertheless be included: an edge that contradicts the existing transitive closure is not inserted until additional evidence is collected. This safeguard affects only the low-probability comparison-error event and is not needed in the main analysis.

D.4 Constructing a Linear Extension

Any topological ordering of $\mathcal{P}_t$ may be used in the exploitation step. A useful deterministic rule is a UCB-prioritized topological sort. At each position, among all devices whose predecessors have already been placed, select the device with the largest reward upper confidence bound $U_i(t)$. Remaining ties may be broken by device index. This rule does not solve an optimistic optimization problem over all linear extensions; such an optimization may be combinatorial. Its purpose is only to produce one admissible extension. The certificate guarantee applies to every linear extension, so the regret analysis does not depend on the tie-breaking rule.

D.5 Horizon-Free Implementation

If the horizon T is unknown, the algorithm can use a doubling schedule. In epoch r, use nominal horizon $T_r = 2^r$ and confidence budgets proportional to $\delta_\pi/2^{r+1}$ and $\delta_\theta/2^{r+1}$. Reward statistics and the partial order need not be reset between epochs. Only the confidence radii and tolerance schedule are updated. Alternatively, time-uniform reward confidence sequences and anytime likelihood boundaries may be used. These modifications change logarithmic factors but not the main algorithmic structure.

D.6 Per-Slot Complexity

With a dense reachability matrix, the principal computational costs are $O(M^2)$ to enumerate unresolved pairs, compute the certificate, and select the probing pair; $O(M + |E_t|)$ to construct a linear extension; $O(M)$ to run the optimistic dynamic program; and $O(M^2)$ to update the transitive closure when a new edge is accepted. The memory requirement is $O(M^2)$ for the reachability matrix and pairwise likelihood statistics.

For larger device sets, bitset representations can reduce the practical cost of transitive-closure updates, and the certificate can be computed only over a dynamically maintained candidate subset of unresolved pairs.

E Proofs for Relevance-Aware Performance Guarantees

E.1 Joint Confidence Event

Let $\mathcal{E}_\pi$ be the event that every directed edge ever inserted into the partial-order graph is correct. For each unordered pair i, j, assign a comparison-error budget δ_{ij} satisfying $\sum_{i<j} \delta_{ij} \le \delta_\pi$. By the anytime likelihood-ratio guarantee of the pairwise comparator,

$$\Pr(\text{an incorrect edge for pair } i, j) \le \delta_{ij}.$$

A union bound therefore yields $\Pr(\mathcal{E}_\pi^c) \le \delta_\pi$. On $\mathcal{E}_\pi$, every edge in $\mathcal{P}_t$ agrees with the true permutation, and hence $\pi \in \mathcal{L}(\mathcal{P}_t)$ for all t.

Similarly, define $\mathcal{E}_\theta = \{\theta_i \in [L_i(t), U_i(t)], \forall i, t\}$. Using the reward confidence construction of Appendix B, $\Pr(\mathcal{E}_\theta^c) \le \delta_\theta$. Therefore,

$$\Pr(\mathcal{E}_\pi \cap \mathcal{E}_\theta) \ge 1 - \delta_\pi - \delta_\theta.$$

E.2 Proof of Theorem 5

Fix slot t and suppose that $\mathcal{E}_\pi \cap \mathcal{E}_\theta$ holds. Let $\pi_1, \pi_2 \in \mathcal{L}(\mathcal{P}_t)$. If a pair $\{i, j\}$ is comparable in $\mathcal{P}_t$, every linear extension must respect the same orientation. Hence, any inversion between π_1 and π_2 must belong to the unresolved set $\mathrm{Inv}(\pi_1, \pi_2) \subseteq \mathcal{U}_t$. Consider any reward vector x satisfying $x_i \in [L_i(t), U_i(t)]$. For an unresolved pair $\{i, j\}$, $|x_i - x_j| \le G_{ij}(t)$. The fixed-action sensitivity result of Appendix C therefore gives, for every admission set A,

$$|F_t(A; \pi_1, x) - F_t(A; \pi_2, x)| \le \sum_{i,j \in \mathrm{Inv}(\pi_1, \pi_2)} a_{i,t} a_{j,t} |x_i - x_j| \le \sum_{i,j \in \mathcal{U}_t} a_{i,t} a_{j,t} G_{ij}(t) = B_t^{\mathrm{ord}}.$$

Taking the supremum over A proves the first claim.

On $\mathcal{E}_\pi$, $\pi \in \mathcal{L}(\mathcal{P}_t)$, and on $\mathcal{E}_\theta$, $\theta_i \in [L_i(t), U_i(t)]$. Setting $\pi_1 = \pi$, $\pi_2 = \hat{\pi}_t$, and $x = \theta$ proves the second claim. □

E.3 Exploitation-Slot Regret Decomposition

Fix an exploitation slot t and let $\hat{\pi} = \hat{\pi}_t$. Let

$$A^\star \in \arg\max_A F_t(A; \pi, \theta)$$

and

$$A_t \in \arg\max_A F_t(A; \widehat{\pi}, U(t)).$$

On the joint good event,

$$r_t = F_t(A^\star; \pi, \theta) - F_t(A_t; \pi, \theta) = [F_t(A^\star; \pi, \theta) - F_t(A^\star; \widehat{\pi}, \theta)] + [F_t(A^\star; \widehat{\pi}, \theta) - F_t(A_t; \pi, \theta)].$$

By Theorem 5,

$$F_t(A^\star; \pi, \theta) - F_t(A^\star; \widehat{\pi}, \theta) \le B_t^{\mathrm{ord}}.$$

Since $U_i(t) \ge \theta_i$,

$$F_t(A^\star; \widehat{\pi}, \theta) \le F_t(A^\star; \widehat{\pi}, U(t)).$$

By optimality of A_t under the optimistic model,

$$F_t(A^\star; \widehat{\pi}, U(t)) \le F_t(A_t; \widehat{\pi}, U(t)).$$

Hence,

$$\begin{aligned} r_t &\le B_t^{\mathrm{ord}} + F_t(A_t; \widehat{\pi}, U(t)) - F_t(A_t; \pi, \theta) \\ &= B_t^{\mathrm{ord}} + [F_t(A_t; \widehat{\pi}, U(t)) - F_t(A_t; \pi, U(t))] + [F_t(A_t; \pi, U(t)) - F_t(A_t; \pi, \theta)]. \end{aligned}$$

Because every $U_i(t)$ belongs to the interval $[L_i(t), U_i(t)]$, Theorem 5 also applies to the optimistic reward vector $U(t)$. Therefore,

$$F_t(A_t; \widehat{\pi}, U(t)) - F_t(A_t; \pi, U(t)) \le B_t^{\mathrm{ord}}.$$

Finally, expected reward is linear in the reward vector under the true priority order:

$$F_t(A_t;\pi,U(t)) - F_t(A_t;\pi,\theta) = \textstyle\sum_i w_{i,t}(A_t;\pi)\big(U_i(t)-\theta_i\big).$$

Combining the inequalities gives

$$r_t \le 2B_t^{\mathrm{ord}} + \textstyle\sum_i w_{i,t}(A_t;\pi)\big(U_i(t)-\theta_i\big).$$

Since exploitation occurs only when $B_t^{\mathrm{ord}} \le \varepsilon_t$, we obtain

$$r_t \le 2\varepsilon_t + \textstyle\sum_i w_{i,t}(A_t;\pi)\big(U_i(t)-\theta_i\big).$$

E.4 Cumulative Reward-Optimism Term

Let $Y_{i,t} = \mathbf{1}\{I_t = i\}$. Conditional on the pre-service history, $\mathbb{E}[Y_{i,t} \mid \mathcal{F}_{t-1}] = w_{i,t}(A_t;\pi)$. Define

$$b_i(n) = \begin{cases} 1, & n = 0, \\ 2\sigma\sqrt{2\log(4MT/\delta_\theta)/n}, & n \ge 1. \end{cases}$$

On $\mathcal{E}_\theta$, $U_i(t) - \theta_i \le b_i(N_i(t-1))$. Therefore,

$$\mathbb{E}\left[\textstyle\sum_{t\notin\mathcal{Q}_T}\sum_i w_{i,t}(A_t;\pi)\big(U_i(t)-\theta_i\big)\right] \le \mathbb{E}\left[\textstyle\sum_{t=1}^T\sum_i Y_{i,t}b_i(N_i(t-1))\right].$$

For each device,

$$\textstyle\sum_{t=1}^T Y_{i,t}b_i(N_i(t-1)) \le 1 + 4\sigma\sqrt{2N_i(T)\log\frac{4MT}{\delta_\theta}}.$$

Summing over devices and using

$$\textstyle\sum_i N_i(T) \le T$$

and Cauchy–Schwarz gives

$$\mathbb{E}\left[\textstyle\sum_{t\notin\mathcal{Q}_T}\sum_i w_{i,t}(A_t;\pi)\big(U_i(t)-\theta_i\big)\right] \le M + 4\sigma\sqrt{2MT\log(4MT/\delta_\theta)}.$$

Notice that rewards collected during priority-probing slots are included in $N_i(t)$. Ignoring their beneficial effect can only enlarge the upper bound.

E.5 Proof of Theorem 6

On the joint good event, every probing slot contributes at most one unit of regret, and every exploitation slot satisfies the bound of Appendix E.3. Hence,

$$R_T \le N_T^{\mathrm{probe}} + 2\textstyle\sum_{t=1}^T \varepsilon_t + \sum_{t\notin\mathcal{Q}_T}\sum_i w_{i,t}(A_t;\pi)\big(U_i(t)-\theta_i\big).$$

Taking expectations and applying Appendix E.4 yields

$$\mathbb{E}[R_T\mathbf{1}\{\mathcal{E}_\pi\cap\mathcal{E}_\theta\}] \le \mathbb{E}[N_T^{\mathrm{probe}}] + 2\textstyle\sum_{t=1}^T\varepsilon_t + C\left[M + \sigma\sqrt{MT\log(4MT/\delta_\theta)}\right].$$

On the complement of the joint good event, cumulative regret is at most T. Since

$$\Pr(\mathcal{E}_\pi^c \cup \mathcal{E}_\theta^c) \le \delta_\pi + \delta_\theta,$$

we obtain

$$\mathbb{E}[R_T] \le \mathbb{E}[N_T^{\mathrm{probe}}] + 2\textstyle\sum_{t=1}^T\varepsilon_t + C\left[M + \sigma\sqrt{MT\log\frac{4MT}{\delta_\theta}}\right] + T(\delta_\pi+\delta_\theta).$$

This proves Theorem 6. □

E.6 Probe Complexity of Relevance-Aware Priority Learning

We now prove Theorem 7. The key observation is that Algorithm 2 cannot select a pair whose relevance–information score is uniformly small once the total residual value contributed by such pairs is also below the certificate tolerance. For each pair e, recall $s_e(t) = c_e(t)J_e(t)$. Let $D \subseteq \mathcal{E}$ be any screenable set, i.e.,

$$\sum_{e\in D} \bar{c}_e + \Big(\max_{e\in D} \bar{s}_e\Big) \sum_{e\notin D} \tfrac{1}{\underline{J}_e} \le \varepsilon_T. \tag{9}$$

Since ε_t is nonincreasing, for all $t \le T$, $\varepsilon_t \ge \varepsilon_T$.

Proof. Work on the good event $\mathcal{G}$. We first show that no pair in D can be selected for probing at any time $t \ge t_0$. Suppose, for contradiction, that Algorithm 2 probes at time $t \ge t_0$ and selects some pair $e_t \in D$. Since Algorithm 2 chooses the unresolved pair with the largest relevance–information score,

$$s_e(t) \le s_{e_t}(t) \qquad \forall e \in \mathcal{U}_t. \tag{10}$$

Decompose the current order certificate as

$$B_t^{\mathrm{ord}} = \sum_{e\in\mathcal{U}_t\cap D} c_e(t) + \sum_{e\in\mathcal{U}_t\setminus D} c_e(t). \tag{11}$$

For the first term, the good-event envelope gives

$$\sum_{e\in\mathcal{U}_t\cap D} c_e(t) \le \sum_{e\in D} \bar{c}_e. \tag{12}$$

For every unresolved pair $e \notin D$,

$$c_e(t) = \tfrac{s_e(t)}{J_e(t)} \le \tfrac{s_{e_t}(t)}{\underline{J}_e}, \tag{13}$$

where we used equation 10 and $J_e(t) \ge \underline{J}_e$. Hence

$$\sum_{e\in\mathcal{U}_t\setminus D} c_e(t) \le s_{e_t}(t) \sum_{e\notin D} \tfrac{1}{\underline{J}_e} \le \Big(\max_{e\in D} \bar{s}_e\Big) \sum_{e\notin D} \tfrac{1}{\underline{J}_e}. \tag{14}$$

Combining equation 11–14 and the screenability condition equation 9,

$$B_t^{\mathrm{ord}} \le \varepsilon_T \le \varepsilon_t.$$

This contradicts the probing rule of Algorithm 2, which probes only when $B_t^{\mathrm{ord}} > \varepsilon_t$. Therefore,

$$e_t \notin D \qquad \text{for every probing round } t \ge t_0.$$

It follows that, on $\mathcal{G}$,

$$N_T^{\mathrm{probe}} \le t_0 + \sum_{e\notin D} N_e(T),$$

where $N_e(T)$ is the number of slots in which pair e is explicitly probed.

Let τ_e denote the pair-specific stopping time of the sequential comparator. Interleaving comparisons of different pairs does not change the pair-specific stopping argument, since the comparator is updated only on the subsequence of slots assigned to e. By Theorem 9,

$$\mathbb{E}[\tau_e] \le L_e(\delta) = O\Big(\tfrac{\log(|\mathcal{E}|T/\delta)}{\underline{J}_e}\Big).$$

Moreover, $N_e(T) \le \tau_e$, even if the horizon ends before the pair is resolved. Thus

$$\mathbb{E}[N_T^{\mathrm{probe}} \mathbf{1}_{\mathcal{G}}] \le t_0 + \sum_{e\notin D} L_e(\delta).$$

On $\mathcal{G}^c$, trivially, $N_T^{\mathrm{probe}} \le T$. Since $\Pr(\mathcal{G}^c) \le \delta$,

$$\mathbb{E}[N_T^{\mathrm{probe}}] \le t_0 + \sum_{e\notin D} L_e(\delta) + \delta T.$$

Since this holds for every screenable D, minimizing over D proves

$$\mathbb{E}[N_T^{probe}] \le t_0 + \mathcal{C}_{\mathrm{RA}}(T, t_0, \delta) + \delta T$$

□

E.7 Favorable Regimes and Corresponding Learning Complexity

In this section, we use

$$q_{ij,t} = \Pr(D_{i,t} = D_{j,t} = 1 \mid \mathcal{F}_{t-1})$$

to denote the joint activity probability of devices i and j. The adjacent-inversion sensitivity analysis in Section 5 implies the conservative bound

$$c_{ij}(t) \le q_{ij,t}\gamma_i\gamma_j G_{ij}(t), \tag{15}$$

where the probability that the resource reaches the pair has been upper bounded by one, and

$$G_{ij}(t) = \max\{U_i(t) - L_j(t),\, U_j(t) - L_i(t)\}.$$

On the reward-confidence event,

$$G_{ij}(t) \le |\theta_i - \theta_j| + 2r_i(t) + 2r_j(t), \tag{16}$$

where $r_i(t)$ is the confidence radius of device i's reward.

Conditional on both devices being active, the outcome distributions under the two possible priority orders are

$$P_{ij}^{i\succ j} = \big(\gamma_i,\, (1-\gamma_i)\gamma_j,\, (1-\gamma_i)(1-\gamma_j)\big)$$

and

$$P_{ij}^{j\succ i} = \big((1-\gamma_j)\gamma_i,\, \gamma_j,\, (1-\gamma_i)(1-\gamma_j)\big),$$

corresponding to service of i, service of j, and no service, respectively. Define

$$I_{ij} = \min\left\{D_{\mathrm{KL}}(P_{ij}^{i\succ j}\|P_{ij}^{j\succ i}), D_{\mathrm{KL}}(P_{ij}^{j\succ i}\|P_{ij}^{i\succ j})\right\}.$$

The information acquired per physical probing slot is therefore

$$J_{ij}(t) = q_{ij,t}I_{ij}$$

E.7.1 Near-homogeneous rewards

Corollary 2 (Near-homogeneous rewards)**.** *Let $D_\Delta \subseteq \mathcal{E}$ be a collection of pairs satisfying*

$$|\theta_i - \theta_j| \le \Delta, \qquad (i,j) \in D_\Delta.$$

Suppose that, after t_0,

$$r_i(t) \le r_0, \qquad q_{ij,t} \le \bar{q}, \qquad \gamma_i \le \bar{\gamma}, \qquad J_{ij}(t) \le \bar{J}.$$

Then every $e \in D_\Delta$ satisfies

$$\bar{c}_e \le A_\Delta(\Delta + 4r_0), \qquad \bar{s}_e \le A_\Delta\bar{J}(\Delta + 4r_0), \tag{17}$$

where $A_\Delta = \bar{q}\bar{\gamma}^2$. Consequently, if

$$A_\Delta(\Delta + 4r_0)\left(|D_\Delta| + \bar{J}\textstyle\sum_{e\notin D_\Delta}\frac{1}{\bar{J}_e}\right) \le \varepsilon_T, \tag{18}$$

then D_Δ is screenable and

$$\mathbb{E}[N_T^{\mathrm{probe}}] \le t_0 + \textstyle\sum_{e\notin D_\Delta} L_e(\delta) + \delta T. \tag{19}$$

Proof. By equation 16, $G_{ij}(t) \le \Delta + 4r_0$ for every $(i,j) \in D_\Delta$. Applying equation 15,

$$c_{ij}(t) \le \bar{q}\bar{\gamma}^2(\Delta + 4r_0) = A_\Delta(\Delta + 4r_0).$$

Multiplying by $J_{ij}(t) \le \bar{J}$ yields the score bound in equation 17. Therefore

$$\Phi_T(D_\Delta) \le A_\Delta(\Delta + 4r_0)\left(|D_\Delta| + \bar{J}\textstyle\sum_{e\notin D_\Delta}\frac{1}{\bar{J}_e}\right).$$

Under equation 18, D_Δ is screenable, and Theorem 7 gives equation 19. □

In the fully homogeneous case $D_\Delta = \mathcal{E}$, the second term in equation 18 vanishes. Thus, once

$$|\mathcal{E}|A_\Delta(\Delta + 4r_0) \le \varepsilon_T,$$

no further priority probing is required. Notice that this effect becomes visible only after the reward confidence radii have contracted. This explains why the empirical gain in the near-homogeneous regime is more moderate than in the low-gate and rare-overlap regimes, where the relevance attenuation is known directly from the system parameters.

E.7.2 LOW GATE PROBABILITY

Setting. We partition the devices into a core set H and a low-impact set L, with $|H| = |L| = M/2$. Core devices satisfy

$$\gamma_i \in [\gamma_{\min}, \gamma_{\max}] \subset (0,1), \qquad i \in H,$$

whereas every low-impact device has

$$\gamma_i = g, \qquad i \in L.$$

This is the setting used in the low-gate experiments, where g is varied across [0.05, 0.6].

Corollary 3 (Low-gate devices)**.** *Assume that joint activity probabilities are bounded above and below by positive constants and that $G_{ij}(t) \le G_{\max}$. As $g \to 0$,*

pair type	c_{ij}	J_{ij}	$s_{ij} = c_{ij} J_{ij}$
$H-L$	$O(g)$	$\Theta(g)$	$O(g^2)$
$L-L$	$O(g^2)$	$\Theta(g^3)$	$O(g^5)$.

(20)

Let D_g contain all $H-L$ and $L-L$ pairs, and let

$$K_{HL} = |H||L|, \qquad K_{LL} = \tbinom{|L|}{2}.$$

There exist constants $A_1, A_2, B_1, B_2 > 0$, independent of g, such that D_g is screenable whenever

$$K_{HL} A_1 g + K_{LL} A_2 g^2 + \max\{B_1 g^2, B_2 g^5\} \textstyle\sum_{e \in H \times H} \frac{1}{J_e} \le \varepsilon_T. \tag{21}$$

Hence

$$\mathbb{E}[N_T^{\text{probe}}] \le t_0 + \textstyle\sum_{e \in H \times H} L_e(\delta) + \delta T \tag{22}$$

whenever equation 21 holds.

Proof. For a core–low pair,

$$c_{ij}(t) \le q_{ij,t} \gamma_i g G_{ij}(t) = O(g).$$

For two low-gate devices,

$$c_{ij}(t) \le q_{ij,t} g^2 G_{ij}(t) = O(g^2). \tag{23}$$

Let the gate probability of core devices be $c \in [\gamma_{\min}, \gamma_{\max}]$. Then we have

$$P^{c \succ g} = (c, (1-c)g, (1-c)(1-g))$$

and

$$P^{g \succ c} = ((1-g)c, g, (1-c)(1-g)).$$

Therefore,

$$D_{\mathrm{KL}}(P^{c \succ g} \| P^{g \succ c}) = -c \log(1-g) + (1-c) g \log(1-c) = g\Big[c + (1-c)\log(1-c)\Big] + O(g^2),$$

and

$$D_{\mathrm{KL}}(P^{g \succ c} \| P^{c \succ g}) = (1-g) c \log(1-g) - g \log(1-c) = g\Big[- c - \log(1-c)\Big] + O(g^2).$$

Since c lies in $(0,1)$, the coefficients of the first order terms are positive, then

$$I_{HL} = \Theta(g).$$

When $c = g$,

$$P^{i \succ j} = (g, (1-g)g, (1-g)^2), \qquad P^{j \succ i} = ((1-g)g, g, (1-g)^2),$$

Therefore,

$$I_{LL} = g \log \tfrac{1}{1-g} + (1-g) g \log(1-g) = -g^2 \log(1-g) = g^3 + O(g^4). \tag{24}$$

This gives $I_{LL} = \Theta(g^3)$. Joint activity probability bounded away from zero gives

$$J_{HL} = \Theta(g), \qquad J_{LL} = \Theta(g^3).$$

Using equation 23, we get

$$s_{HL} = O(g^2), \qquad s_{LL} = O(g^5).$$

Finally, using the screenability condition in Theorem 7 proves the corollary. □

The contrast with exact identification is particularly sharp. Resolving an $H - L$ pair requires $\Theta\Big(\frac{\log(1/\delta)}{g}\Big)$ probing slots, whereas resolving an $L - L$ pair requires $\Theta\Big(\frac{\log(1/\delta)}{g^3}\Big)$. Thus, as g decreases, these relations become increasingly expensive to identify exactly while simultaneously becoming less relevant to the admission value. Relevance-aware learning exploits precisely this divergence between identification difficulty and decision importance.

E.7.3 RARE CONTEXTUAL OVERLAP

Setting. Devices are partitioned into four groups. Within-group devices have a joint activity probability bounded away from zero, while a cross-group pair is jointly active with probability $q_{ij} = \rho$.

Corollary 4 (Rare cross-group activity)**.** *Assume that gate probabilities are bounded away from* 0 *and* 1*, and that reward contrasts are uniformly bounded. For a cross-group pair,*

$$c_{ij} = O(\rho), \qquad J_{ij} = \Theta(\rho), \qquad s_{ij} = O(\rho^2).$$

Let D_ρ denote the set of all cross-group pairs and let $K_{\text{cross}} = |D_\rho|$. There exist constants $A, B > 0$, independent of ρ, such that D_ρ is screenable whenever

$$K_{\text{cross}} A\rho + B\rho^2 \textstyle\sum_{e \in \mathcal{E}_{\text{within}}} \frac{1}{J_e} \le \varepsilon_T. \tag{25}$$

Under this condition,

$$\mathbb{E}[N_T^{\text{probe}}] \le t_0 + \textstyle\sum_{e \in \mathcal{E}_{\text{within}}} L_e(\delta) + \delta T. \tag{26}$$

Proof. For a cross-group pair, equation 15 gives

$$c_{ij}(t) \le \rho\, \gamma_i \gamma_j G_{ij}(t) = O(\rho).$$

Since the gate parameters remain bounded away from their degenerate values, $I_{ij} = \Theta(1)$, and therefore

$$J_{ij} = \rho I_{ij} = \Theta(\rho).$$

Hence

$$s_{ij} = c_{ij} J_{ij} = O(\rho^2).$$

Summing the relevance envelope over D_ρ and applying Theorem 7 completes the proof. □

Again, exact identification behaves in the opposite direction: distinguishing a cross-group relation requires $\Theta\left(\frac{\log(1/\delta)}{\rho I_{ij}}\right)$ probing slots. Hence rare contextual overlap simultaneously makes a cross-group relation harder to identify and less useful to identify.

E.7.4 SUMMARY OF FAVORABLE-REGIME SCALINGS

The three corollaries above reveal a common mechanism behind relevance-aware priority learning. Table 1 summarizes how the pairwise decision relevance c_{ij}, the per-slot priority information J_{ij}, and their product $s_{ij} = c_{ij} J_{ij}$ scale with the structural parameter of each favorable regime. For comparison, the last column reports the number of probing slots required for exact pair identification, up to constant and logarithmic factors. Throughout the table, all parameters other than the displayed regime parameter are assumed to remain in fixed nondegenerate ranges, and reward-confidence inflation is omitted for clarity.

The key observation is that decision relevance and exact-identification difficulty may move in opposite directions. For nearly homogeneous rewards, decreasing the reward contrast Δ reduces the consequence of a priority error without making the pair statistically easier to identify. The separation is even sharper for low-gate and rare-overlap pairs: as g or ρ decreases, the priority relation becomes less consequential for the admission value, while the information available for resolving it also vanishes. Exact-priority learning therefore spends increasingly many probes on relations whose downstream value is simultaneously decreasing.

In contrast, the relevance-aware score $s_{ij} = c_{ij} J_{ij}$ vanishes rapidly in these regimes. In particular, it scales as $O(g^5)$ for two low-gate devices and as $O(\rho^2)$ for rarely co-active cross-group pairs. These scalings explain why such pairs satisfy the screening condition in Theorem 7 and can remain unresolved, while priority relations among high-impact devices continue to be learned. This theoretical separation is also reflected in the resolved-pair and regret experiments reported in Section F.2.1.

E.8 DECISION-REGRET LOWER BOUND

In this section, we analyze the lower bound on decision regret arising from priority uncertainty. To simplify the analysis and highlight the primary challenges, we assume that the device reward parameters are known; consequently, the algorithm needs only to learn the priorities to make correct decisions. In this setting, we construct two priority instances whose optimal admission decisions differ. Distinguishing the two instances is then necessary for achieving subpolynomial regret.

The two device hard instance. Consider the stationary two-device subsystem i, j. Recall that $a_i = p_i\gamma_i, a_j = p_j\gamma_j$, and the rewards are Bernoulli with means θ_i, θ_j. We assume

$$\theta_i > \theta_j > 0, \qquad \theta_j < a_i\theta_i.$$

Define two environments

$$\nu^+ : \quad i \succ j,$$
$$\nu^- : \quad j \succ i.$$

These two environments have the same demand, gate and reward parameters. The only difference is the priority order.

When admitting only device i,

$$V(\{i\}) = a_i\theta_i.$$

When admitting only device j,

$$V(\{j\}) = a_j\theta_j.$$

In environment ν^+, admitting both devices gives

$$V_+(\{i, j\}) = a_i\theta_i + (1 - a_i)a_j\theta_j.$$

Therefore,

$$V_+(\{i, j\}) - V(\{i\}) = (1 - a_i)a_j\theta_j > 0,$$

and

$$V_+(\{i, j\}) - V(\{j\}) = a_i(\theta_i - a_j\theta_j) > 0.$$

Then the optimal admission set is

$$S^*_+ = \{i, j\}.$$

While in environment ν^-,

$$V_-(\{i, j\}) = a_j\theta_j + (1 - a_j)a_i\theta_i.$$

Therefore,

$$V(\{i\}) - V_-(\{i, j\}) = a_j(a_i\theta_i - \theta_j) =: \Delta_- > 0.$$

Moreover, since $\theta_j < a_i\theta_i$, the optimal admission set is

$$S^*_- = \{i\}.$$

Thus the two instances induce different optimal actions:

$$\nu^+ : \text{ admit both}, \qquad \nu^- : \text{ admit only } i.$$

Next, we will use the above hard instances to prove a lower bound on the decision regret for any uniformly efficient algorithm. When proving the theorem, we will use the property of a "uniformly efficient" algorithm.

Definition 1 (uniformly efficient algorithm)**.** *With known reward parameters, an algorithm only needs to learn the priority parameters to make correct decision. An algorithm is called "uniformly efficient" if for any $\alpha > 0$ and any environment ν, $\mathbb{E}_\nu[R_T] = o(T^\alpha)$.*

We restrict attention to uniformly efficient policies. This is a standard regularity class excluding policies that incur polynomial regret on some fixed instance. Achieving such efficiency requires the probability of persistent priority misidentification to vanish with the horizon, and therefore generally requires logarithmically growing statistical evidence. We have the following decision-regret lower bound for these algorithms.

Theorem 10 (Decision-regret lower bound. Formal version of Theorem 8)**.** *Consider the instances ν^+ and ν^- defined above. Let*

$$I_{ij}^{-+} = D_{\mathrm{KL}}\Big(P_{ij}^{j\succ i} \| P_{ij}^{i\succ j}\Big)$$

be the conditional KL divergence between the two service-outcome laws when both devices are active, and define the per-round information rate

$$J_{ij}^{-+} = p_i p_j I_{ij}^{-+}. \tag{27}$$

Then every uniformly efficient policy satisfies

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu^-}[R_T]}{\log T} \geq \frac{\Delta_-}{J_{ij}^{-+}}, \tag{28}$$

where

$$\Delta_- = a_j(a_i\theta_i - \theta_j).$$

This theorem states that a pair whose order changes the optimal admission decision imposes an unavoidable regret cost equal, up to logarithmic factors, to its decision consequence divided by the information available for distinguishing the two orders.

Proof. Step 1: only admitting both devices reveals the priority information. Let $N_{ij}(T)$ denote the number of rounds up to T in which the policy admits both i and j.

The two instances ν^+ and ν^- differ only in the relative priority of i and j. Hence, if an action does not contain both devices, its observation distribution is identical under the two instances. When both are admitted, the two observation laws differ only on the event that both devices are active, which occurs with probability $p_i p_j$. Conditional on this event, their KL divergence is I_{ij}^{-+}.

For a fixed learning policy $\mathcal{A}$, let $\mathbb{P}_{\nu,\mathcal{A}}^T$ denote the probability law of the complete interaction history up to time T under instance ν. We suppress $\mathcal{A}$ when there is no ambiguity. By the chain rule for KL divergence under adaptive sampling,

$$D_{\mathrm{KL}}\big(\mathbb{P}_{\nu^-}^T \| \mathbb{P}_{\nu^+}^T\big) = J_{ij}^{-+}\mathbb{E}_{\nu^-}[N_{ij}(T)]. \tag{29}$$

Step 2: a uniformly efficient algorithm behaves differently under the two instances. Define the event

$$A_T = \left\{N_{ij}(T) < \tfrac{T}{2}\right\}.$$

In environment ν^-, admitting both is suboptimal, and leads to the instantaneous regret of at least Δ_- Therefore,

$$\Delta_-\mathbb{E}_{\nu^-}[N_{ij}(T)] \leq \mathbb{E}_{\nu^-}[R_T]. \tag{30}$$

Uniform efficiency implies, for any $\alpha > 0$,

$$\mathbb{E}_{\nu^-}[N_{ij}(T)] = o(T^\alpha).$$

Therefore, using Markov inequality gives

$$\mathbb{P}_{\nu^-}(A_T^c) = \mathbb{P}_{\nu^-}\left(N_{ij}(T) \geq T/2\right) = o(T^{\alpha-1}), \tag{31}$$

which means

$$\mathbb{P}_{\nu^-}(A_T) \to 1.$$

On the other hand, in environment ν^+, $\{i,j\}$ is the only optimal decision. Let

$$\Delta_+ = V_+(\{i,j\}) - \max_{S\neq\{i,j\}} V_+(S) > 0.$$

then

$$\Delta_+\mathbb{E}_{\nu^+}[T - N_{ij}(T)] \leq \mathbb{E}_{\nu^+}[R_T].$$

Uniform efficiency implies

$$\mathbb{E}_{\nu^+}[T - N_{ij}(T)] = o(T^\alpha),$$

as a result,

$$\mathbb{P}_{\nu^+}(A_T) \leq \mathbb{P}_{\nu^+}\left(T - N_{ij}(T) > T/2\right) = o(T^{\alpha-1}). \tag{32}$$

This means that event A_T has asymptotically separated probabilities in these two environments:

$$P_-(A_T) \to 1, \qquad P_+(A_T) \to 0.$$

Step 3: information inequality. Using data-processing inequality for KL divergence, we can get

$$D_{\mathrm{KL}}(\mathbb{P}^T_{\nu^-} \| \mathbb{P}^T_{\nu^+}) \geq \mathrm{kl}\left(P_-(A_T), P_+(A_T)\right), \tag{33}$$

where $\mathrm{kl}(\cdot,\cdot)$ is binary relative entropy. Using equation 31–32, for any fixed $\alpha \in (0,1)$,

$$\liminf_{T\to\infty} \frac{\mathrm{kl}(P_-(A_T), P_+(A_T))}{\log T} \geq 1-\alpha. \tag{34}$$

Combining equation 29,

$$\liminf_{T\to\infty} \frac{J_{ij}^{-+}\mathbb{E}_{\nu^-}[N_{ij}(T)]}{\log T} \geq 1-\alpha.$$

Let $\alpha \downarrow 0$, we get

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu^-}[N_{ij}(T)]}{\log T} \geq \frac{1}{J_{ij}^{-+}}. \tag{35}$$

Finally, using equation 30 gives

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu^-}[R_T]}{\log T} \geq \frac{\Delta_-}{J_{ij}^{-+}}.$$

□

Under the two priority orders, the value difference of admitting both devices is

$$C_{ij}^{\mathrm{swap}} = V_+(\{i,j\}) - V_-(\{i,j\}) = a_i a_j(\theta_i - \theta_j).$$

This is exactly the local priority inversion sensitivity. We can construct a family of reward parameters as

$$\theta_j = \kappa a_i \theta_i, \quad 0 < \kappa < 1.$$

Then

$$\Delta_- = a_i a_j \theta_i (1-\kappa),$$

and

$$C_{ij}^{\mathrm{swap}} = a_i a_j \theta_i (1-\kappa a_i).$$

Therefore,

$$\Delta_- = \frac{1-\kappa}{1-\kappa a_i} C_{ij}^{\mathrm{swap}}$$

If a_i and κ are bounded away from degenerate values, this ratio is a positive constant. Then we can get the following corollary.

Corollary 5 (Relevance–information lower bound)**.** *For the above family of instances,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R_T]}{\log T} = \Omega\left(\frac{C_{ij}^{\mathrm{swap}}}{p_i p_j I_{ij}^{-+}}\right).$$

Thus the same two primitives that govern relevance-aware exploration—decision sensitivity and statistical information—also appear in the intrinsic difficulty of the admission problem.

F Detailed Experiment Settings and Results

We evaluate all algorithms in a stochastic priority-gated admission environment with a unit-capacity service resource. There are M devices. At each slot t, the learner observes a demand-probability vector $p_t = (p_{1,t}, \ldots, p_{M,t})$ and chooses an admission set $A_t \subseteq [M]$. There is no admission cost and no explicit cardinality constraint, except for heuristic baselines that impose one by design.

For each device i, demand is realized as

$$D_{i,t} \sim \text{Bernoulli}(p_{i,t}).$$

If device $i \in A_t$ has demand, an independent priority gate is realized as

$$G_{i,t} \sim \text{Bernoulli}(\gamma_i),$$

where γ_i is the known gate-success probability. Among admitted devices with $D_{i,t}G_{i,t} = 1$, the unit service is allocated to the highest-priority device under the latent priority permutation π. If no admitted device is effectively active, no service is provided. When device i is served, the reward has mean θ_i. Unless otherwise stated, rewards are Bernoulli with mean θ_i.

The learner observes p_t before making the admission decision. After the decision, it observes demands of admitted devices, the served device if any, and the realized reward if service occurs. Latent gate outcomes and unadmitted-device demands are not revealed. All algorithms are evaluated by pseudo-regret relative to the single-slot dynamic-programming oracle that knows π, θ, γ, and p_t:

$$R_T = \sum_{t=1}^{T} \left[V_t^\star - F_t(A_t; \pi, \theta, \gamma, p_t) \right],$$

where $V_t^\star$ is the optimal expected reward at slot t, and $F_t(A_t; \cdot)$ is the expected reward of the algorithm's chosen admission set. The oracle is computed in $O(M)$ by processing devices from low to high priority and comparing each device reward θ_i with the continuation value of lower-priority devices.

To reduce Monte Carlo noise in algorithm comparisons, we use common random numbers. For each fixed parameter instance and stochastic repetition, the demand, gate, and potential reward sequences are pre-generated and shared across all algorithms. Random seeds are derived deterministically from a global seed, the condition or reward-family index, the instance index, and the repetition index.

F.1 Overall Comparison Experiments

The overall comparison experiments are designed to compare the proposed algorithms and baselines under ordinary controlled stochastic-priority settings. The default configuration uses $M = 16, T = 10{,}000$. For each reward family, we generate 20 independent parameter instances and 10 stochastic repetitions per instance. Each reported trajectory is recorded at 101 evenly spaced checkpoints.

We consider two reward families. In the well-separated family, the true reward means are generated from evenly spaced values between 0.15 and 0.85 and then randomly assigned to devices. In the clustered family, rewards are generated around cluster centers $0.30, 0.45, 0.60, 0.75$, with small uniform perturbations clipped to $[0.05, 0.95]$. The priority permutation is sampled uniformly at random for every instance. The default demand model is static: each device has a fixed demand probability sampled independently from $\text{Unif}(0.3, 0.7)$.

The default algorithm set contains Known-Priority UCB-DP, Naive-Importance UCB, Exact-Priority Two-Phase, Relevance-Aware Priority-UCB, and Random Explore-Then-Commit with exploration budgets $B = 300$ and $B = 1000$. The Known-Priority algorithm knows the true priority order and only learns rewards, serving as a reward-learning reference. Exact-Priority Two-Phase first learns the full priority order using sequential pairwise likelihood tests and then runs UCB-DP. Relevance-Aware Priority-UCB probes priority pairs only when the current priority uncertainty certificate exceeds a tolerance. Random ETC probes random priority pairs for a fixed budget and then commits to a linear extension of the learned partial order. The Naive-Importance baseline does not learn priority; it scores devices by reward UCB multiplied by importance coefficients linearly ranging from 0.3 to 3.0, and admits a fixed number of devices, default 4. The importance coefficients are calculated using the true priority order.

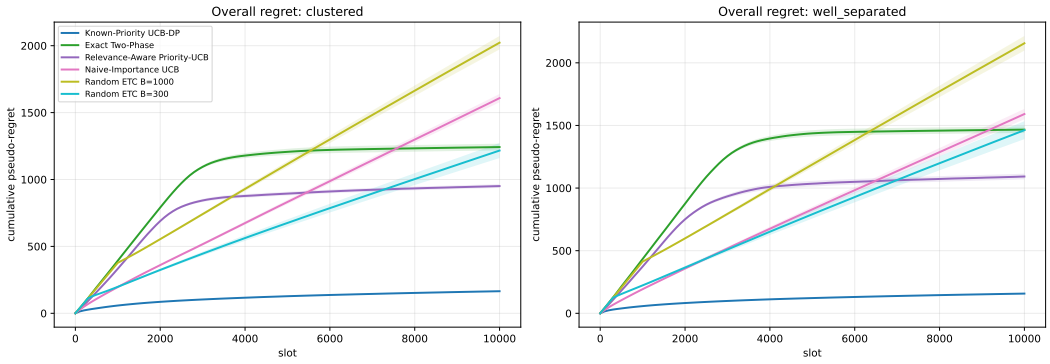


Figure 2: Evolution of cumulative pseudo-regret under clustered and well-separated reward configurations.

Unless otherwise stated, the reward-UCB confidence parameter is $\delta_\theta = 0.05$, and the priority-test confidence parameter is also $\delta_{\text{priority}} = 0.05$. For Relevance-Aware Priority-UCB, the default tolerance schedule is

$$\varepsilon_t = c\sqrt{\frac{M \log T}{t+1}}, \qquad c = 2.0,$$

and probe pairs are selected by the product of decision relevance and statistical information rate.

F.1.1 Additional Results for the Overall Comparison

We first examine the evolution of cumulative pseudo-regret over the entire horizon, rather than only its terminal value. We then inspect a representative run of the Relevance-Aware algorithm to illustrate how its order-sensitivity certificate translates into reduced priority exploration.

Regret Evolution over Time. Figure 2 reports the cumulative pseudo-regret as a function of time under both the clustered and well-separated reward configurations. The experimental setup and the compared algorithms are identical to those used for the final-regret comparison in the main text.

The trajectories provide additional insight into where the regret of different methods originates. As expected, Known-Priority UCB-DP has the slowest regret growth throughout the horizon, since it does not incur any cost for learning the priority structure and only needs to estimate the reward parameters.

The behavior of Exact-Priority Two-Phase clearly reflects its two-phase construction. Its regret increases rapidly during the initial priority-learning phase, when the algorithm deliberately probes the system to identify the complete priority order. Once the priority structure has been sufficiently resolved, the growth becomes substantially slower, since the remaining uncertainty is primarily due to reward estimation. This confirms that a non-negligible fraction of its overall regret is attributable to the cost of exact priority identification.

The Relevance-Aware Priority-UCB algorithm exhibits a similar qualitative transition, but accumulates noticeably less regret. Rather than insisting on exact recovery of the entire ordering, it gradually suppresses probing once the remaining priority uncertainty becomes sufficiently unimportant for the admission decision. As a result, the algorithm enters an exploitation-dominated regime earlier while still retaining enough priority information to make near-optimal decisions. This effect is visible in both reward configurations and explains the lower terminal regret reported in the main text.

The remaining baselines exhibit a qualitatively different behavior. Naive-Importance UCB does not explicitly learn the priority structure, and therefore continues to suffer from structural decision errors even as the reward estimates improve. Similarly, Random ETC relies on a fixed exploration budget rather than adapting exploration to the statistical difficulty or decision relevance of the instance. A large budget such as $B = 1000$ incurs a substantial up-front exploration cost, whereas a smaller budget such as $B = 300$ may commit to an imperfect priority estimate. In both cases, the resulting regret can continue to grow at a considerably faster rate than that of the proposed algorithms.

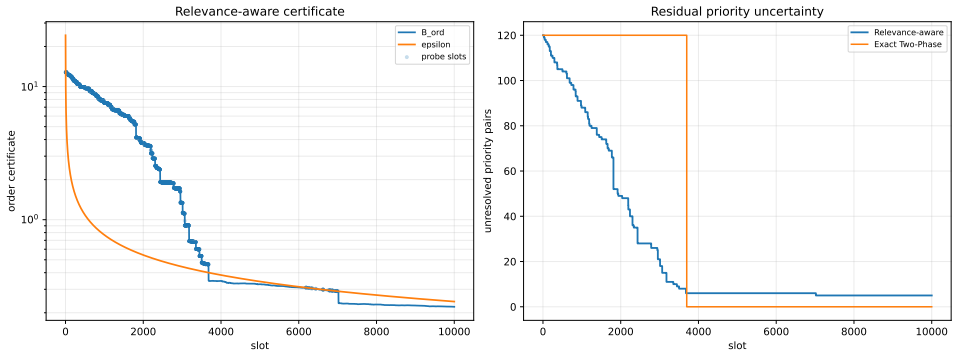


Figure 3: Illustration of the relevance-aware priority-learning mechanism in a representative run. Left: evolution of the order-sensitivity certificate B_t^{ord} and the threshold ε_t, with markers indicating explicit priority-probing slots. Priority probing is suppressed once the remaining order uncertainty satisfies $B_t^{\mathrm{ord}} \leq \varepsilon_t$. Right: number of unresolved priority pairs over time. Exact-Priority Two-Phase eventually resolves the entire priority order, whereas Relevance-Aware learning deliberately retains several low-impact unresolved pairs whose possible effect on the admission value is already negligible.

These trajectories therefore reveal an important distinction that is less visible from final-regret statistics alone: the advantage of Relevance-Aware learning is not merely a smaller constant exploration budget, but an earlier transition from priority learning to decision-oriented exploitation that is determined by the value impact of the remaining uncertainty.

The Relevance-Aware Certificate Reduces Priority Probing. We also inspect the internal behavior of the Relevance-Aware algorithm in a representative run. Figure 3 visualizes two complementary quantities.

The left panel shows the order-sensitivity bound B_t^{ord} together with the time-varying certificate threshold ε_t. Recall that B_t^{ord} quantifies an upper bound on the possible decision-value loss induced by the currently unresolved priority relations. The Relevance-Aware rule compares this bound with the admissible tolerance ε_t: when $B_t^{\mathrm{ord}} \leq \varepsilon_t$, the remaining priority ambiguity is sufficiently small from the perspective of the current decision problem, and additional priority probing is unnecessary. The markers in the figure indicate slots in which explicit priority probes are performed.

At the beginning of the horizon, the priority structure is highly uncertain and B_t^{ord} is large. The algorithm therefore actively collects priority information, causing the certificate to decrease substantially. As learning proceeds, many pairwise relations are either resolved or become sufficiently low-impact that the aggregate order-sensitivity bound approaches the tolerance level. Once the certificate condition is satisfied, the algorithm can suppress additional probes even though the complete priority permutation has not been identified.

The right panel makes this distinction particularly clear by plotting the number of unresolved priority pairs over time. With $M = 16$ devices, there are initially $\binom{16}{2} = 120$ unknown pairwise relations. The Exact-Priority Two-Phase baseline ultimately eliminates all of this uncertainty, since exact identification of the full order is a prerequisite for leaving its priority-learning phase. In contrast, Relevance-Aware learning progressively resolves those relations that matter for decision quality but deliberately leaves a small number of low-impact pairs unresolved even at the end of the horizon.

Importantly, these residual pairs do not represent a failure of the learning procedure. They are precisely the type of uncertainty that the proposed method is designed to tolerate: although their exact ordering remains statistically unresolved, the certificate guarantees that their potential effect on the admission value is already below the prescribed relevance threshold. Resolving them would therefore consume additional probing slots while providing little corresponding reduction in decision regret.

Taken together, the two panels provide a direct mechanistic explanation for the regret improvement observed in Figure 2. Exact-Priority learning treats full permutation recovery as the learning objective, whereas Relevance-Aware learning treats priority information as an intermediate resource and

stops acquiring it once the remaining uncertainty is sufficiently benign for the downstream admission problem.

F.2 Favorable-Regime Experiments

The favorable-regime experiments are designed to test settings in which exact recovery of the full priority order is statistically expensive but not necessary for near-optimal admission. These experiments focus on the comparison between Exact-Priority Two-Phase and Relevance-Aware Priority-UCB, with Known-Priority UCB-DP as an upper reference. By default, the favorable-regime script uses

$$M = 16, \qquad T = 10{,}000,$$

with 20 independent instances and 20 stochastic repetitions per condition. The same reward-learning and priority-learning confidence parameters as in the main validation experiments are used.

We study four types of conditions. First, the reward-homogeneity regime varies the reward spread. For spread parameter Δ, rewards are generated as

$$\theta_i = \theta_0 + U_i, \qquad U_i \sim \mathrm{Unif}(-\Delta, \Delta),$$

with $\theta_0 = 0.6$, clipped to $[0.01, 0.99]$. When $\Delta = 0$, all devices have identical rewards. The sweep uses

$$\Delta \in \{0, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}.$$

Demand probabilities are static and sampled from $\mathrm{Unif}(0.3, 0.7)$, and gates are sampled from $\mathrm{Unif}(0.5, 0.9)$. This regime tests whether the relevance-aware algorithm avoids resolving priority relations whose reward gaps are negligible.

Second, the small-effective-probability regime creates a set of low-gate devices. A fraction 0.5 of devices are core devices with $\gamma_i \sim \mathrm{Unif}(0.6, 0.9)$. The remaining devices have gate probabilities near a swept value γ_{low}, sampled within a radius 0.02 and clipped to $[0.01, 0.99]$. The sweep uses

$$\gamma_{\mathrm{low}} \in \{0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.60\}.$$

Rewards are sampled from $\mathrm{Unif}(0.2, 0.8)$, and static demand probabilities are sampled from $\mathrm{Unif}(0.3, 0.7)$. This regime tests whether priority relations involving rarely gate-active devices can remain unresolved without significant regret.

Third, the rare-contextual-overlap regime partitions devices into 4 balanced random groups. Time is divided into blocks of length 200. In each block, one group has high demand probability $p_{\mathrm{high}} = 0.7$, while all other groups have low demand probability p_{low}. The active group rotates periodically across blocks. The sweep uses

$$p_{\mathrm{low}} \in \{0.01, 0.03, 0.05, 0.10, 0.20, 0.30, 0.40\}.$$

Rewards are sampled from $\mathrm{Unif}(0.2, 0.8)$, and gates are sampled from $\mathrm{Unif}(0.6, 0.9)$. This regime tests whether the algorithm avoids spending priority probes on device pairs that rarely co-occur in demand.

Fourth, the ordinary-control regime uses moderate rewards, gates, and static demands:

$$\theta_i \sim \mathrm{Unif}(0.2, 0.8), \quad \gamma_i \sim \mathrm{Unif}(0.5, 0.9), \quad p_i \sim \mathrm{Unif}(0.3, 0.7).$$

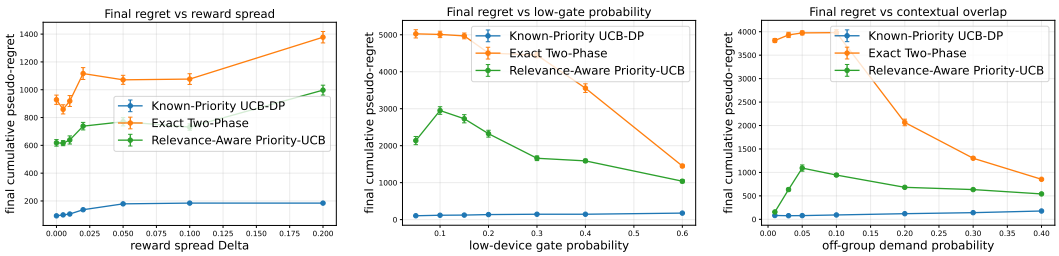
This condition serves as a baseline where priority information is generally more relevant.

Across all experiments, curves report means across generated instances and stochastic repetitions. Error bands or error bars use normal-approximation 95% intervals based on standard errors. Final-regret summaries, probe counts, unresolved-pair counts, checkpoint trajectories, representative runs, and the exact run command are saved in the experiment output directory to support reproducibility.

F.2.1 Additional Experiments in Favorable Regimes

We provide additional experiments to examine the behavior of Relevance-Aware Priority-UCB (RA) in the favorable regimes motivated by our priority-misordering sensitivity analysis.

Near-homogeneous rewards. We first vary the reward spread Δ, which controls the separation between neighboring device reward means. Small Δ produces nearly homogeneous rewards, so exchanging the order of two devices tends to have a limited effect on the expected admission value.



(a) Near-homogeneous rewards. (b) Low-gate regime. (c) Rare contextual overlap.

Figure 4: Final cumulative pseudo-regret as the structural parameters of the three favorable regimes are varied.

Figure 4a shows that RA consistently incurs less regret than Exact-Priority Two-Phase across the entire range of reward spreads. Both algorithms generally incur larger regret as the reward profile becomes more heterogeneous, since priority mistakes involving devices with more distinct reward values can have larger consequences. At small Δ, by contrast, unresolved pairwise relations can be tolerated more aggressively because many local inversions have only a limited value impact.

The dependence is not perfectly monotone, which is expected in finite-horizon experiments: changing Δ affects not only the sensitivity of the admission value to priority errors but also the reward-learning dynamics and the identity of the most relevant devices. The important observation is therefore not strict monotonicity, but that RA remains systematically below Exact-Priority while avoiding the requirement of completely resolving low-consequence ordering differences.

Low gate probabilities. We next study a regime in which half of the devices form a core set with relatively large gate probabilities, while the remaining half are designated as low-impact devices. The gate probability of the latter group, denoted by γ_{low}, is varied along the horizontal axis. This regime creates a particularly interesting statistical tradeoff. When γ_{low} is small, a low-gate device has a smaller effect on the service outcome and hence many priority relations involving it have reduced decision relevance. At the same time, such relations can be statistically difficult to identify because informative service outcomes involving a low-gate device occur less frequently.

As shown in Figure 4b, this combination is especially costly for Exact-Priority Two-Phase. For small γ_{low}, Exact-Priority spends substantial exploration effort identifying relations involving devices that rarely affect the final reward. Its regret therefore increases sharply as the low-gate devices become harder to identify.

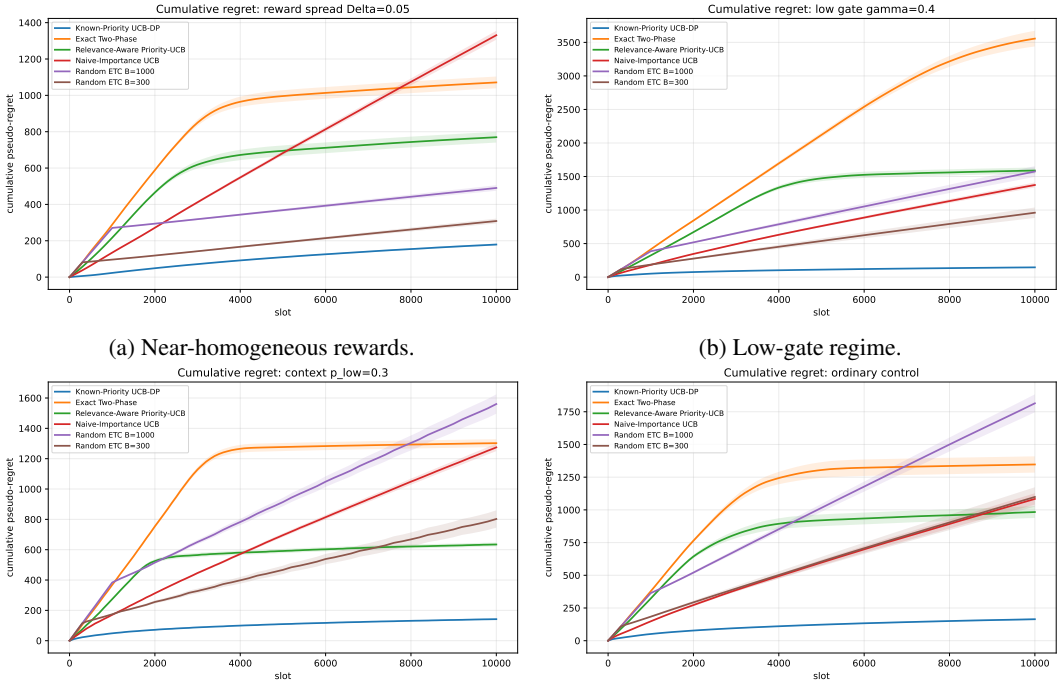
RA substantially mitigates this effect. Because it evaluates unresolved relations through their downstream value impact, it can stop probing many low-gate pairs before their exact ordering is statistically resolved. The difference between RA and Exact-Priority is consequently particularly pronounced in the low-γ_{low} region.

The RA curve itself need not be monotone. Decreasing the gate probability simultaneously makes a relation less relevant and makes it harder to statistically resolve. The observed behavior reflects this competition between information scarcity and decision irrelevance, which is precisely the situation for which relevance-aware exploration is designed.

Rare contextual overlap. The third regime varies the frequency with which devices from different contextual groups become simultaneously active. We partition the devices into four groups. Devices within a group have relatively strong co-activation, while the off-group demand/co-activation probability p_{off} controls how often devices belonging to different groups compete in the same context.

Figure 4c reveals the strongest manifestation of the relevance-aware effect. When p_{off} is very small, cross-group priority relations are both difficult to learn and rarely relevant to actual admission decisions. Exact-Priority nevertheless attempts to recover these relations as part of the complete permutation, leading to very large exploration regret.

RA instead leaves many such cross-group relations unresolved. At very small overlap probabilities, its regret can therefore remain close to the known-priority benchmark, even though a nontrivial fraction of the global priority order is still unknown.



(a) Near-homogeneous rewards.

(b) Low-gate regime.

(c) Rare contextual overlap.

(d) Ordinary control setting.

Figure 5: Cumulative pseudo-regret for representative favorable regimes and the ordinary control.

As p_{off} increases, cross-group pairs compete more frequently and their relative priority becomes increasingly relevant. RA consequently has stronger incentives to resolve them. The performance gap between RA and Exact-Priority then narrows as the problem approaches a setting in which most pairwise relations are genuinely decision-relevant.

These parameter sweeps highlight the key distinction between the two algorithms. Exact-Priority pays according to the difficulty of recovering the entire latent order, whereas RA pays primarily for the subset of priority information that is important for the admission objective.

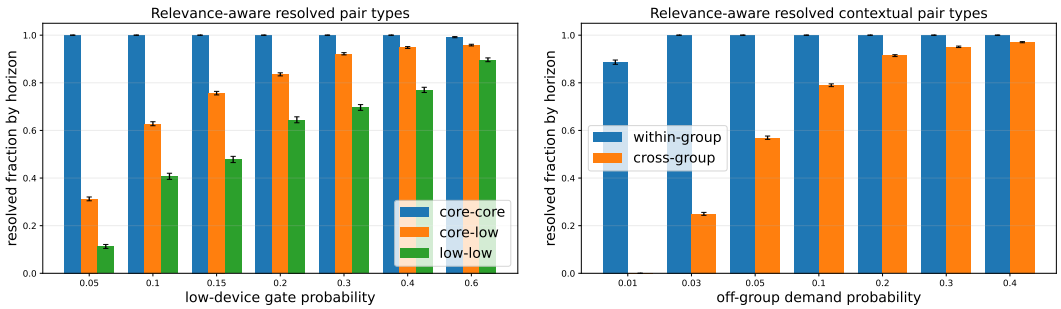
Representative Regret Trajectories. Figure 5 complements the final-regret results by showing cumulative pseudo-regret over time for representative instances from the three favorable regimes. We use $\Delta = 0.05$ for near-homogeneous rewards, $\gamma_{\text{low}} = 0.4$ for the low-gate regime, and $p_{\text{off}} = 0.3$ for rare contextual overlap. An ordinary setting without deliberately introduced low-relevance structure is included for comparison.

Across all three favorable regimes, the trajectory of Exact-Priority Two-Phase exhibits the characteristic cost of full order identification: a substantial amount of regret is accumulated while the algorithm continues probing priority relations before transitioning to reward-oriented exploitation. RA transitions earlier because its stopping criterion does not require every remaining pair to be resolved.

In the near-homogeneous reward instance, RA reduces the exploration cost relative to Exact-Priority and subsequently exhibits a much slower regret growth. This is consistent with the fact that many unresolved inversions involve devices with similar reward values and therefore have limited influence on the admission objective.

The effect is stronger in the low-gate instance. Exact-Priority continues to accumulate substantial regret while attempting to resolve low-gate relations, whereas RA begins to flatten considerably earlier. This illustrates the mismatch between identification difficulty and decision relevance: low-gate pairs can be expensive to identify even though their possible effect on reward is small.

A similar phenomenon appears in the contextual-overlap experiment. Once RA has learned the priority relations that are repeatedly relevant within frequently interacting groups, its regret growth becomes small even though some cross-group relations remain unresolved. Exact-Priority instead incurs the additional cost required to complete the global order.



(a) Low-gate regime.

(b) Rare contextual overlap.

Figure 6: Priority relations resolved by Relevance-Aware learning under structured favorable regimes.

The ordinary control shows a smaller, although still visible, advantage for RA. This comparison is useful because it confirms that the particularly large gains in the structured regimes do not arise merely from a generic tuning advantage; they are associated with the presence of relations whose exact identification is disproportionately expensive relative to their decision impact.

The heuristic baselines further illustrate that no fixed exploration budget is uniformly optimal. For some individual favorable instances, Random ETC with $B = 300$ can perform competitively, or even outperform RA, because its predetermined budget happens to match the amount of priority information needed in that instance. Similarly, priority-oblivious heuristics can perform reasonably well when priority itself has weak decision impact. However, such behavior is instance dependent and requires an exploration budget chosen without directly measuring the residual decision relevance. RA instead provides an adaptive mechanism that determines when additional priority information is worth acquiring.

Thus, these representative trajectories are not intended to establish pointwise dominance over every heuristic on every instance. Rather, they illustrate the central comparison of interest: relative to exact priority identification, RA can terminate costly priority exploration substantially earlier while preserving decision quality.

Pair resolution under heterogeneous gate probabilities. Figure 6a separates pairs into three classes:

$$\text{core–core}, \qquad \text{core–low}, \qquad \text{low–low}.$$

Half of the devices belong to the high-gate core set, while the remaining half have gate probability γ_{low}. The result provides a direct visualization of relevance-aware exploration. RA resolves essentially all core–core relations throughout the tested range. These pairs involve devices that frequently influence the service outcome and therefore remain important for admission decisions. In contrast, when γ_{low} is small, RA resolves only a fraction of the relations involving low-gate devices. The effect is strongest for low–low pairs, whose ordering has the smallest potential impact. Core–low pairs occupy an intermediate position. As γ_{low} increases, this separation gradually disappears. Low-gate devices become more capable of affecting the final service outcome, making their priority relations increasingly relevant. Accordingly, the fractions of resolved core–low and low–low pairs approach one. This experiment provides a more direct validation of the proposed exploration principle than aggregate regret alone: RA does not simply perform fewer comparisons uniformly. Instead, it selectively preserves high-value priority information while suppressing exploration of relations whose influence on the downstream objective is small.

Pair resolution under rare contextual overlap. Figure 6b reports an analogous diagnostic for the contextual-overlap regime. We distinguish within-group from cross-group priority pairs. When the off-group activity probability is very small, RA resolves almost all of the within-group relations while leaving most cross-group relations unresolved. At the smallest overlap level, essentially no cross-group ordering information needs to be identified, whereas the majority of within-group relations are still learned. As p_{off} increases, cross-group devices become simultaneously relevant more frequently. The resolved fraction of cross-group pairs therefore increases steadily and eventually approaches that of within-group pairs. This behavior is exactly what the sensitivity analysis predicts. A pairwise priority relation matters only when the two devices can jointly affect the same admission decision. Rare contextual overlap attenuates this effect, allowing RA to retain uncertainty about

cross-group relations without paying substantial policy regret. Together, Figure 6 shows that the empirical advantage of RA arises from structured selective learning, rather than from simply reducing the total number of probes. The algorithm learns high-impact relations almost completely, while deliberately tolerating uncertainty precisely where the priority-misordering sensitivity is small.