SAMPLE-EFFICIENT LLM EVALUATION VIA PURE EXPLORATION UNDER TASK HETEROGENEITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study sample-efficient LLM evaluation from preference feedback under task heterogeneity, where LLM comparisons are actively collected across diverse tasks to identify the best-performing model overall. We formulate this problem as fixed-confidence pure exploration under a contextual multinomial logit model with task-dependent model features, inducing heterogeneous preferences coupled through a shared latent parameter. Using a change-of-measure argument and a local quadratic approximation, we derive a Fisher-geometric separation criterion that characterizes the identification difficulty. A tractable relaxation of this criterion admits an explicit directional-variance representation, suggesting that comparisons should be allocated to reduce the worst uncertainty over decision-changing alternatives. Building on this principle, we develop MULTI-FW, a global Frank–Wolfe procedure that coordinates task-conditioned comparison allocations. Theoretically, combining estimator concentration and consistency with our proposed task-aware tracking, we establish asymptotically optimal separation growth despite exogenous task arrivals. For any fixed $\varepsilon \in (0, 1)$, suitably tuning forced exploration yields an $O_{\mathrm{a.s.}}(\log^{1+\varepsilon}(1/\delta))$ sample-complexity upper bound for MULTI-FW, approaching the lower bound's logarithmic confidence dependence. Finally, we evaluate MULTI-FW in a semi-real environment constructed from the Arena Human Preference 140K dataset, using logged human preferences and real LLM responses. MULTI-FW achieves faster separation growth than task-agnostic Elo-style, uniform-random, and dueling-style baselines, with at least a 15.6-percentage-point gain in trajectory-averaged identification accuracy and more than a 62% reduction in the comparison budget needed for stable 90% accuracy.

1 INTRODUCTION

Preference-based evaluation is widely used to compare LLMs, since evaluators can often choose the preferred output from a small set of candidates more reliably than assign calibrated numerical scores. This paradigm appears in arena-style LLM evaluation (Chiang et al., 2024), as well as recommendation, advertising, and content ranking. Since comparisons are costly, evaluation systems should actively allocate them to the LLM subsets most informative for the final decision.

A key feature of LLM evaluation is task heterogeneity. For example, in arena-style LLM evaluation, LLMs are compared across diverse prompt categories such as summarization, code generation, and mathematical reasoning. LLM performance and comparison informativeness can vary substantially across these tasks, since the same LLMs may produce different task-specific outputs and preference probabilities. Similar heterogeneity arises in recommendation and advertising across user segments, query types, or market contexts. Thus, comparisons are collected under heterogeneous tasks, while the evaluation procedure often targets a global decision under a task distribution.

This paper studies preference-based LLM evaluation under task heterogeneity as a fixed-confidence pure-exploration problem. At each round, a task is drawn from a target evaluation distribution. After observing the task, the learner selects a subset of candidate LLMs to compare, obtains their task-specific outputs, and observes only the preferred output. We model this winner-only feedback using a contextual MNL model with task-dependent output features and a shared latent preference parameter. The goal is to identify the LLM with the highest task-averaged utility with high confidence using as few comparisons as possible. This task-averaged objective is important because many evaluation

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.

platforms ultimately require a single deployment, recommendation, or leaderboard decision under a target evaluation distribution, even though the evidence used to support that decision comes from heterogeneous tasks. The central question is:

Can we actively allocate task-dependent comparisons to identify the task-averaged best LLM with high confidence using as few comparisons as possible?

This problem introduces four unique challenges. First, task-dependent preference heterogeneity requires preserving task-specific differences while exploiting cross-task statistical coupling. Second, task-specific information must support a task-averaged decision, requiring coordinated allocations rather than independent within-task optimization. Third, exogenous task arrivals constrain both the tracking of evolving allocations and information acquisition, since comparisons are available only within the realized task. Finally, sample-efficient fixed-confidence stopping requires valid confidence certification under adaptively collected winner-only feedback without excessive comparison costs. These challenges motivate globally coordinated comparison designs through the aggregate Fisher geometry, with task-aware execution and reliable stopping.

Our problem is related to pure exploration and dueling bandits, but differs in its task-dependent comparison structure. Existing bandit pure-exploration methods typically assume a fixed design space with static arm features, including multi-armed, linear, generalized linear, and structured MNL/Plackett-Luce settings (Kaufmann et al., 2016; Garivier & Kaufmann, 2016; Soare et al., 2014; Xu et al., 2018; Fiez et al., 2019; Jedra & Proutiere, 2020; Kazerouni & Wein, 2021; Jun et al., 2021; Rivera & Tewari, 2024; Gupta et al., 2021). Dueling-bandit methods usually posit a single global preference matrix or latent utility vector (Yue et al., 2012; Bengs et al., 2021; Saha & Gopalan, 2019; 2020; Yue & Joachims, 2011; Bengs et al., 2024; Saad et al., 2026). Recent active preference-learning methods target different feedback models or objectives, such as absolute-feedback evaluation, ranking-error reduction, or regret/suboptimality minimization (Wu et al., 2026; Lanctot et al., 2026; Mehta et al., 2025). In contrast, we study fixed-confidence best-LLM identification with task-dependent preference feedback, where comparisons must be coordinated across heterogeneous task blocks to identify the task-averaged best LLM.

Our methodological starting point is an information-theoretic characterization of the difficulty of identifying the task-averaged best LLM from task-specific comparison data. Any correct procedure must collect enough Kullback–Leibler information to rule out admissible alternative parameters under which a different LLM is task-averaged optimal. A local quadratic approximation to this change-of-measure constraint motivates a Fisher-geometric separation criterion, whose empirical counterpart is half the squared distance from the current estimate to the admissible alternative set in the cumulative Fisher metric. Relaxing the parameter-class restriction on alternatives yields an unrestricted surrogate, for which we establish the directional-variance representation

$$\widetilde{\mathrm{Sep}}_t = \frac{1}{2}\left(\max_{i \neq \hat{i}_t} a_{t,i}^\top H_t(\hat{\theta}_t)^{-1} a_{t,i}\right)^{-1},$$

where $\hat{i}_t$ is the unique empirically best LLM, $H_t(\hat{\theta}_t)$ is the cumulative Fisher information, and $a_{t,i}$ is the normalized task-averaged feature difference between $\hat{i}_t$ and competitor i. This suggests a direct allocation principle: reduce the worst directional uncertainty over decision-changing alternatives.

Building on this surrogate, we develop MULTI-FW, a task-aware active allocation algorithm for LLM evaluation. Using the shared-parameter estimate, it jointly optimizes task-conditioned target allocations to minimize the worst directional variance under the aggregate Fisher information. Despite this coupling, its Frank–Wolfe linear minimization oracle decomposes across task blocks, selecting within each task a comparison set most informative along the same globally hardest direction. The resulting mixtures are formed before task arrival and implemented online through task-aware deficit tracking, while forced exploration ensures sufficient information growth. Thus, MULTI-FW combines global blockwise design with task-conditioned execution under exogenous task arrivals.

On the theoretical side, we establish concentration and consistency for the constrained MLE under forced exploration. Optimal-design stability and task-aware tracking yield allocation convergence and asymptotically optimal separation growth. These results justify δ-correct stopping and almost-sure sample-complexity upper bounds for MULTI-FW under both restricted and unrestricted alternatives. In the restricted case, the characteristic-time term has the same dependence as the local quadratic

lower bound. Tuning forced exploration yields $\tau_\delta = O_{\text{a.s.}}(\log^{1+\varepsilon}(1/\delta))$ for any fixed $\varepsilon \in (0, 1)$ as $\delta \downarrow 0$, approaching the lower bound's logarithmic confidence dependence.

Finally, we evaluate MULTI-FW in a semi-real environment constructed from the Arena Human Preference 140K dataset. We derive task-dependent features from real LLM responses and sample feedback afresh from a contextual MNL model fitted to logged human preferences. Across 100 replications, MULTI-FW achieves faster separation growth than task-agnostic Elo-style, uniform-random, and dueling-style baselines, with at least a 15.6-percentage-point gain in trajectory-averaged identification accuracy and more than a 62% reduction in the budget for stable 90% accuracy.

In summary, our contributions are fourfold: (i) We formulate preference-based LLM evaluation under task heterogeneity as a fixed-confidence pure-exploration problem, requiring coordinated task-conditioned comparison allocations. (ii) We introduce a Fisher-geometric separation criterion for best-LLM identification and derive a directional-variance representation of its unrestricted empirical surrogate, motivating allocation rules that reduce the worst uncertainty over decision–changing alternatives. (iii) We develop MULTI-FW, combining global blockwise Frank–Wolfe design with task-aware tracking, and establish an almost-sure sample-complexity upper bound nearly matching the lower bound's logarithmic confidence dependence. (iv) Arena-derived semi-real experiments demonstrate that MULTI-FW achieves faster separation growth and higher identification accuracy with fewer comparisons than all three baselines.

2 Related Work and Our Distinctions

Pure exploration under structured models. A large body of work studies pure exploration with adaptive sampling, including multi-armed bandits (Kaufmann et al., 2016; Garivier & Kaufmann, 2016), linear bandits (Soare et al., 2014; Xu et al., 2018; Fiez et al., 2019; Jedra & Proutiere, 2020), generalized linear and logistic bandits (Kazerouni & Wein, 2021; Jun et al., 2021; Rivera & Tewari, 2024), and MNL pure exploration under Plackett–Luce feedback (Gupta et al., 2021). These works provide algorithmic foundations for adaptive allocation. However, they are primarily fixed-design problems: the learner samples within a single static geometry. In contrast, our setting involves randomly realized heterogeneous tasks with task-specific feature matrices and comparison geometries. Efficient allocation therefore requires coordinating comparison effort across task blocks, rather than optimizing within one fixed design space.

Pure exploration in dueling bandits. Our work is also related to pure exploration in dueling bandits (Yue et al., 2012; Bengs et al., 2021), where feedback is comparative. Parametric dueling methods include PAC identification under preferences induced by Plackett–Luce models (Saha & Gopalan, 2019; 2020). Nonparametric formulations consider broader preference-matrix assumptions (Yue & Joachims, 2011; Bengs et al., 2024; Saad et al., 2026). These works highlight the importance of comparison feedback, but their preference structure is static. Our setting instead induces a family of task-dependent preference structures. The learner must use comparison feedback collected under different tasks to estimate a shared preference parameter and identify a global task-averaged best LLM, which is not captured by standard dueling-bandit formulations.

Model evaluation and active preference data collection. Recent work studies sample-efficient model evaluation and active preference collection for modern AI systems. Active contextual dueling bandits address contextual preference learning and target regret/suboptimality (Mehta et al., 2025). The FAQ estimator improves efficiency for fixed benchmarks with absolute feedback (Wu et al., 2026), while active evaluation of general agents targets ranking-error reduction over discrete tasks (Lanctot et al., 2026). These works share our motivation of reducing evaluation cost, but differ in both feedback model and statistical objective. In our setting, efficient evaluation requires allocating comparisons across task-conditioned preference problems, rather than estimating model scores, minimizing regret, or reducing ranking error.

3 Problem Formulation

Task-dependent contextual preference model. We consider N candidate LLMs indexed by $[N] = \{1, \dots, N\}$ and J task types indexed by $[J] = \{1, \dots, J\}$. At each round t, a task $q_t \in [J]$ is

drawn i.i.d. from an evaluation distribution $p = (p_1, \ldots, p_J)$, where $p_j = \mathbb{P}(q_t = j)$. We assume $p_j > 0$ for all $j \in [J]$, since zero-probability tasks can be omitted. We treat p as known because it represents the target benchmark or deployment mixture defining the task-averaged evaluation objective, as in standard LLM evaluation suites that aggregate performance over pre-specified task categories (Hendrycks et al., 2021; Liang et al., 2022; Chiang et al., 2024).

Each task induces a distinct feature representation for the candidate LLM outputs: let $x_{j,i} \in \mathbb{R}^d$ denote the feature vector of the response produced by LLM i under task j, and let $X_j = [x_{j,1}, \ldots, x_{j,N}]^\top \in \mathbb{R}^{N \times d}$. For example, $x_{j,i}$ may encode response embeddings, rubric-based features, or evaluator-model features. Throughout, we assume $\|x_{j,i}\|_2 \le 1$ for all $j \in [J]$ and $i \in [N]$.

At round t, after observing the realized task q_t, the learner selects a comparison set, or assortment, $S_t \in \mathcal{S} := \{S \subseteq [N] : |S| = K\}$ of fixed size K. We assume a shared preference parameter $\theta^* \in \Theta := B_2(0, W) = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le W\}$ for known $W > 0$, and consider winner-only preference feedback: the user or evaluator returns a preferred LLM $Y_t \in S_t$ according to the contextual multinomial logit (MNL) choice model (Chen et al., 2020; Oh & Iyengar, 2021; Zhang & Sugiyama, 2023; Luo et al., 2026)

$$\mathbb{P}_{\theta^*}\big(Y_t = i \mid q_t = j,\, S_t = S\big) = \rho_{j,S,i}(\theta^*) := \frac{\exp(x_{j,i}^\top \theta^*)}{\sum_{\ell \in S} \exp(x_{j,\ell}^\top \theta^*)}, \qquad i \in S.$$

For $\theta \in \Theta$, define the task-specific utility of LLM i under task j as $u_{j,i}(\theta) := x_{j,i}^\top \theta$. Thus, task heterogeneity enters through the feature matrices $\{X_j\}_{j=1}^J$, while the shared parameter θ^* statistically couples preferences across tasks.

Task-averaged identification and fixed-confidence objective. Since LLM performance varies across tasks, we define optimality at the level of the task mixture. For any $\theta \in \Theta$, the expected utility of LLM i under p is $\bar{u}_i(\theta) := \mathbb{E}_{q \sim p}[u_{q,i}(\theta)] = \bar{x}_i^\top \theta$, where $\bar{x}_i := \sum_{j=1}^J p_j x_{j,i}$. Accordingly, define the task-averaged best-LLM map $i^*(\theta) := \arg\max_{i \in [N]} \bar{u}_i(\theta) = \arg\max_{i \in [N]} \bar{x}_i^\top \theta$. We focus on parameters for which the task-averaged best LLM is unique and assume throughout that $\theta^* \in \mathcal{R}_\Theta := \{\theta \in \Theta : i^*(\theta) \text{ is unique}\}$, so the identification target is $i^* := i^*(\theta^*)$.

Importantly, the identification target depends only on the task-averaged features $\{\bar{x}_i\}_{i=1}^N$, whereas statistical information is generated through the task-specific feature matrices $\{X_j\}_{j=1}^J$. Thus, the decision geometry and the information geometry are misaligned: decision boundaries are defined by the task-averaged directions $\bar{x}_i - \bar{x}_{i'}$, while each comparison contributes information only through the realized task-specific matrix X_{q_t}. This local-to-global mismatch motivates allocation rules that coordinate comparisons across tasks rather than treating each task independently.

To achieve this pure-exploration objective, a sequential evaluation policy must decide which comparisons to collect, when to stop, and which LLM to recommend. Let $\mathcal{F}_{t-1} := \sigma((q_s, S_s, Y_s) : 1 \le s \le t-1)$ denote the history before round t. After observing the realized task q_t, the learner selects a comparison set $S_t \in \mathcal{S}$ according to a sampling rule measurable with respect to $\sigma(\mathcal{F}_{t-1}, q_t)$. For a prescribed confidence level $\delta \in (0,1)$, the policy stops at a stopping time τ_δ and, whenever $\tau_\delta < \infty$, outputs a final recommendation $\hat{i}_{\tau_\delta}$. The policy is δ-correct if $\mathbb{P}_\theta\Big(\tau_\delta < \infty,\, \hat{i}_{\tau_\delta} \notin \arg\max_{i \in [N]} \bar{x}_i^\top \theta\Big) \le \delta,\ \forall\, \theta \in \Theta$. Among δ-correct policies, the goal is to minimize the expected sample complexity $\mathbb{E}_{\theta^*}[\tau_\delta]$. Thus, the task-dependent pure-exploration problem is to actively allocate task-conditioned comparisons so as to identify the task-averaged best LLM using as few comparisons as possible.

4 Information-Theoretic Characterization and Separation Geometry

We first derive an information-theoretic lower bound that motivates a separation criterion and the associated population design problem, then develop a tractable unrestricted surrogate with an explicit directional-variance representation that underlies our active allocation algorithm.

Information-theoretic lower bound. Recall that $i^*(\theta^*)$ is the task-averaged best LLM and $\theta^* \in \mathcal{R}_\Theta$. Thus, identification only requires ruling out parameters in Θ under which another LLM strictly

overtakes $i^*(\theta^*)$. For any $\theta \in \mathcal{R}_\Theta$, define the restricted strict alternative set $\mathrm{Alt}_\Theta(\theta) := \{\lambda \in \Theta : \exists\, i \neq i^*(\theta),\ \bar{x}_i^\top \lambda > \bar{x}_{i^*(\theta)}^\top \lambda\}$, where strict improvement makes $i^*(\theta)$ suboptimal.

For $t \geq 1$, define $N_{j,S}(t) := \sum_{s=1}^{t} \mathbf{1}\{q_s = j, S_s = S\}$ and $N_j(t) := \sum_{S \in \mathcal{S}} N_{j,S}(t)$. The following lemma states that reliable identification requires sufficient accumulated KL information to distinguish the true parameter from every decision-changing alternative.

Lemma 1 (Change-of-measure lower bound). *For any δ-correct policy satisfying $\mathbb{E}_{\theta^*}[\tau_\delta] < \infty$ and any $\lambda \in \mathrm{Alt}_\Theta(\theta^*)$, $\sum_{j=1}^{J} \sum_{S \in \mathcal{S}} \mathbb{E}_{\theta^*}[N_{j,S}(\tau_\delta)] \mathrm{KL}(P_{\theta^*}(\cdot \mid j, S) \,\|\, P_\lambda(\cdot \mid j, S)) \gtrsim \log(1/\delta)$.*

For the contextual MNL model, let $\mu_{j,S}(\theta) := \sum_{i \in S} \rho_{j,S,i}(\theta) x_{j,i}$ and $\Sigma_{j,S}(\theta) := \sum_{i \in S} \rho_{j,S,i}(\theta)(x_{j,i} - \mu_{j,S}(\theta))(x_{j,i} - \mu_{j,S}(\theta))^\top$. For λ near θ^*, the KL information admits the local quadratic approximation $\mathrm{KL}(P_{\theta^*}(\cdot \mid j, S) \,\|\, P_\lambda(\cdot \mid j, S)) \approx \frac{1}{2}(\lambda - \theta^*)^\top \Sigma_{j,S}(\theta^*)(\lambda - \theta^*)$.

Because task arrivals are exogenous, the population design is parameterized by conditional comparison allocations. For each task j, let $w_j = (w_{j,S})_{S \in \mathcal{S}} \in \Delta(\mathcal{S})$, with $w_{j,S} := \mathbb{E}_{\theta^*}[N_{j,S}(\tau_\delta)] / \mathbb{E}_{\theta^*}[N_j(\tau_\delta)]$, and write $\mathcal{W} := \Delta(\mathcal{S})^J$. For $w = (w_1, \ldots, w_J) \in \mathcal{W}$, define $V(w; \theta) := \sum_{j=1}^{J} p_j \sum_{S \in \mathcal{S}} w_{j,S} \Sigma_{j,S}(\theta)$. Lemma 1 and the local approximation yield:

Theorem 1 (Population separation). *Under the local quadratic approximation above, any δ-correct policy satisfies*

$$\mathbb{E}_{\theta^*}[\tau_\delta] \gtrsim \log(1/\delta) \left(\sup_{w \in \mathcal{W}} \inf_{\lambda \in \mathrm{Alt}_\Theta(\theta^*)} \frac{1}{2} \|\theta^* - \lambda\|^2_{V(w;\theta^*)} \right)^{-1}.$$

Separation and population design. For $\theta \in \mathcal{R}_\Theta$ and $w \in \mathcal{W}$, define the restricted separation

$$\psi_\Theta(\theta, w) := \frac{1}{2} \inf_{\lambda \in \mathrm{Alt}_\Theta(\theta)} \|\theta - \lambda\|^2_{V(w;\theta)}. \tag{1}$$

It measures how well w separates θ from the nearest admissible decision-changing alternative. Since larger separation corresponds to smaller intrinsic sample complexity, define $\psi^*_\Theta(\theta) := \max_{w \in \mathcal{W}} \psi_\Theta(\theta, w)$ and $C^*_\Theta(\theta) := \arg\max_{w \in \mathcal{W}} \psi_\Theta(\theta, w)$. Whenever $\psi^*_\Theta(\theta) > 0$, set $T^*_\Theta(\theta) := (\psi^*_\Theta(\theta))^{-1}$, so Theorem 1 reads $\mathbb{E}_{\theta^*}[\tau_\delta] \gtrsim T^*_\Theta(\theta^*) \log(1/\delta)$. The resulting population design is global: each w_j controls comparisons only within task j, whereas decision boundaries are determined by task-averaged utilities and Fisher information aggregates all task-specific comparison matrices. Thus, comparison effort must be coordinated across tasks rather than optimized within each task.

Empirical separation and directional-variance surrogate. Since the population criterion depends on the unknown parameter θ^*, we use its plug-in empirical counterpart. Let $\hat{\theta}_t \in \Theta$ denote the current estimate based on observations collected before round t, and define $H_t(\theta) := \sum_{s=1}^{t-1} \Sigma_{q_s, S_s}(\theta)$. The corresponding restricted empirical separation over the bounded parameter class is

$$\mathrm{Sep}_t^\Theta := \frac{1}{2} \inf_{\lambda \in \mathrm{Alt}_\Theta(\hat{\theta}_t)} \|\hat{\theta}_t - \lambda\|^2_{H_t(\hat{\theta}_t)}. \tag{2}$$

The restriction to Θ, however, makes the inner optimization less tractable. We therefore relax $\mathrm{Alt}_\Theta(\theta)$ to $\mathrm{Alt}_\infty(\theta) := \{\lambda \in \mathbb{R}^d : \exists\, i \neq i^*(\theta),\ \bar{x}_i^\top \lambda > \bar{x}_{i^*(\theta)}^\top \lambda\}$ and define the corresponding unrestricted surrogate $\widetilde{\mathrm{Sep}}_t := \frac{1}{2} \inf_{\lambda \in \mathrm{Alt}_\infty(\hat{\theta}_t)} \|\hat{\theta}_t - \lambda\|^2_{H_t(\hat{\theta}_t)}$. This relaxation admits an explicit directional-variance representation.

Theorem 2 (Directional-variance representation). *Suppose that $\hat{\imath}_t := i^*(\hat{\theta}_t)$ is unique and $H_t(\hat{\theta}_t)$ is positive definite. For each $i \neq \hat{\imath}_t$, let $a_{t,i} := (\bar{x}_{\hat{\imath}_t} - \bar{x}_i)/((\bar{x}_{\hat{\imath}_t} - \bar{x}_i)^\top \hat{\theta}_t)$. Then*

$$\widetilde{\mathrm{Sep}}_t = \frac{1}{2} \left(\max_{i \neq \hat{\imath}_t} a_{t,i}^\top H_t(\hat{\theta}_t)^{-1} a_{t,i} \right)^{-1}. \tag{3}$$

Thus, maximizing the surrogate is equivalent to suppressing the worst Fisher-metric variance over task-averaged decision directions, yielding the allocation principle of directing comparison effort toward the hardest decision-relevant directions across tasks. For population design, define

$$\widetilde{\psi}(\theta, w) := \frac{1}{2} \inf_{\lambda \in \mathrm{Alt}_\infty(\theta)} \|\theta - \lambda\|^2_{V(w;\theta)}. \tag{4}$$

Since $\mathrm{Alt}_\Theta(\theta) = \mathrm{Alt}_\infty(\theta) \cap \Theta \subseteq \mathrm{Alt}_\infty(\theta)$, we have $\widetilde{\psi}(\theta, w) \le \psi_\Theta(\theta, w)$ for every $w \in \mathcal{W}$. Thus, the restricted criterion provides the lower-bound benchmark, while its unrestricted surrogate underlies MULTI-FW. Section 6 analyzes both within a common framework.

5 MULTI-FW: TASK-AWARE ACTIVE ALLOCATION

In this section, we develop MULTI-FW (Algorithm 1), which combines a global task-conditioned Frank–Wolfe design with online tracking and forced exploration.

Task-conditioned target design. At round t, we use the constrained MLE $\hat{\theta}_t \in \arg\min_{\theta\in\Theta} L_t(\theta)$ based on the past negative log-likelihood $L_t(\theta) := -\sum_{s=1}^{t-1} \log \rho_{q_s,S_s,Y_s}(\theta)$. The unrestricted empirical surrogate in (3) depends on the accumulated Fisher information $H_t(\hat{\theta}_t)$. For long-run design, we use its population counterpart $\widetilde{\psi}$ in (4) under a task-conditioned blockwise allocation $\pi = (\pi_1, \dots, \pi_J) \in \mathcal{W}$. At time t, before observing q_t, with $\hat{\theta}_t$, $\hat{\imath}_t$, and $\{a_{t,i}\}_{i\neq\hat{\imath}_t}$ fixed, define

$$g(\pi; \hat{\theta}_t) := \max_{i\neq\hat{\imath}_t} a_{t,i}^\top V(\pi; \hat{\theta}_t)^{-1} a_{t,i}.$$

Whenever $V(\pi; \hat{\theta}_t)$ is positive definite, Theorem 2 gives $\widetilde{\psi}(\hat{\theta}_t, \pi) = \frac{1}{2} g(\pi; \hat{\theta}_t)^{-1}$. Hence the practical MULTI-FW target is $\pi^{\mathrm{FW}}(\hat{\theta}_t) \in \arg\min_{\pi\in\mathcal{W}} g(\pi; \hat{\theta}_t)$, equivalently maximizing $\widetilde{\psi}(\hat{\theta}_t, \pi)$.

Frank–Wolfe optimization on a product of simplices. The objective $g(\pi; \hat{\theta}_t)$ is convex in π, and the feasible region is the product of simplices $\Delta(\mathcal{S})^J$, so we solve it with a block-structured Frank-Wolfe procedure. Crucially, despite the coupling through aggregate Fisher information, the linear minimization oracle decomposes across task blocks along a shared globally hardest direction. At iteration m, let $V^{(m)} := V(\pi^{(m)}; \hat{\theta}_t)$, choose a current worst competitor $i^{(m)} \in \arg\max_{i\neq\hat{\imath}_t} a_{t,i}^\top (V^{(m)})^{-1} a_{t,i}$, and set $u^{(m)} := (V^{(m)})^{-1} a_{t,i^{(m)}}$. Since the derivative with respect to $\pi_{j,S}$ is proportional to $-p_j u^{(m)\top} \Sigma_{j,S}(\hat{\theta}_t) u^{(m)}$, the oracle separates over task blocks: for each task type j, it selects

$$S_j^{(m)} \in \arg\max_{S\in\mathcal{S}} u^{(m)\top} \Sigma_{j,S}(\hat{\theta}_t) u^{(m)}.$$

Let $e_{j,S_j^{(m)}}$ denote the corresponding vertex of $\Delta(\mathcal{S})$. The blockwise Frank–Wolfe update is

$$\pi_j^{(m+1)} = (1-\eta_m)\pi_j^{(m)} + \eta_m e_{j,S_j^{(m)}}$$

for every $j \in [J]$. The resulting iterate is used as the task-conditioned target design.

Execution with tracking and forced exploration. For $t \ge 2$, before observing q_t, we form the round-t target $w(t-1)$ using the Frank–Wolfe iterate above when the empirical best LLM is unique; otherwise, we select a target in $\mathcal{W}$ using a fixed, δ-independent rule. The target is then implemented by task-wise deficit tracking. Define the task-gated cumulative target mass $C_{j,S}(t) := \sum_{s=2}^{t} \mathbf{1}\{q_s = j\}\, w_{j,S}(s-1)$, $C_j(t) := \big(C_{j,S}(t)\big)_{S\in\mathcal{S}}$, with $C_{j,S}(1) := 0$. After observing $q_t = j$, on a non-forced round the policy selects $S_t \in \arg\min_{S\in\mathrm{supp}(C_j(t))} \{N_{j,S}(t-1) - C_{j,S}(t)\}$.

To guarantee sufficient information growth, we use a task-aware forced-exploration safeguard. Define $\Sigma_{j,S}^{\mathrm{unw}} := \frac{1}{2}\sum_{i,i'\in S}(x_{j,i} - x_{j,i'})(x_{j,i} - x_{j,i'})^\top$ and $U_t := \sum_{s=1}^{t} \Sigma_{q_s,S_s}^{\mathrm{unw}}$. Suppose $c_X := \lambda_{\min}\Big(\binom{N}{K}^{-1} \sum_{j=1}^{J} p_j \sum_{S\in\mathcal{S}} \Sigma_{j,S}^{\mathrm{unw}}\Big) > 0$. For a fixed $\alpha \in (0,1)$, tracking is overridden by task-wise round robin whenever $t = 1$ or $\lambda_{\min}(U_{t-1}) < c_X (t-1)^\alpha$. Despite exogenous task arrivals, the predictable trigger ensures that task proportions among forced rounds converge to p. Task-wise round robin then pools complementary information across potentially rank-deficient tasks. This guarantees $\lambda_{\min}(U_t) \gtrsim c_X t^\alpha$ for all sufficiently large t, a.s.; see Appendix B.1.

6 THEORETICAL GUARANTEES

We establish a fixed-confidence sample-complexity upper bound for a generic version of MULTI-FW that allows any fixed alternative domain $\mathcal{D}$ satisfying $\Theta \subseteq \mathcal{D} \subseteq \mathbb{R}^d$. For $\theta \in \mathcal{R}_\Theta$, let $\mathrm{Alt}_\mathcal{D}(\theta) :=$

Algorithm 1 MULTI-FW: Global Frank–Wolfe Design with Tracking and Forced Exploration

Initialize U_0, $\hat{\theta}_1$, counts $N_{j,S}(0)$, targets $C_{j,S}(1) = 0$, and round-robin pointers.
for $t = 1, 2, \ldots$ **do**
  Update the constrained MLE $\hat{\theta}_t$ from $\mathcal{F}_{t-1}$.
  if the fixed-confidence stopping condition of Section 6 holds **then**
    stop and output $\hat{i}_t$.
  end if
  if $t \geq 2$ **then**
    Compute the Frank–Wolfe iterate $w(t-1)$ if the empirical best LLM is unique; otherwise use a fixed, δ-independent fallback rule.
  end if
  Observe task type $q_t = j$.
  if $t \geq 2$ **then**
    Update $C_{j,S}(t) = C_{j,S}(t-1) + w_{j,S}(t-1)$ for all S, with other blocks unchanged.
  end if
  Select S_t by round-robin if $t = 1$ or $\lambda_{\min}(U_{t-1}) < c_X(t-1)^{\alpha}$; otherwise by deficit tracking.
  Observe winner $Y_t \in S_t$ and update U_t and the realized counts.
end for

$\{\lambda \in \mathcal{D} : \exists\, i \neq i^*(\theta) \text{ such that } \bar{x}_i^\top \lambda > \bar{x}_{i^*(\theta)}^\top \lambda\}$ and define $\psi_{\mathcal{D}}(\theta, w) := \frac{1}{2} \inf_{\lambda \in \mathrm{Alt}_{\mathcal{D}}(\theta)} \|\theta - \lambda\|^2_{V(w;\theta)}, \psi^*_{\mathcal{D}}(\theta) := \max_{w \in \mathcal{W}} \psi_{\mathcal{D}}(\theta, w), C^*_{\mathcal{D}}(\theta) := \arg\max_{w \in \mathcal{W}} \psi_{\mathcal{D}}(\theta, w)$. Whenever $\hat{\theta}_{t+1} \in \mathcal{R}_\Theta$, the generic algorithm selects $w(t) \in C^*_{\mathcal{D}}(\hat{\theta}_{t+1})$, thus recovering the unrestricted target used by MULTI-FW in Section 5 when $\mathcal{D} = \mathbb{R}^d$.

Constrained-MLE confidence and consistency Recall the constrained MLE $\hat{\theta}_t \in \arg\min_{\theta \in \Theta} L_t(\theta)$. Let $D_X := \max_{j \in [J],\, S \in \mathcal{S}} \max_{i,i' \in S} \|x_{j,i} - x_{j,i'}\|_2 \leq 2$. For $\delta \in (0,1)$ and $\lambda > 0$, define $\gamma_t^\lambda(\delta) := \frac{\sqrt{\lambda}}{2} + \frac{4}{\sqrt{\lambda}} \log\Big[\frac{2^{3d/2}}{\delta}\big(1 + \frac{2tK}{\lambda d}\big)^{d/2}\Big]$.

Proposition 1 (Constrained-MLE confidence)**.** *For any $\lambda > 0$, with probability at least $1 - \delta$, $\|\hat{\theta}_t - \theta^*\|_{H_t(\hat{\theta}_t)} \leq e^{3WD_X} \gamma_t^\lambda(\delta)$ holds simultaneously for every t such that $\lambda_{\min}(H_t(\theta^*)) \geq \lambda$.*

The proof first controls the score at θ^* in the true Fisher metric via a self-normalized inequality. The constrained-MLE optimality condition, together with a self-concordant Hessian comparison along the segment between θ^* and $\hat{\theta}_t$, then converts this score bound into the plug-in Fisher-norm guarantee.

Combined with the information growth guaranteed by Proposition S1, this immediately yields strong consistency: for any fixed $\alpha \in (0,1)$, $\hat{\theta}_t \to \theta^*$ almost surely; see Corollary S2 in Appendix C.1.5.

Design Stability and Task-Aware Tracking In our setting, the mismatch between task-specific information and task-averaged decisions precludes directly applying the regularity argument in Jedra & Proutiere (2020). We instead establish joint continuity of $\psi_{\mathcal{D}}$ and optimal-design stability through multiplicative Fisher comparisons; see Proposition S4 in Appendix C.2.3.

Whenever $N_j(t) > 0$, define the empirical conditional allocation $\widehat{w}_{j,S}(t) := N_{j,S}(t)/N_j(t)$. Combined with task-aware tracking, the stability above yields the following allocation convergence.

Proposition 2 (Task-aware allocation convergence)**.** *Suppose that $\theta^* \in \mathcal{R}_\Theta$ and $p_j > 0$ for every $j \in [J]$. Under the generic* MULTI-FW *considered above,* $\mathrm{dist}(\widehat{w}(t), C^*_{\mathcal{D}}(\theta^*)) \longrightarrow 0$ *a.s.*

Unlike standard tracking where every action is directly available, here only comparisons within the exogenously realized task can be selected. Crucially, since the target design $w(t)$ is chosen before q_{t+1} is realized, a martingale strong law aligns the task-gated targets with their long-run task frequencies, while deficit tracking and forced-exploration overrides contribute only sublinear error.

Separation Growth and Fixed-Confidence Guarantee Proposition 2 gives convergence of the within-task allocations, while $N_j(t)/t \to p_j$ aligns the task frequencies with the target mixture. Together with $\hat{\theta}_t \to \theta^*$ and joint continuity of $\psi_{\mathcal{D}}$, this yields Proposition 3 below, which establishes

the optimal growth rate of the empirical separation $\mathrm{Sep}_t^{\mathcal{D}} := \frac{1}{2}\inf_{\lambda \in \mathrm{Alt}_{\mathcal{D}}(\hat{\theta}_t)} \|\hat{\theta}_t - \lambda\|^2_{H_t(\hat{\theta}_t)}$, defined whenever $\hat{\theta}_t \in \mathcal{R}_\Theta$, and links allocation convergence to fixed-confidence sample complexity.

Proposition 3 (Asymptotic growth of separation)**.** *Under the conditions of Proposition 2,* $\mathrm{Sep}_t^{\mathcal{D}}/(t-1) \to \psi^*_{\mathcal{D}}(\theta^*)$ *almost surely.*

For a given $\delta \in (0,1)$, fix any $\lambda_\delta > 0$ and define the confidence threshold $\beta_t^{\lambda_\delta}(\delta) := \frac{1}{2}e^{6WD_X}\left(\gamma_t^{\lambda_\delta}(\delta)\right)^2$. Set $\kappa_{W,K} := \left[1 + (K-1)e^{2W}\right]^{-2}$. We use the stopping rule

$$\tau_\delta := \inf\left\{ t \geq 2 : \begin{array}{l} \hat{i}_t := i^*(\hat{\theta}_t) \text{ is unique,} \\ \lambda_{\min}(U_{t-1}) \geq \kappa_{W,K}^{-1}\lambda_\delta, \\ \mathrm{Sep}_t^{\mathcal{D}} > \beta_t^{\lambda_\delta}(\delta) \end{array} \right\}, \tag{5}$$

and output $\hat{i}_{\tau_\delta}$ whenever $\tau_\delta < \infty$. We establish its δ-correctness in Appendix C.3.2.

We now turn to sample complexity. By exploiting the separation-growth result of Proposition 3 and setting $\lambda_\delta := \log^{\frac{2\alpha}{1+\alpha}}(1/\delta)$ in the stopping rule, we obtain Theorem 3 below.

Theorem 3 (Almost-sure sample-complexity upper bound)**.** *Suppose that* $\theta^* \in \mathcal{R}_\Theta$ *satisfies* $\psi^*_{\mathcal{D}}(\theta^*) > 0$*, and define* $T^*_{\mathcal{D}}(\theta^*) := 1/\psi^*_{\mathcal{D}}(\theta^*)$*. Fix* $\alpha \in (0,1)$ *and specialize* $\lambda_\delta := \log^{\frac{2\alpha}{1+\alpha}}(1/\delta)$*. Under the generic* MULTI-FW *considered above and the stopping rule* (5)*,*

$$\mathbb{P}_{\theta^*}\left(\limsup_{\delta \downarrow 0} \frac{\tau_\delta}{\log^{\frac{2}{1+\alpha}}(1/\delta)} \leq \max\left\{ 8e^{6WD_X} T^*_{\mathcal{D}}(\theta^*), \left(\frac{1}{\kappa_{W,K} c_X}\right)^{1/\alpha} \right\} \right) = 1. \tag{6}$$

The choice of λ_δ balances the two stopping requirements: information growth requires $t \gtrsim \lambda_\delta^{1/\alpha}$, whereas confidence certification with linear separation growth requires $t \gtrsim \log^2(1/\delta)/\lambda_\delta$. The first term in the upper bound depends on the $\mathcal{D}$-based characteristic time $T^*_{\mathcal{D}}(\theta^*)$, while the second reflects the forced-information acquisition requirement. Moreover, for any fixed $\varepsilon \in (0,1)$, choosing $\alpha = (1-\varepsilon)/(1+\varepsilon)$ yields $\tau_\delta = O_{\mathrm{a.s.}}(\log^{1+\varepsilon}(1/\delta))$ as $\delta \downarrow 0$, bringing the confidence dependence arbitrarily close to the $\log(1/\delta)$ lower bound of Theorem 1; see Corollary S4 in Appendix C.3.4.

Two choices of $\mathcal{D}$ are instructive. For $\mathcal{D} = \mathbb{R}^d$, the theorem provides a sample-complexity guarantee for MULTI-FW in Section 5 under the stated target-selection rule. For $\mathcal{D} = \Theta$, the first term has the same $T^*_\Theta(\theta^*)$ dependence as the local quadratic lower bound in Theorem 1.

7 ARENA-DERIVED SEMI-REAL EVALUATION

7.1 CONSTRUCTION OF THE ARENA ENVIRONMENT

Arena data and evaluation population. We construct the semi-real environment from the public *Arena Human Preference 140K* dataset released by LM Arena, which contains 135,634 logged pairwise evaluations from real user interactions, including user conversations, candidate LLM responses, and human preferences. We retain English conversations with decisive A-versus-B preferences and focus on the 15 LLMs with the largest decisive-comparison exposure. This subset is densely supported by the logs: all $\binom{15}{2} = 105$ LLM pairs are observed, with at least 47 decisive votes per pair. Using the Arena code indicator, we partition queries into two mutually exclusive tasks, code and non-code, providing a simple but well-supported instance of task heterogeneity. For each retained comparison, we use the full user conversation and the final response from each candidate LLM to construct the task-dependent representations below.

Task-dependent features and semi-real environment. For each retained final LLM response $r_{t,i}$, we compute a 384-dimensional semantic embedding $e_{t,i} := E(r_{t,i}) \in \mathbb{R}^{384}$ using `all-MiniLM-L6-v2`. We average these embeddings within each task–LLM pair to obtain $\widetilde{x}_{k,i} := n_{k,i}^{-1} \sum_{t:\, k_t = k,\, i \in S_t} e_{t,i}$, for $k \in \{1,2\}$ and $i \in [15]$. Because pairwise MNL feedback depends only on within-task feature differences, these representations induce a 28-dimensional identifiable subspace; we work in an orthonormal basis of this subspace and denote the resulting task-dependent features by $x_{k,i} \in \mathbb{R}^{28}$. Using the retained decisive Arena comparisons, we fit

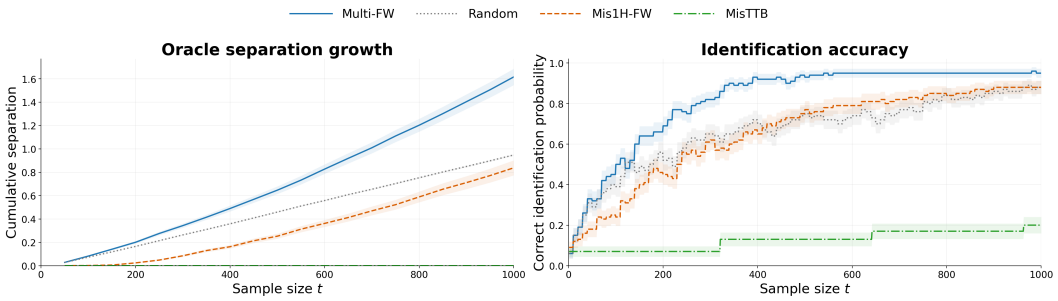

Figure 1: Performance trajectories on the Arena-derived semi-real environment: cumulative separation (left) and identification accuracy (right). Shaded regions denote ± 1 standard error over 100 replications.

Table 1: Identification performance on the Arena-derived semi-real environment.

Description	Method	Accuracy @ $t = 600$	Checkpoint AUC	$\tau_{0.8}^{\text{first}}$	$\tau_{0.9}^{\text{stable}}$
Task-dependent active allocation	MULTI-FW	**0.950**	**0.844**	**270**	**380**
Uniform random allocation	RANDOM	0.740	0.688	750	> 1000
Task-agnostic Elo-style	MIS1H-FW	0.790	0.679	620	> 1000
Non-contextual dueling-style	MISTTB	0.130	0.130	> 1000	> 1000

the shared contextual MNL model of Section 3, obtaining $\widehat{\theta}_{\text{Arena}}$ and task-dependent preference probabilities $\widehat{P}_{k,ij} := \Pr_{\widehat{\theta}_{\text{Arena}}}(i \succ j \mid k)$. These probabilities define the semi-real environment: when an algorithm selects an LLM pair under task k, the winner is newly sampled according to $\widehat{P}_{k,ij}$. Thus, task-dependent feature heterogeneity and preference structure are estimated from real Arena interactions, while the sequential comparison trajectory is generated afresh for each experimental run.

7.2 PERFORMANCE ON THE ARENA ENVIRONMENT

We compare MULTI-FW with three baselines over 100 independent replications with horizon $T = 1000$. *(i)* RANDOM*:* a uniform random allocation over all candidate comparison sets. *(ii)* MIS1H-FW*:* a task-agnostic Elo-style baseline that deliberately collapses the problem into a misspecified single-task surrogate, representing each LLM by the same one-hot feature vector across all tasks and therefore ignoring both task identity and task-dependent LLM features. *(iii)* MISTTB*:* a non-contextual dueling-style baseline, implemented as a fixed-budget adaptation of Trace-the-Best (Saha & Gopalan, 2019). It maintains an incumbent LLM and compares it against challengers in a cyclic batch procedure, without modeling task heterogeneity or exploiting task-dependent LLM features.

As shown in Figure 1, MULTI-FW substantially outperforms all three baselines in identification accuracy and achieves the fastest separation growth. The poor performance of MISTTB reflects its long-batch incumbent–challenger structure, which limits pairwise coverage over the horizon.

To summarize the identification-accuracy trajectories quantitatively, we report four complementary metrics in Table 1. Accuracy at $t = 600$ provides a representative mid-horizon snapshot; checkpoint AUC averages identification accuracy over $t = 50, 100, \ldots, 1000$; $\tau_{0.8}^{\text{first}}$ is the first 80% crossing; and $\tau_{0.9}^{\text{stable}}$ is the first time after which accuracy remains at or above 90%.

As shown in Table 1, MULTI-FW leads all baselines in both snapshot and trajectory-averaged identification accuracy, achieving 0.950 at $t = 600$ and a checkpoint AUC of 0.844, compared with baseline maxima of 0.790 and 0.688, respectively. It also reaches 80% accuracy at $t = 270$, versus the earliest baseline crossing at $t = 620$, requiring approximately 56% fewer comparisons. Moreover, MULTI-FW enters a stable 90%-accuracy regime by $t = 380$, whereas no baseline does so within the horizon, corresponding to a reduction of more than 62% in $\tau_{0.9}^{\text{stable}}$.

AI USE STATEMENT

We used generative AI tools to assist with the code-level implementation of author-specified methods, adapting data-preprocessing and experimental scripts, and organizing experimental outputs. We also used them to draft proofs in LaTeX based on proof outlines and key steps established and confirmed by the authors beforehand. We did not use these tools to formulate mathematical claims, supply key proof arguments, or design or provide feedback on the research methodology or experimental design. We also did not use them to develop theoretical models or conceptual frameworks, propose or refine research hypotheses, draw experimental conclusions, or generate synthetic datasets. Additional uses included clarification of theoretical background, language polishing, and checks of notation and typographical consistency. All retained AI-assisted material was reviewed by the authors. The authors checked the final mathematical arguments, and the implementation was tested through numerical consistency and reproducibility comparisons. The authors take full responsibility for the final text, theoretical claims, experimental results, and released code.

REPRODUCIBILITY STATEMENT

Proofs of our theoretical results are provided in Appendices A–C. Section 7.1 describes the Arena-derived environment, and Appendix D provides further details on its construction and the experimental implementation. The supplementary code includes the construction pipeline, the frozen environment used in the paper, and scripts for running the experiments and generating the figures in Section 7.2. The accompanying READMEs document dependencies, external data and encoder sources, and commands for rebuilding the environment or running the experiments directly from the bundled instance.

REFERENCES

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.

Viktor Bengs, Björn Haddenhorst, and Eyke Hüllermeier. Identifying copeland winners in dueling bandits with indifferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 226–234. PMLR, 2024.

Xi Chen, Yining Wang, and Yuan Zhou. Dynamic assortment optimization with changing contextual information. *Journal of machine learning research*, 21(216):1–44, 2020.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pp. 3052–3060. PMLR, 2020.

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pp. 998–1027. PMLR, 2016.

Shubham Gupta, Aadirupa Saha, and Sumeet Katariya. Pure exploration with structured preference feedback. *arXiv preprint arXiv:2104.05294*, 2021.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In *International Conference on Machine Learning*, pp. 5148–5157. PMLR, 2021.

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

Abbas Kazerouni and Lawrence M Wein. Best arm identification in generalized linear bandits. *Operations Research Letters*, 49(3):365–371, 2021.

Marc Lanctot, Kate Larson, Ian Gemp, and Michael Kaisers. Active evaluation of general agents: Problem definition and comparison of baseline algorithms. *arXiv preprint arXiv:2601.07651*, 2026.

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

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Yiyun Luo, Will Wei Sun, and Yufeng Liu. Rate-optimal online learning for dynamic assortment selection with positioning. *Operations Research*, 74(1):224–242, 2026.

Viraj Mehta, Syrine Belakaria, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Barbara E Engelhardt, Stefano Ermon, Jeff Schneider, and Willie Neiswanger. Sample efficient preference alignment in llms via active exploration. In *Second Conference on Language Modeling*, 2025.

Min-hwan Oh and Garud Iyengar. Multinomial logit contextual bandits: Provable optimality and practicality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 9205–9213, 2021.

Eduardo Ochoa Rivera and Ambuj Tewari. Near optimal pure exploration in logistic bandits. *arXiv preprint arXiv:2410.20640*, 2024.

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

Aadirupa Saha and Aditya Gopalan. Pac battling bandits in the plackett-luce model. In *Algorithmic Learning Theory*, pp. 700–737. PMLR, 2019.

Aadirupa Saha and Aditya Gopalan. From pac to instance-optimal sample complexity in the plackett-luce model. In *International Conference on Machine Learning*, pp. 8367–8376. PMLR, 2020.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in neural information processing systems*, 27, 2014.

Quoc Tran-Dinh, Yen-Huan Li, and Volkan Cevher. Composite convex minimization involving self-concordant-like cost functions. In *Modelling, Computation and Optimization in Information Systems and Management Sciences: Proceedings of the 3rd International Conference on Modelling, Computation and Optimization in Information Systems and Management Sciences-MCO 2015-Part I*, pp. 155–168. Springer, 2015.

Skyler Wu, Yash Nair, and Emmanuel J Candés. Efficient evaluation of llm performance with statistical guarantees. *arXiv preprint arXiv:2601.20251*, 2026.

Liyuan Xu, Junya Honda, and Masashi Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 843–851. PMLR, 2018.

Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 241–248, 2011.

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

Yu-Jie Zhang and Masashi Sugiyama. Online (multinomial) logistic bandit: Improved regret and constant computation cost. *Advances in Neural Information Processing Systems*, 36:29741–29782, 2023.

A Proofs for Section 4

A.1 Proof of Lemma 1: Change-of-Measure Lower Bound

Proof. Fix any alternative parameter $\lambda \in \mathrm{Alt}_\Theta(\theta^*)$, and write $i^* := i^*(\theta^*)$. By the definition of the restricted alternative set, there exists some $i \neq i^*$ such that $\bar{x}_i^\top \lambda > \bar{x}_{i^*}^\top \lambda$. Hence $i^* \notin \arg\max_{k\in[N]} \bar{x}_k^\top \lambda$.

Let $A := \{\tau_\delta < \infty,\ \hat{i}_{\tau_\delta} = i^*\}$ be the event that the algorithm stops and recommends the model that is best under θ^*. Since $\mathbb{E}_{\theta^*}[\tau_\delta] < \infty$, we have $\mathbb{P}_{\theta^*}(\tau_\delta < \infty) = 1$. Therefore, by δ-correctness under θ^*,

$$\mathbb{P}_{\theta^*}(A^c) = \mathbb{P}_{\theta^*}(\tau_\delta = \infty) + \mathbb{P}_{\theta^*}\Big(\tau_\delta < \infty,\ \hat{i}_{\tau_\delta} \neq i^*\Big) \leq \delta.$$

On the other hand, since $i^* \notin \arg\max_{k\in[N]} \bar{x}_k^\top \lambda$, the event A implies a finite incorrect recommendation under λ. Thus, by δ-correctness under λ,

$$\mathbb{P}_\lambda(A) \leq \mathbb{P}_\lambda\Big(\tau_\delta < \infty,\ \hat{i}_{\tau_\delta} \notin \arg\max_{k\in[N]} \bar{x}_k^\top \lambda\Big) \leq \delta.$$

Combining the two inequalities gives $\mathbb{P}_{\theta^*}(A^c) + \mathbb{P}_\lambda(A) \leq 2\delta$.

Let $\mathbb{P}_{\theta^*}^{\tau_\delta}$ and $\mathbb{P}_\lambda^{\tau_\delta}$ denote the laws of the stopped history generated by the algorithm under parameters θ^* and λ, respectively. Applying the Bretagnolle–Huber inequality to the event A under these two stopped-history measures yields

$$\mathbb{P}_{\theta^*}(A^c) + \mathbb{P}_\lambda(A) \geq \frac{1}{2}\exp(-\mathrm{KL}(\mathbb{P}_{\theta^*}^{\tau_\delta} \,\|\, \mathbb{P}_\lambda^{\tau_\delta}))\,.$$

Together with the previous upper bound, this implies $2\delta \geq \frac{1}{2}\exp\big(-\mathrm{KL}(\mathbb{P}_{\theta^*}^{\tau_\delta} \| \mathbb{P}_\lambda^{\tau_\delta})\big)$. Rearranging gives the stopped-history KL lower bound

$$\mathrm{KL}(\mathbb{P}_{\theta^*}^{\tau_\delta} \,\|\, \mathbb{P}_\lambda^{\tau_\delta}) \geq \log\frac{1}{4\delta}. \tag{7}$$

It remains to decompose this stopped-history KL in our setting. This is the standard divergence decomposition (Lattimore & Szepesvári, 2020), with each "arm" replaced by a task-assortment pair (j, S).

Recall that $N_{j,S}(\tau_\delta) := \sum_{s=1}^{\tau_\delta} \mathbf{1}\{q_s = j,\ S_s = S\}$. Let $\mathcal{H}_{\tau_\delta} := (q_1, S_1, Y_1, \ldots, q_{\tau_\delta}, S_{\tau_\delta}, Y_{\tau_\delta}, \tau_\delta)$ denote the stopped history, and let $\mathbb{P}_{\theta^*}^{\tau_\delta}$ and $\mathbb{P}_\lambda^{\tau_\delta}$ be the laws of $\mathcal{H}_{\tau_\delta}$ under parameters θ^* and λ, respectively.

At each round s, conditional on the past history $\mathcal{H}_{s-1}$, the task q_s is drawn from the same distribution under both parameters, and the sampling rule selects S_s using the same parameter-independent policy kernel. Hence these terms cancel in the likelihood ratio. The only parameter-dependent term is the conditional distribution of the winner Y_s given (q_s, S_s). Moreover, since τ_δ is a stopping time, the event $\{s \leq \tau_\delta\}$ is determined before observing any data after round s, so the stopped likelihood ratio is obtained by multiplying the one-step likelihood ratios over the realized rounds $s \leq \tau_\delta$. Therefore,

$$\log\frac{d\mathbb{P}_{\theta^*}^{\tau_\delta}}{d\mathbb{P}_\lambda^{\tau_\delta}} = \sum_{s=1}^{\tau_\delta} \log\frac{P_{\theta^*}(Y_s \mid q_s, S_s)}{P_\lambda(Y_s \mid q_s, S_s)}.$$

Taking expectation under $\mathbb{P}_{\theta^*}$, we obtain

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\theta^*}^{\tau_\delta} \,\|\, \mathbb{P}_\lambda^{\tau_\delta}) &= \mathbb{E}_{\theta^*}\left[\sum_{s=1}^{\tau_\delta} \log\frac{P_{\theta^*}(Y_s \mid q_s, S_s)}{P_\lambda(Y_s \mid q_s, S_s)}\right] \\
&= \mathbb{E}_{\theta^*}\left[\sum_{s=1}^{\tau_\delta} \mathbb{E}_{\theta^*}\left[\log\frac{P_{\theta^*}(Y_s \mid q_s, S_s)}{P_\lambda(Y_s \mid q_s, S_s)} \,\middle|\, \mathcal{H}_{s-1}, q_s, S_s\right]\right].
\end{aligned}$$

Under $\mathbb{P}_{\theta^*}$, conditional on $(\mathcal{H}_{s-1}, q_s, S_s)$, the winner satisfies $Y_s \sim P_{\theta^*}(\cdot \mid q_s, S_s)$. Hence

$$\mathbb{E}_{\theta^*}\left[\log \frac{P_{\theta^*}(Y_s \mid q_s, S_s)}{P_\lambda(Y_s \mid q_s, S_s)} \;\middle|\; \mathcal{H}_{s-1}, q_s, S_s\right] = \mathrm{KL}(P_{\theta^*}(\cdot \mid q_s, S_s) \,\|\, P_\lambda(\cdot \mid q_s, S_s)).$$

Therefore,

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\theta^*}^{\tau_\delta} \,\|\, \mathbb{P}_\lambda^{\tau_\delta}) &= \mathbb{E}_{\theta^*}\left[\sum_{s=1}^{\tau_\delta} \mathrm{KL}(P_{\theta^*}(\cdot \mid q_s, S_s) \,\|\, P_\lambda(\cdot \mid q_s, S_s))\right] \\
&= \mathbb{E}_{\theta^*}\left[\sum_{j=1}^{J} \sum_{S \in \mathcal{S}} N_{j,S}(\tau_\delta) \mathrm{KL}(P_{\theta^*}(\cdot \mid j, S) \,\|\, P_\lambda(\cdot \mid j, S))\right] \\
&= \sum_{j=1}^{J} \sum_{S \in \mathcal{S}} \mathbb{E}_{\theta^*}[N_{j,S}(\tau_\delta)] \,\mathrm{KL}(P_{\theta^*}(\cdot \mid j, S) \,\|\, P_\lambda(\cdot \mid j, S)). \qquad (8)
\end{aligned}$$

This proves the desired divergence decomposition.

Combining (7) and (8) gives

$$\sum_{j=1}^{J} \sum_{S \in \mathcal{S}} \mathbb{E}_{\theta^*}[N_{j,S}(\tau_\delta)] \mathrm{KL}(P_{\theta^*}(\cdot \mid j, S) \,\|\, P_\lambda(\cdot \mid j, S)) \geq \log \frac{1}{4\delta} \gtrsim \log(1/\delta),$$

which proves the claim. □

A.2 Proof of Theorem 1: Population Separation

Proof. If $\mathbb{E}_{\theta^*}[\tau_\delta] = +\infty$, the claimed lower bound holds trivially. Hence, in the remainder of the proof, assume $\mathbb{E}_{\theta^*}[\tau_\delta] < \infty$.

We start from Lemma 1. For any $\lambda \in \mathrm{Alt}_\Theta(\theta^*)$, any δ-correct policy must satisfy

$$\sum_{j=1}^{J} \sum_{S \in \mathcal{S}} \mathbb{E}_{\theta^*}[N_{j,S}(\tau_\delta)] \mathrm{KL}(P_{\theta^*}(\cdot \mid j, S) \,\|\, P_\lambda(\cdot \mid j, S)) \gtrsim \log(1/\delta).$$

Under the local quadratic approximation of the contextual MNL likelihood, for λ close to θ^*, we treat $\mathrm{KL}(P_{\theta^*}(\cdot \mid j, S) \,\|\, P_\lambda(\cdot \mid j, S)) \approx \frac{1}{2}(\lambda - \theta^*)^\top \Sigma_{j,S}(\theta^*)(\lambda - \theta^*)$. Substituting this approximation into the change-of-measure lower bound gives, for every $\lambda \in \mathrm{Alt}_\Theta(\theta^*)$,

$$\frac{1}{2}(\lambda - \theta^*)^\top \left(\sum_{j=1}^{J} \sum_{S \in \mathcal{S}} \mathbb{E}_{\theta^*}[N_{j,S}(\tau_\delta)] \Sigma_{j,S}(\theta^*)\right)(\lambda - \theta^*) \gtrsim \log(1/\delta). \qquad (9)$$

We now rewrite the matrix inside the quadratic form in terms of the task-conditioned allocation. For each task j, let $N_j(\tau_\delta) := \sum_{S \in \mathcal{S}} N_{j,S}(\tau_\delta)$ be the number of times task j appears before stopping. Since the task types are drawn i.i.d. from the known distribution $p = (p_1, \ldots, p_J)$, and since the stopping rule is adapted to the observed history, the process $\sum_{s=1}^{t}(\mathbf{1}\{q_s = j\} - p_j)$ is a martingale. Hence, by the optional stopping/Wald identity, $\mathbb{E}_{\theta^*}[N_j(\tau_\delta)] = p_j \,\mathbb{E}_{\theta^*}[\tau_\delta]$.

Define the conditional allocation proportions $w_{j,S} := \mathbb{E}_{\theta^*}[N_{j,S}(\tau_\delta)]/\mathbb{E}_{\theta^*}[N_j(\tau_\delta)]$, for $S \in \mathcal{S}$ and $j \in [J]$. Then $w_j = (w_{j,S})_{S \in \mathcal{S}} \in \Delta(\mathcal{S})$, since $w_{j,S} \geq 0$ and $\sum_{S \in \mathcal{S}} w_{j,S} = \sum_{S \in \mathcal{S}} \mathbb{E}_{\theta^*}[N_{j,S}(\tau_\delta)]/\mathbb{E}_{\theta^*}[N_j(\tau_\delta)] = 1$.

For notational convenience, define the task-wise information matrix $V_j(w_j; \theta^*) := \sum_{S \in \mathcal{S}} w_{j,S} \Sigma_{j,S}(\theta^*)$. Recalling the definition of the population Fisher information, $V(w; \theta^*) =$

$\sum_{j=1}^{J} p_j V_j(w_j;\theta^*)$, where $w=(w_1,\dots,w_J)\in\mathcal{W}$, we obtain

$$\begin{aligned}\sum_{j=1}^{J}\sum_{S\in\mathcal{S}}\mathbb{E}_{\theta^*}[N_{j,S}(\tau_\delta)]\Sigma_{j,S}(\theta^*) &= \sum_{j=1}^{J}\mathbb{E}_{\theta^*}[N_j(\tau_\delta)]\sum_{S\in\mathcal{S}} w_{j,S}\Sigma_{j,S}(\theta^*)\\ &= \sum_{j=1}^{J} p_j\mathbb{E}_{\theta^*}[\tau_\delta]\, V_j(w_j;\theta^*)\\ &= \mathbb{E}_{\theta^*}[\tau_\delta]\, V(w;\theta^*). \end{aligned} \tag{10}$$

Substituting (10) into (9), we get that for every $\lambda\in\mathrm{Alt}_\Theta(\theta^*)$,

$$\frac{\mathbb{E}_{\theta^*}[\tau_\delta]}{2}(\lambda-\theta^*)^\top V(w;\theta^*)(\lambda-\theta^*)\gtrsim\log(1/\delta).$$

Since the preceding display holds for every $\lambda\in\mathrm{Alt}_\Theta(\theta^*)$, it also holds for the least favorable alternative. Hence

$$\frac{\mathbb{E}_{\theta^*}[\tau_\delta]}{2}\inf_{\lambda\in\mathrm{Alt}_\Theta(\theta^*)}(\lambda-\theta^*)^\top V(w;\theta^*)(\lambda-\theta^*)\gtrsim\log(1/\delta).$$

Thus, for policy π with the induced conditional allocation w, we obtain

$$\mathbb{E}_{\theta^*}[\tau_\delta]\gtrsim\frac{\log(1/\delta)}{\inf_{\lambda\in\mathrm{Alt}_\Theta(\theta^*)}\frac{1}{2}\|\theta^*-\lambda\|^2_{V(w;\theta^*)}}.$$

Since $w\in\mathcal{W}$, we have

$$\inf_{\lambda\in\mathrm{Alt}_\Theta(\theta^*)}\frac{1}{2}\|\theta^*-\lambda\|^2_{V(w;\theta^*)}\le\sup_{w\in\mathcal{W}}\inf_{\lambda\in\mathrm{Alt}_\Theta(\theta^*)}\frac{1}{2}\|\theta^*-\lambda\|^2_{V(w;\theta^*)}.$$

Therefore, every δ-correct policy π satisfies the policy-independent lower bound

$$\mathbb{E}_{\theta^*}[\tau_\delta]\gtrsim\frac{\log(1/\delta)}{\sup_{w\in\mathcal{W}}\inf_{\lambda\in\mathrm{Alt}_\Theta(\theta^*)}\frac{1}{2}\|\theta^*-\lambda\|^2_{V(w;\theta^*)}}.$$

This proves the claimed population separation lower bound. □

A.3 Proof of Theorem 2: Directional-Variance Representation

Proof. For notational simplicity, write $H_t:=H_t(\hat\theta_t)$ and $\hat\imath:=\hat\imath_t$. Since $\hat\imath_t$ is unique, for every $i\neq\hat\imath$, $(\bar x_{\hat\imath}-\bar x_i)^\top\hat\theta_t>0$. Define $d_{t,i}:=\bar x_{\hat\imath}-\bar x_i$ and $a_{t,i}:=d_{t,i}/(d_{t,i}^\top\hat\theta_t)$. Then $a_{t,i}^\top\hat\theta_t=1$.

By the definition of the unrestricted strict alternative set,

$$\mathrm{Alt}_\infty(\hat\theta_t)=\bigcup_{i\neq\hat\imath}\left\{\lambda\in\mathbb{R}^d:\bar x_i^\top\lambda>\bar x_{\hat\imath}^\top\lambda\right\}=\bigcup_{i\neq\hat\imath}\left\{\lambda\in\mathbb{R}^d:d_{t,i}^\top\lambda<0\right\}.$$

For each $i\neq\hat\imath$, the closed half-space $\{\lambda: d_{t,i}^\top\lambda\le 0\}$ is the closure of the strict half-space $\{\lambda: d_{t,i}^\top\lambda<0\}$. Since the map $\lambda\mapsto\|\hat\theta_t-\lambda\|^2_{H_t}$ is continuous,

$$\inf_{\lambda:\,d_{t,i}^\top\lambda<0}\|\hat\theta_t-\lambda\|^2_{H_t}=\inf_{\lambda:\,d_{t,i}^\top\lambda\le 0}\|\hat\theta_t-\lambda\|^2_{H_t}.$$

Therefore, since there are finitely many competitors,

$$\inf_{\lambda\in\mathrm{Alt}_\infty(\hat\theta_t)}\|\hat\theta_t-\lambda\|^2_{H_t}=\min_{i\neq\hat\imath}\inf_{\lambda:\,d_{t,i}^\top\lambda\le 0}\|\hat\theta_t-\lambda\|^2_{H_t}. \tag{11}$$

Fix a competitor $i \neq \hat{i}$. We now compute $\inf_{\lambda:\, d_{t,i}^\top \lambda \leq 0} \|\hat{\theta}_t - \lambda\|_{H_t}^2$. Let $\Delta := \hat{\theta}_t - \lambda$. The constraint $d_{t,i}^\top \lambda \leq 0$ is equivalent to $d_{t,i}^\top(\hat{\theta}_t - \Delta) \leq 0$, or $d_{t,i}^\top \Delta \geq d_{t,i}^\top \hat{\theta}_t$. Dividing both sides by the positive scalar $d_{t,i}^\top \hat{\theta}_t$, this becomes $a_{t,i}^\top \Delta \geq 1$. Thus

$$\inf_{\lambda:\, d_{t,i}^\top \lambda \leq 0} \|\hat{\theta}_t - \lambda\|_{H_t}^2 = \inf_{\Delta:\, a_{t,i}^\top \Delta \geq 1} \|\Delta\|_{H_t}^2.$$

Since H_t is positive definite, the Cauchy–Schwarz inequality in the H_t-inner product gives, for any $\Delta \in \mathbb{R}^d$,

$$(a_{t,i}^\top \Delta)^2 = \left(\Delta^\top H_t^{1/2} H_t^{-1/2} a_{t,i}\right)^2 \leq (\Delta^\top H_t \Delta)(a_{t,i}^\top H_t^{-1} a_{t,i}).$$

Therefore, for any Δ satisfying $a_{t,i}^\top \Delta \geq 1$, $\|\Delta\|_{H_t}^2 = \Delta^\top H_t \Delta \geq (a_{t,i}^\top \Delta)^2 / (a_{t,i}^\top H_t^{-1} a_{t,i}) \geq 1/(a_{t,i}^\top H_t^{-1} a_{t,i})$.

This lower bound is attained for the corresponding closed half-space problem. Indeed, take

$$\Delta_i^\star := \frac{H_t^{-1} a_{t,i}}{a_{t,i}^\top H_t^{-1} a_{t,i}}.$$

Then $a_{t,i}^\top \Delta_i^\star = (a_{t,i}^\top H_t^{-1} a_{t,i})/(a_{t,i}^\top H_t^{-1} a_{t,i}) = 1$, and $\|\Delta_i^\star\|_{H_t}^2 = (\Delta_i^\star)^\top H_t \Delta_i^\star = 1/(a_{t,i}^\top H_t^{-1} a_{t,i})$. Hence

$$\inf_{\lambda:\, d_{t,i}^\top \lambda \leq 0} \|\hat{\theta}_t - \lambda\|_{H_t}^2 = \frac{1}{a_{t,i}^\top H_t^{-1} a_{t,i}}. \tag{12}$$

Taking the minimum over competitors, i.e., substituting (12) into (11), gives

$$\begin{aligned} \inf_{\lambda \in \mathrm{Alt}_\infty(\hat{\theta}_t)} \|\hat{\theta}_t - \lambda\|_{H_t}^2 &= \min_{i \neq \hat{i}} \frac{1}{a_{t,i}^\top H_t^{-1} a_{t,i}} \\ &= \left(\max_{i \neq \hat{i}} a_{t,i}^\top H_t^{-1} a_{t,i}\right)^{-1}. \end{aligned} \tag{13}$$

By the definition of the unrestricted empirical surrogate, $\widetilde{\mathrm{Sep}}_t = \frac{1}{2} \inf_{\lambda \in \mathrm{Alt}_\infty(\hat{\theta}_t)} \|\hat{\theta}_t - \lambda\|_{H_t(\hat{\theta}_t)}^2$. Substituting (13) and recalling that $H_t = H_t(\hat{\theta}_t)$, we obtain

$$\widetilde{\mathrm{Sep}}_t = \frac{1}{2} \left(\max_{i \neq \hat{i}_t} a_{t,i}^\top H_t(\hat{\theta}_t)^{-1} a_{t,i}\right)^{-1},$$

which is exactly (3). This proves the claimed directional-variance representation. □

B Proofs for Section 5: Algorithms

B.1 Information Growth under Forced Exploration

Proposition S1 (Information growth)**.** *Suppose that the task types are drawn independently according to $p = (p_1, \ldots, p_J)$, with $p_j > 0$ for every $j \in [J]$. Define*

$$\Sigma_{j,S}^{\mathrm{unw}} := \frac{1}{2} \sum_{i,i' \in S} (x_{j,i} - x_{j,i'})(x_{j,i} - x_{j,i'})^\top, \qquad j \in [J], \quad S \in \mathcal{S},$$

and

$$U_t := \sum_{s=1}^{t} \Sigma_{q_s,S_s}^{\mathrm{unw}}, \qquad U_0 := 0.$$

Let

$$c_X := \lambda_{\min}\left(\binom{N}{K}^{-1} \sum_{j=1}^{J} p_j \sum_{S \in \mathcal{S}} \Sigma_{j,S}^{\mathrm{unw}}\right) > 0.$$

Fix $\alpha \in (0,1)$ and consider any nonanticipating baseline assortment sequence $\{B_t\}_{t\geq 1}$, where B_t may depend on past observations and the current task. Play B_t whenever $\lambda_{\min}(U_{t-1}) \geq c_X(t-1)^\alpha$ and $t \geq 2$; otherwise play task-wise round-robin $\mathrm{RR}_{q_t}(t)$. Each task uses a fixed cyclic ordering of $\mathcal{S}$, and its round-robin pointer advances only on its forced rounds.

Let $F(t)$ denote the number of forced-exploration overrides among rounds $2,\dots,t$, excluding initialization at round 1. Then, almost surely,

$$\limsup_{t\to\infty} \frac{F(t)}{t^\alpha} \leq 1, \qquad \liminf_{t\to\infty} \frac{\lambda_{\min}(U_t)}{t^\alpha} \geq c_X.$$

Proof. Fix an admissible baseline rule. Let $m := |\mathcal{S}| = \binom{N}{K}$ and define

$$G_j := \sum_{S\in\mathcal{S}} \Sigma^{\mathrm{unw}}_{j,S}, \qquad \Gamma_X := \frac{1}{m}\sum_{j=1}^{J} p_j G_j.$$

Thus $\lambda_{\min}(\Gamma_X) = c_X > 0$.

Let $\mathcal{F}_t$ denote the history through round t, including any algorithmic randomization, which is generated independently of the exogenous task sequence. The task-arrival model gives $\mathbb{P}(q_s = j \mid \mathcal{F}_{s-1}) = p_j$. For $s \geq 2$, define

$$I_s := \mathbf{1}\{\lambda_{\min}(U_{s-1}) < c_X(s-1)^\alpha\}.$$

Then I_s is $\mathcal{F}_{s-1}$-measurable, and $F(t) = \sum_{s=2}^{t} I_s$, with $F(1) = 0$.

Step 1: Task frequencies among forced rounds. For each $j \in [J]$, let

$$F_j(t) := \sum_{s=2}^{t} I_s \mathbf{1}\{q_s = j\}.$$

We show that the task frequencies among forced overrides converge to p almost surely on the event $\{F(t) \to \infty\}$.

Fix $j \in [J]$ and consider

$$M_j(t) := \sum_{s=2}^{t} \frac{I_s\big(\mathbf{1}\{q_s = j\} - p_j\big)}{1 + F(s)}, \qquad M_j(1) := 0.$$

Since $F(s) = F(s-1) + I_s$, both I_s and $F(s)$ are $\mathcal{F}_{s-1}$-measurable. Moreover, $M_j(t)$ is adapted and square integrable at every finite time. Writing $\Delta M_j(s) := M_j(s) - M_j(s-1)$, we obtain

$$\begin{aligned}\mathbb{E}[\Delta M_j(s) \mid \mathcal{F}_{s-1}] &= \frac{I_s}{1+F(s)}\left(\mathbb{P}(q_s = j \mid \mathcal{F}_{s-1}) - p_j\right) \\ &= 0.\end{aligned}$$

Thus $M_j(t)$ is a square-integrable martingale. Its conditional second moments satisfy

$$\mathbb{E}\Big[\big(\Delta M_j(s)\big)^2 \mid \mathcal{F}_{s-1}\Big] = p_j(1-p_j)\frac{I_s}{(1+F(s))^2}.$$

Whenever $I_s = 1$, the counter $F(s)$ increases by one. Consequently, for every t,

$$\sum_{s=2}^{t} \frac{I_s}{(1+F(s))^2} = \sum_{r=1}^{F(t)} \frac{1}{(1+r)^2} \leq \sum_{r=1}^{\infty} \frac{1}{(1+r)^2} < \infty.$$

By orthogonality of martingale increments,

$$\begin{aligned}\mathbb{E}[M_j(t)^2] &= \sum_{s=2}^{t} \mathbb{E}\Big[\big(\Delta M_j(s)\big)^2\Big] \\ &= p_j(1-p_j)\mathbb{E}\left[\sum_{r=1}^{F(t)} \frac{1}{(1+r)^2}\right] \\ &\leq p_j(1-p_j)\sum_{r=1}^{\infty} \frac{1}{(1+r)^2}.\end{aligned}$$

Hence $\sup_t \mathbb{E}[M_j(t)^2] < \infty$. The L^2-bounded martingale convergence theorem therefore implies that $M_j(t)$ converges almost surely. Since J is finite, there is an event Ω_{F} of probability one on which all the martingales $M_j(t)$ converge. All subsequent arguments are carried out on this event.

Consider a sample path in Ω_{F} on which $F(t) \to \infty$, and let σ_r be the round of the r-th counted forced override. Then $\sigma_r \to \infty$ and

$$M_j(\sigma_r) = \sum_{\ell=1}^{r} \frac{\mathbf{1}\{q_{\sigma_\ell} = j\} - p_j}{1+\ell}.$$

Thus the series on the right converges as $r \to \infty$.

We use Kronecker's lemma in the following deterministic form: if (a_r) is a real sequence, (b_r) is a positive increasing sequence with $b_r \to \infty$, and $\sum_{r=1}^{\infty} a_r/b_r$ converges, then $b_n^{-1} \sum_{r=1}^{n} a_r \to 0$. Applying this result with $a_r = \mathbf{1}\{q_{\sigma_r} = j\} - p_j$ and $b_r = r+1$ gives

$$\frac{1}{r+1} \sum_{\ell=1}^{r} \big(\mathbf{1}\{q_{\sigma_\ell} = j\} - p_j\big) \longrightarrow 0.$$

Multiplying by $(r+1)/r \to 1$ yields the same convergence with denominator r. Finally, $\sum_{\ell=1}^{F(t)} \mathbf{1}\{q_{\sigma_\ell} = j\} = F_j(t)$, and $F(t) \to \infty$. Substituting $r = F(t)$ therefore proves the first anchor:

$$\frac{F_j(t)}{F(t)} \longrightarrow p_j, \qquad j \in [J], \quad \text{a.s. on } \{F(t) \to \infty\}. \tag{14}$$

Step 2: Information per forced round. For each $j \in [J]$ and $S \in \mathcal{S}$, define

$$F_{j,S}(t) := \sum_{s=2}^{t} I_s \mathbf{1}\{q_s = j,\ S_s = S\}, \qquad U_t^{\mathrm{F}} := \sum_{s=2}^{t} I_s \Sigma^{\mathrm{unw}}_{q_s, S_s}.$$

Thus $U_t^{\mathrm{F}} = \sum_{j,S} F_{j,S}(t) \Sigma^{\mathrm{unw}}_{j,S}$. Because every information increment is positive semidefinite, $U_t \succeq U_t^{\mathrm{F}}$ for every t.

For each task, the counted forced selections follow a fixed cyclic ordering of $\mathcal{S}$. Excluding initialization only changes the starting position in that ordering. Every complete cycle selects each assortment once, and the remaining incomplete cycle selects each assortment at most once. Consequently, $F_{j,S}(t)$ equals either $\lfloor F_j(t)/m \rfloor$ or $\lceil F_j(t)/m \rceil$, so

$$\left| F_{j,S}(t) - \frac{F_j(t)}{m} \right| \le 1.$$

Set $C_{\mathrm{RR}} := \sum_{j=1}^{J} \sum_{S \in \mathcal{S}} \|\Sigma^{\mathrm{unw}}_{j,S}\|_{\mathrm{op}} < \infty$. The triangle inequality for the operator norm gives

$$\left\| U_t^{\mathrm{F}} - \frac{1}{m} \sum_{j=1}^{J} F_j(t) G_j \right\|_{\mathrm{op}} \le C_{\mathrm{RR}}.$$

On a sample path in Ω_{F} with $F(t) \to \infty$, it follows that

$$\begin{aligned} \left\| \frac{U_t^{\mathrm{F}}}{F(t)} - \Gamma_X \right\|_{\mathrm{op}} &\le \frac{C_{\mathrm{RR}}}{F(t)} + \frac{1}{m} \sum_{j=1}^{J} \left| \frac{F_j(t)}{F(t)} - p_j \right| \|G_j\|_{\mathrm{op}} \\ &\longrightarrow 0, \end{aligned}$$

where the convergence uses (14) and the finiteness of J. Moreover, the variational characterization of the minimum eigenvalue yields

$$\left| \frac{\lambda_{\min}(U_t^{\mathrm{F}})}{F(t)} - c_X \right| = \left| \lambda_{\min}\left(\frac{U_t^{\mathrm{F}}}{F(t)} \right) - \lambda_{\min}(\Gamma_X) \right| \le \left\| \frac{U_t^{\mathrm{F}}}{F(t)} - \Gamma_X \right\|_{\mathrm{op}}.$$

We have therefore established the second anchor:

$$\frac{U_t^{\mathrm{F}}}{F(t)} \longrightarrow \Gamma_X, \qquad \frac{\lambda_{\min}(U_t^{\mathrm{F}})}{F(t)} \longrightarrow c_X > 0, \quad \text{a.s. on } \{F(t) \to \infty\}. \tag{15}$$

Step 3: Sublinear number of forced rounds. Fix a sample path in Ω_{F}. Since $F(t)$ is nondecreasing, it either remains bounded or tends to infinity. If it remains bounded, then $F(t)/t^\alpha \to 0$. We therefore consider the case $F(t) \to \infty$.

Fix $\eta \in (0, c_X)$. By (15), for all sufficiently large u,

$$\lambda_{\min}(U_u^{\mathrm{F}}) \geq (c_X - \eta)F(u).$$

For sufficiently large t, set $r := F(t)$. The round $\sigma_r \leq t$ is the most recent counted forced round, and exactly $r-1$ counted forced overrides precede it. Hence $F(\sigma_r - 1) = r - 1$. Also, since counted overrides begin at round 2, $\sigma_r \geq r + 1$. Thus $\sigma_r - 1 \to \infty$ as $t \to \infty$, and the preceding information bound applies at $u = \sigma_r - 1$.

Combining that bound with $U_{\sigma_r - 1} \succeq U_{\sigma_r - 1}^{\mathrm{F}}$ and the forcing condition at σ_r, we obtain

$$\begin{aligned}(c_X - \eta)(F(t) - 1) &\leq \lambda_{\min}(U_{\sigma_r - 1}^{\mathrm{F}}) \\ &\leq \lambda_{\min}(U_{\sigma_r - 1}) \\ &< c_X(\sigma_r - 1)^\alpha \\ &\leq c_X t^\alpha.\end{aligned}$$

Therefore, for all sufficiently large t, $F(t) \leq 1 + \frac{c_X}{c_X - \eta} t^\alpha$. Dividing by t^α and taking upper limits gives

$$\limsup_{t\to\infty} \frac{F(t)}{t^\alpha} \leq \frac{c_X}{c_X - \eta}.$$

Letting $\eta \downarrow 0$ proves that this upper limit is at most one.

Together with the bounded-count case, this shows that, on every sample path in Ω_{F}, $F(t) \leq 2t^\alpha$ for all sufficiently large t. In particular, $F(t) = O(t^\alpha)$, and

$$0 \leq \frac{F(t)}{t} \leq 2t^{\alpha - 1} \longrightarrow 0,$$

because $\alpha < 1$. This proves the third anchor:

$$\limsup_{t\to\infty} \frac{F(t)}{t^\alpha} \leq 1, \qquad F(t) = O(t^\alpha) = o(t) \quad \text{a.s.} \tag{16}$$

Step 4: Growth of the cumulative information. Continue to work on a sample path in Ω_{F}. By (16), $F(t)/t \to 0$, and hence $F(t) < t - 1$ for all sufficiently large t. There are $t - 1$ rounds in $\{2, \ldots, t\}$, so at least one of them is non-forced. For such t, define the most recent non-forced round by

$$s_t := \max\{s \in \{2, \ldots, t\} : I_s = 0\}.$$

Every round in $\{s_t + 1, \ldots, t\}$ is forced. Consequently,

$$t - s_t = \sum_{s=s_t+1}^{t} I_s \leq F(t),$$

and therefore

$$0 \leq 1 - \frac{s_t}{t} \leq \frac{F(t)}{t} \longrightarrow 0.$$

Thus $s_t/t \to 1$, and also $(s_t - 1)/t = s_t/t - 1/t \to 1$.

Since round s_t is non-forced, the sampling rule gives

$$\lambda_{\min}(U_{s_t - 1}) \geq c_X(s_t - 1)^\alpha.$$

Furthermore, $U_t \succeq U_{s_t - 1}$ because every intervening information increment is positive semidefinite. Hence

$$\frac{\lambda_{\min}(U_t)}{t^\alpha} \geq \frac{\lambda_{\min}(U_{s_t - 1})}{t^\alpha} \geq c_X \left(\frac{s_t - 1}{t}\right)^\alpha.$$

The right-hand side converges to c_X, since $(s_t - 1)/t \to 1$ and $x \mapsto x^\alpha$ is continuous at 1. Taking lower limits, and recalling that $\mathbb{P}(\Omega_{\mathrm{F}}) = 1$, establishes the final anchor:

$$\liminf_{t\to\infty} \frac{\lambda_{\min}(U_t)}{t^\alpha} \geq c_X \quad \text{a.s.} \tag{17}$$

□

C Proofs for Section 6: Theoretical Guarantees

C.1 Constrained-MLE Confidence and Consistency

C.1.1 Uniform Fisher Nondegeneracy

Lemma S1 (Uniform Fisher invertibility under forced exploration)**.** *Suppose that $\|x_{j,i}\|_2 \le 1$ for all $j \in [J]$ and $i \in [N]$, and let $\Theta := B_2(0, W)$ with $\theta^* \in \Theta$. Let $c_X > 0$ be defined as in Proposition S1, and suppose that the forced-exploration safeguard therein is used with fixed $\alpha \in (0,1)$. Define $\kappa_{W,K} := \big[1 + (K-1)e^{2W}\big]^{-2} \ge e^{-4W}/K^2$. Then, for every $t \ge 1$,*

$$H_t(\theta) \succeq \kappa_{W,K} U_{t-1}, \qquad \forall \theta \in \Theta.$$

Furthermore, almost surely,

$$\liminf_{t\to\infty} \frac{\inf_{\theta\in\Theta} \lambda_{\min}\big(H_t(\theta)\big)}{t^\alpha} \ge \kappa_{W,K} c_X > 0.$$

Thus, almost surely, the cumulative Fisher information is uniformly positive definite over Θ for all sufficiently large t.

Proof. Fix $t \ge 1$ and $\theta \in \Theta$.

Part I: Uniform Fisher–proxy comparison. For any task $j \in [J]$, assortment $S \in \mathcal{S}$, and $i \in S$, $\rho_{j,S,i}(\theta) = \exp(x_{j,i}^\top \theta) / \sum_{\ell\in S} \exp(x_{j,\ell}^\top \theta)$. Since $\|x_{j,i}\|_2 \le 1$ and $\|\theta\|_2 \le W$, we have $-W \le x_{j,i}^\top \theta \le W$. It follows that

$$\rho_{j,S,i}(\theta) \ge \frac{e^{-W}}{e^{-W} + (K-1)e^{W}} = \frac{1}{1 + (K-1)e^{2W}} =: \underline{\rho}_{W,K}.$$

Therefore, for every $i, \ell \in S$, $\rho_{j,S,i}(\theta)\rho_{j,S,\ell}(\theta) \ge \underline{\rho}_{W,K}^2 = \kappa_{W,K}$. Moreover, since $1+(K-1)e^{2W} \le Ke^{2W}$, we have $\kappa_{W,K} = [1 + (K-1)e^{2W}]^{-2} \ge e^{-4W}/K^2$.

For notational simplicity, write $\rho_i := \rho_{j,S,i}(\theta)$, $x_i := x_{j,i}$, and $\mu := \sum_{i\in S} \rho_i x_i$. The single-observation Fisher information admits the pairwise representation

$$\Sigma_{j,S}(\theta) = \frac{1}{2} \sum_{i,\ell\in S} \rho_i \rho_\ell (x_i - x_\ell)(x_i - x_\ell)^\top. \tag{18}$$

Indeed,

$$\begin{aligned}
&\frac{1}{2} \sum_{i,\ell\in S} \rho_i \rho_\ell (x_i - x_\ell)(x_i - x_\ell)^\top \\
&\quad = \sum_{i\in S} \rho_i x_i x_i^\top - \left(\sum_{i\in S} \rho_i x_i\right)\left(\sum_{\ell\in S} \rho_\ell x_\ell\right)^\top \\
&\quad = \sum_{i\in S} \rho_i x_i x_i^\top - \mu\mu^\top \\
&\quad = \sum_{i\in S} \rho_i (x_i - \mu)(x_i - \mu)^\top = \Sigma_{j,S}(\theta).
\end{aligned}$$

Recall that $\Sigma_{j,S}^{\mathrm{unw}} := \frac{1}{2}\sum_{i,\ell\in S}(x_{j,i} - x_{j,\ell})(x_{j,i} - x_{j,\ell})^\top$. Using the uniform lower bound on the choice probabilities and (18), we obtain

$$\Sigma_{j,S}(\theta) \succeq \kappa_{W,K} \Sigma_{j,S}^{\mathrm{unw}}.$$

This comparison holds uniformly over all $j \in [J]$, $S \in \mathcal{S}$, and $\theta \in \Theta$. Summing over the observations collected before time t gives

$$\begin{aligned} H_t(\theta) &= \sum_{s=1}^{t-1} \Sigma_{q_s,S_s}(\theta) \\ &\succeq \kappa_{W,K} \sum_{s=1}^{t-1} \Sigma^{\mathrm{unw}}_{q_s,S_s} \\ &= \kappa_{W,K} U_{t-1}. \end{aligned}$$

Therefore,

$$H_t(\theta) \succeq \kappa_{W,K} U_{t-1}, \qquad \forall t \geq 1, \quad \theta \in \Theta.$$

Part II: Uniform Fisher information growth. By Proposition S1, there exists an event Ω_U of probability one on which

$$\liminf_{n\to\infty} \frac{\lambda_{\min}(U_n)}{n^\alpha} \geq c_X.$$

Fix a sample path in Ω_U. The uniform Fisher–proxy comparison established in Part I gives $\lambda_{\min}(H_t(\theta)) \geq \kappa_{W,K} \lambda_{\min}(U_{t-1})$ for every $t \geq 1$ and $\theta \in \Theta$. Taking the infimum over $\theta \in \Theta$ and dividing by t^α, we obtain, for every $t \geq 2$,

$$\frac{\inf_{\theta\in\Theta} \lambda_{\min}\big(H_t(\theta)\big)}{t^\alpha} \geq \kappa_{W,K} \frac{\lambda_{\min}(U_{t-1})}{(t-1)^\alpha} \left(\frac{t-1}{t}\right)^\alpha.$$

Since $((t-1)/t)^\alpha \to 1$ and the information matrices are positive semidefinite, taking lower limits yields

$$\liminf_{t\to\infty} \frac{\inf_{\theta\in\Theta} \lambda_{\min}\big(H_t(\theta)\big)}{t^\alpha} \geq \kappa_{W,K} c_X > 0.$$

In particular, on this sample path, there exists a finite time T such that $\inf_{\theta\in\Theta} \lambda_{\min}(H_t(\theta)) \geq \frac{1}{2}\kappa_{W,K} c_X t^\alpha > 0$ for every $t \geq T$. Since $\mathbb{P}(\Omega_U) = 1$, the asserted growth bound holds almost surely, and the cumulative Fisher information is uniformly positive definite over Θ for all sufficiently large t, almost surely. □

C.1.2 Self-Concordant Hessian Comparison

Self-concordant-like functions. Following Tran-Dinh et al. (2015), a convex function $f \in C^3(\mathbb{R}^d)$ is said to be M_f-self-concordant-like, for some $M_f > 0$, if, for every $\theta, u \in \mathbb{R}^d$, the univariate function $\phi(s) := f(\theta + su)$ satisfies $|\phi'''(s)| \leq M_f \|u\|_2 \, \phi''(s)$ for all $s \in \mathbb{R}$.

The following Hessian comparison is the main generic consequence of self-concordance used in our analysis.

Proposition S2 (Generic Hessian comparison Tran-Dinh et al., 2015, Theorem 4)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an M_f-self-concordant-like function. Then, for every $\theta_1, \theta_2 \in \mathrm{dom}(f)$,*

$$e^{-M_f\|\theta_2-\theta_1\|_2} \nabla^2 f(\theta_1) \preceq \nabla^2 f(\theta_2).$$

Interchanging θ_1 and θ_2 in Proposition S2 yields the two-sided comparison

$$e^{-M_f\|\theta_2-\theta_1\|_2} \nabla^2 f(\theta_1) \preceq \nabla^2 f(\theta_2) \preceq e^{M_f\|\theta_2-\theta_1\|_2} \nabla^2 f(\theta_1).$$

Model-specific loss and Hessian. For a task $j \in [J]$, an assortment $S \in \mathcal{S}$, and a realized winner $y \in S$, write the single-round negative log-likelihood as $\ell_{j,S,y}(\theta) := -x_{j,y}^\top \theta + A_{j,S}(\theta)$, where $A_{j,S}(\theta) := \log\big(\sum_{i\in S} \exp(x_{j,i}^\top \theta)\big)$. Since the observation-dependent term is linear, $\nabla^2 \ell_{j,S,y}(\theta) = \nabla^2 A_{j,S}(\theta) = \Sigma_{j,S}(\theta)$. Consequently, for the cumulative loss $L_t(\theta) := \sum_{s=1}^{t-1} \ell_{q_s,S_s,Y_s}(\theta)$, we have

$$\nabla^2 L_t(\theta) = \sum_{s=1}^{t-1} \Sigma_{q_s,S_s}(\theta) = H_t(\theta).$$

A standard log-sum-exp argument yields the classical $\sqrt{6K}$ self-concordance constant. For the no-outside-option model considered here, however, the following diameter-based argument gives the sharper constant used in the subsequent analysis.

For a task–assortment pair (j, S), define its within-assortment feature diameter by $D_{j,S} := \max_{i,\ell \in S} \|x_{j,i} - x_{j,\ell}\|_2$.

Lemma S2 (Refined diameter-based self-concordance bound)**.** *For every $j \in [J]$, $S \in \mathcal{S}$, and $y \in S$, the single-round loss $\ell_{j,S,y}$ is $D_{j,S}$-self-concordant-like. Namely, for every $\theta, u \in \mathbb{R}^d$, the function $\phi(s) := \ell_{j,S,y}(\theta + su)$ satisfies*

$$|\phi'''(s)| \le D_{j,S}\|u\|_2\, \phi''(s), \qquad \forall s \in \mathbb{R}.$$

In particular, under $\|x_{j,i}\|_2 \le 1$, we have $D_{j,S} \le 2$, so every single-round loss is also 2-self-concordant-like.

Proof. Fix $j \in [J]$, $S \in \mathcal{S}$, $y \in S$, and $\theta, u \in \mathbb{R}^d$. Since the term $-x_{j,y}^\top \theta$ is linear, it contributes neither to the second nor to the third directional derivative. It therefore suffices to consider $\psi(s) := A_{j,S}(\theta + su) = \log\big(\sum_{i \in S} \exp(a_i s + b_i)\big)$, where $a_i := x_{j,i}^\top u$ and $b_i := x_{j,i}^\top \theta$.

For each $s \in \mathbb{R}$, let Z_s be the discrete random variable defined by $\mathbb{P}(Z_s = a_i) = \rho_{j,S,i}(\theta + su)$ for $i \in S$. The derivatives of the log-partition function are the corresponding cumulants: $\psi'(s) = \mathbb{E}[Z_s]$, $\psi''(s) = \mathrm{Var}(Z_s)$, and $\psi'''(s) = \mathbb{E}[(Z_s - \mathbb{E}[Z_s])^3]$. Therefore,

$$\begin{aligned}
|\psi'''(s)| &\le \mathbb{E}\big[|Z_s - \mathbb{E}[Z_s]|^3\big] \\
&\le \left(\max_{i \in S} |a_i - \mathbb{E}[Z_s]|\right) \mathbb{E}\big[(Z_s - \mathbb{E}[Z_s])^2\big] \\
&= \left(\max_{i \in S} |a_i - \mathbb{E}[Z_s]|\right) \psi''(s).
\end{aligned}$$

Since $\mathbb{E}[Z_s]$ is a convex combination of $\{a_i : i \in S\}$, it lies between their minimum and maximum, so $\max_{i \in S} |a_i - \mathbb{E}[Z_s]| \le \max_{i,\ell \in S} |a_i - a_\ell|$. Moreover, $\max_{i,\ell \in S} |a_i - a_\ell| = \max_{i,\ell \in S} |u^\top (x_{j,i} - x_{j,\ell})| \le D_{j,S}\|u\|_2$. Hence

$$|\psi'''(s)| \le D_{j,S}\|u\|_2\, \psi''(s).$$

Since the linear observation term has zero second and third derivatives, the same inequality holds for $\phi(s) = \ell_{j,S,y}(\theta + su)$.

Finally, under the bounded-feature condition, $D_{j,S} \le \max_{i,\ell \in S}(\|x_{j,i}\|_2 + \|x_{j,\ell}\|_2) \le 2$. This proves the result. □

Cumulative self-concordance and Hessian comparison. Recall $D_X := \max_{j \in [J],\, S \in \mathcal{S}} D_{j,S} \le 2$ under the bounded-feature condition.

Proposition S3 (Self-concordance of the cumulative loss)**.** *For every $t \ge 1$, the cumulative loss $L_t(\theta) = \sum_{\tau=1}^{t-1} \ell_{q_\tau, S_\tau, Y_\tau}(\theta)$ is D_X-self-concordant-like. Namely, for every $\theta, u \in \mathbb{R}^d$, the function $\Phi_t(r) := L_t(\theta + ru)$ satisfies*

$$|\Phi_t'''(r)| \le D_X\|u\|_2\, \Phi_t''(r), \qquad \forall r \in \mathbb{R}.$$

Proof. Fix $t \ge 1$ and $\theta, u \in \mathbb{R}^d$. For each $\tau \in \{1, \dots, t-1\}$, define $\phi_\tau(r) := \ell_{q_\tau, S_\tau, Y_\tau}(\theta + ru)$, so that $\Phi_t(r) = \sum_{\tau=1}^{t-1} \phi_\tau(r)$. By Lemma S2, $|\phi_\tau'''(r)| \le D_{q_\tau, S_\tau}\|u\|_2\, \phi_\tau''(r) \le D_X\|u\|_2\, \phi_\tau''(r)$. Moreover, each single-round loss is convex, so $\phi_\tau''(r) \ge 0$. Therefore,

$$\begin{aligned}
|\Phi_t'''(r)| &= \left|\sum_{\tau=1}^{t-1} \phi_\tau'''(r)\right| \\
&\le \sum_{\tau=1}^{t-1} |\phi_\tau'''(r)| \\
&\le D_X\|u\|_2 \sum_{\tau=1}^{t-1} \phi_\tau''(r) \\
&= D_X\|u\|_2\, \Phi_t''(r).
\end{aligned}$$

Thus, L_t is D_X-self-concordant-like. □

Lemma S3 (Self-concordant Hessian comparison)**.** *For every $t \geq 1$ and $\theta_1, \theta_2 \in \mathbb{R}^d$,*

$$e^{-D_X\|\theta_2-\theta_1\|_2} H_t(\theta_1) \preceq H_t(\theta_2) \preceq e^{D_X\|\theta_2-\theta_1\|_2} H_t(\theta_1).$$

Proof. By Proposition S3, the cumulative loss L_t is D_X-self-concordant-like. Applying Proposition S2 in both directions gives

$$e^{-D_X\|\theta_2-\theta_1\|_2} \nabla^2 L_t(\theta_1) \preceq \nabla^2 L_t(\theta_2) \preceq e^{D_X\|\theta_2-\theta_1\|_2} \nabla^2 L_t(\theta_1).$$

Since $\nabla^2 L_t(\theta) = H_t(\theta)$, the claimed comparison follows. □

Corollary S1 (Uniform Hessian comparison over the parameter domain)**.** *Suppose that $\Theta = B_2(0, W)$ and $\|x_{j,i}\|_2 \leq 1$ for all $j \in [J]$ and $i \in [N]$. Then, for every $t \geq 1$ and $\theta_1, \theta_2 \in \Theta$,*

$$e^{-4W} H_t(\theta_1) \preceq H_t(\theta_2) \preceq e^{4W} H_t(\theta_1).$$

Corollary S1 provides a uniform multiplicative comparison between Fisher matrices evaluated at any two parameters in Θ.

C.1.3 Self-Normalized Score Concentration

We establish a uniform-in-time self-normalized bound for the score at θ^*, which provides the stochastic input to the constrained-MLE confidence bound.

Lemma S4 (Self-normalized score concentration)**.** *Let $\delta \in (0, 1)$ and fix $\lambda > 0$. Suppose that $\|x_{j,i}\|_2 \leq 1$ for all tasks $j \in [J]$ and items $i \in [N]$. Define the score at the true parameter by*

$$Z_t := -\nabla L_t(\theta^*) = \sum_{s=1}^{t-1} \left[x_{q_s,Y_s} - \sum_{i \in S_s} \rho_{q_s,S_s,i}(\theta^*) x_{q_s,i} \right].$$

Then, with probability at least $1 - \delta$, for all $t \geq 1$ such that $\lambda_{\min}(H_t(\theta^*)) \geq \lambda$,

$$\|Z_t\|_{H_t(\theta^*)^{-1}} \leq \gamma_t^\lambda(\delta) := \frac{\sqrt{\lambda}}{2} + \frac{4}{\sqrt{\lambda}} \log\left[\frac{2^{3d/2}}{\delta} \left(1 + \frac{2tK}{\lambda d}\right)^{d/2} \right].$$

Proof. Differentiating the MNL negative log-likelihood at θ^* gives $Z_t = \sum_{s=1}^{t-1} z_s$, where

$$z_s := x_{q_s,Y_s} - \sum_{i \in S_s} \rho_{q_s,S_s,i}(\theta^*) x_{q_s,i}.$$

Define the pre-feedback filtration

$$\widetilde{\mathcal{F}}_t := \sigma\big((q_s, S_s, Y_s)_{s=1}^{t-1}, q_t, S_t\big),$$

which contains the past observations and the current task-assortment pair, but not the current winner Y_t. Conditional on $\widetilde{\mathcal{F}}_s$, the winner Y_s follows the contextual MNL distribution on S_s. Hence $\mathbb{E}[z_s \mid \widetilde{\mathcal{F}}_s] = 0$ and

$$\mathbb{E}\Big[z_s z_s^\top \mid \widetilde{\mathcal{F}}_s \Big] = \Sigma_{q_s,S_s}(\theta^*).$$

Writing $\Sigma_s^* := \Sigma_{q_s,S_s}(\theta^*)$, we therefore have $H_t(\theta^*) = \sum_{s=1}^{t-1} \Sigma_s^*$.

For any fixed $\xi \in B(0, 1/2)$, define $M_1(\xi) := 1$ and, for $t \geq 2$,

$$M_t(\xi) := \exp\Big\{ \xi^\top Z_t - \|\xi\|^2_{H_t(\theta^*)} \Big\}.$$

We show that $\{M_t(\xi)\}_{t \geq 1}$ is a nonnegative supermartingale with respect to $\{\widetilde{\mathcal{F}}_t\}_{t \geq 1}$. Since $\|x_{j,i}\|_2 \leq 1$, we have $\|z_t\|_2 \leq 2$, and hence $|\xi^\top z_t| \leq 1$ for every $\xi \in B(0, 1/2)$. Using $e^x \leq 1 + x + x^2$ for $|x| \leq 1$ together with the conditional mean and covariance identities above gives

$$\mathbb{E}\Big[e^{\xi^\top z_t} \mid \widetilde{\mathcal{F}}_t \Big] \leq 1 + \xi^\top \Sigma_t^* \xi \leq e^{\xi^\top \Sigma_t^* \xi}.$$

Because $Z_{t+1} = Z_t + z_t$ and $H_{t+1}(\theta^*) = H_t(\theta^*) + \Sigma_t^*$, it follows that $\mathbb{E}[M_{t+1}(\xi) \mid \widetilde{\mathcal{F}}_t] \le M_t(\xi)$.

Fix $\lambda > 0$. Let h be the density of the centered isotropic normal distribution with precision $2\lambda I_d$, truncated on $B(0, 1/2)$:

$$h(\xi) = \frac{e^{-\lambda\|\xi\|_2^2}\mathbf{1}\{\|\xi\|_2 \le 1/2\}}{N(h)}, \qquad N(h) := \int_{\|\xi\|_2 \le 1/2} e^{-\lambda\|\xi\|_2^2}\, d\xi.$$

Define the mixture process $\overline{M}_t := \int_{B(0,1/2)} M_t(\xi)h(\xi)\, d\xi$. By Tonelli's theorem, $\{\overline{M}_t\}_{t\ge1}$ is again a nonnegative supermartingale and $\overline{M}_1 = 1$. Therefore, Ville's maximal inequality yields

$$\mathbb{P}\left(\exists t \ge 1 : \overline{M}_t \ge \frac{1}{\delta}\right) \le \delta.$$

Thus, with probability at least $1 - \delta$, $\log \overline{M}_t < \log(1/\delta)$ simultaneously for all $t \ge 1$.

Fix a time t on this event and define $A_t := H_t(\theta^*) + \lambda I_d$ and $f_t(\xi) := \xi^\top Z_t - \xi^\top A_t \xi$. Let $\xi_* \in \arg\max_{\|\xi\|_2 \le 1/4} f_t(\xi)$. Since f_t is quadratic with Hessian $-2A_t$,

$$f_t(\xi) = f_t(\xi_*) + (\xi - \xi_*)^\top \nabla f_t(\xi_*) - (\xi - \xi_*)^\top A_t (\xi - \xi_*).$$

After the change of variables $\eta = \xi - \xi_*$, the inclusion $B(0, 1/4) \subseteq \{\eta : \|\eta + \xi_*\|_2 \le 1/2\}$ allows us to restrict the integral defining $\overline{M}_t$ to $B(0, 1/4)$.

Let φ_t denote the centered normal distribution with precision $2A_t$, truncated on $B(0, 1/4)$, and let $N(\varphi_t)$ denote its normalizing constant. Rewriting the restricted integral under φ_t and applying Jensen's inequality gives

$$\overline{M}_t \ge e^{f_t(\xi_*)} \frac{N(\varphi_t)}{N(h)}.$$

Indeed, the linear term disappears because φ_t is centered and truncated on the symmetric ball $B(0, 1/4)$, so $\mathbb{E}_{\varphi_t}[\eta] = 0$. Consequently, on the same high-probability event,

$$\max_{\|\xi\|_2 \le 1/4} \left(\xi^\top Z_t - \|\xi\|_{A_t}^2\right) < \log(1/\delta) + \log\left(\frac{N(h)}{N(\varphi_t)}\right).$$

We use the following direct adaptation of Lemma 6 in Faury et al. (2020); the proof differs only through the truncation radii and precision parameters and is therefore omitted.

Lemma S5 (Ratio of truncated Gaussian normalizing constants)**.** *Fix $t \ge 1$ and $\lambda > 0$, and let $A_t := H_t(\theta^*) + \lambda I_d$. Let h be the centered isotropic normal distribution with precision $2\lambda I_d$, truncated on $B(0, 1/2)$, and let φ_t be the centered normal distribution with precision $2A_t$, truncated on $B(0, 1/4)$. With $N(h) := \int_{\|\xi\|_2 \le 1/2} e^{-\lambda\|\xi\|_2^2}\, d\xi$ and $N(\varphi_t) := \int_{\|\eta\|_2 \le 1/4} e^{-\eta^\top A_t \eta}\, d\eta$, we have*

$$\log\left(\frac{N(h)}{N(\varphi_t)}\right) \le \log\left(\frac{\det(A_t)^{1/2}}{\lambda^{d/2}}\right) + d\log 2.$$

Now suppose that $\lambda_{\min}(H_t(\theta^*)) \ge \lambda$, i.e., $H_t(\theta^*) \succeq \lambda I_d$. If $Z_t = 0$, the claimed bound is immediate. Otherwise, choose

$$\xi_0 := \frac{\sqrt{\lambda}}{4} \frac{H_t(\theta^*)^{-1} Z_t}{\|Z_t\|_{H_t(\theta^*)^{-1}}}.$$

Since $H_t(\theta^*) \succeq \lambda I_d$, we have $\|\xi_0\|_2 \le 1/4$, so ξ_0 is feasible in the preceding maximum. Moreover,

$$\xi_0^\top Z_t = \frac{\sqrt{\lambda}}{4}\|Z_t\|_{H_t(\theta^*)^{-1}}, \qquad \|\xi_0\|_{A_t}^2 \le \frac{\lambda}{8}.$$

Hence

$$\|Z_t\|_{H_t(\theta^*)^{-1}} < \frac{\sqrt{\lambda}}{2} + \frac{4}{\sqrt{\lambda}}\left[\log(1/\delta) + \log\left(\frac{N(h)}{N(\varphi_t)}\right)\right].$$

Applying Lemma S5 therefore gives

$$\|Z_t\|_{H_t(\theta^*)^{-1}} < \frac{\sqrt{\lambda}}{2} + \frac{4}{\sqrt{\lambda}} \log\left(\frac{2^d}{\delta}\frac{\det(A_t)^{1/2}}{\lambda^{d/2}}\right).$$

Since $H_t(\theta^*) \succeq \lambda I_d$, $A_t \preceq 2H_t(\theta^*)$, and hence $\det(A_t)^{1/2} \leq 2^{d/2}\det(H_t(\theta^*))^{1/2}$. By the determinant–trace inequality and the bounded-feature condition,

$$\det(H_t(\theta^*))^{1/2} \leq \left(\frac{\mathrm{tr}(H_t(\theta^*))}{d}\right)^{d/2} \leq \left(\lambda + \frac{2tK}{d}\right)^{d/2}.$$

Substituting these bounds into the preceding inequality yields

$$\|Z_t\|_{H_t(\theta^*)^{-1}} < \frac{\sqrt{\lambda}}{2} + \frac{4}{\sqrt{\lambda}}\log\left[\frac{2^{3d/2}}{\delta}\left(1 + \frac{2tK}{\lambda d}\right)^{d/2}\right] = \gamma_t^\lambda(\delta).$$

This completes the proof. □

C.1.4 Proof of Proposition 1: Constrained-MLE Confidence

The proof combines the uniform-in-time score bound from Lemma S4 with the distance-sensitive Hessian comparison in Lemma S3 along the segment joining θ^* and $\hat{\theta}_t$.

Proof. Fix $\lambda > 0$. By Lemma S4, there exists an event of probability at least $1 - \delta$ on which, simultaneously for every t satisfying $\lambda_{\min}(H_t(\theta^*)) \geq \lambda$,

$$\|Z_t\|_{H_t(\theta^*)^{-1}} \leq \gamma_t^\lambda(\delta).$$

We work on this event and fix any such t.

Write $h_t := \hat{\theta}_t - \theta^*$ and $\overline{H}_t := \int_0^1 H_t(\theta^* + sh_t)\,ds$. Since Θ is convex and $\hat{\theta}_t$ minimizes L_t over Θ, the first-order optimality condition gives $\nabla L_t(\hat{\theta}_t)^\top h_t \leq 0$. By the fundamental theorem of calculus, $\nabla L_t(\hat{\theta}_t) - \nabla L_t(\theta^*) = \overline{H}_t h_t$. Since $Z_t = -\nabla L_t(\theta^*)$, we therefore have $\nabla L_t(\hat{\theta}_t) = -Z_t + \overline{H}_t h_t$, and hence

$$h_t^\top \overline{H}_t h_t \leq Z_t^\top h_t.$$

Because $\theta^*, \hat{\theta}_t \in \Theta$ and Θ is convex, the entire segment $\{\theta^* + sh_t : s \in [0,1]\}$ lies in Θ. Moreover, $\|h_t\|_2 \leq 2W$. Thus, Lemma S3 implies, for every $s \in [0,1]$,

$$H_t(\theta^* + sh_t) \succeq e^{-D_X s\|h_t\|_2} H_t(\theta^*) \succeq e^{-2WD_X} H_t(\theta^*).$$

Integrating over s gives $\overline{H}_t \succeq e^{-2WD_X} H_t(\theta^*)$. Consequently,

$$e^{-2WD_X}\|h_t\|^2_{H_t(\theta^*)} \leq Z_t^\top h_t \leq \|Z_t\|_{H_t(\theta^*)^{-1}}\|h_t\|_{H_t(\theta^*)},$$

where the last inequality is Cauchy–Schwarz in the $H_t(\theta^*)$ metric. If $h_t = 0$, the desired bound is immediate. Otherwise, dividing by $\|h_t\|_{H_t(\theta^*)}$ yields

$$\|h_t\|_{H_t(\theta^*)} \leq e^{2WD_X}\|Z_t\|_{H_t(\theta^*)^{-1}}.$$

Applying Lemma S3 once more at the two endpoints gives $H_t(\hat{\theta}_t) \preceq e^{D_X\|h_t\|_2} H_t(\theta^*) \preceq e^{2WD_X} H_t(\theta^*)$. Hence

$$\|h_t\|_{H_t(\hat{\theta}_t)} \leq e^{WD_X}\|h_t\|_{H_t(\theta^*)} \leq e^{3WD_X}\|Z_t\|_{H_t(\theta^*)^{-1}} \leq e^{3WD_X}\gamma_t^\lambda(\delta).$$

Since t was arbitrary on the same probability-$1-\delta$ event, the bound holds simultaneously for every t satisfying $\lambda_{\min}(H_t(\theta^*)) \geq \lambda$, as claimed. □

C.1.5 Almost-sure consistency

Corollary S2 (Almost-sure consistency)**.** *Under the forced-exploration safeguard of Proposition S1 with fixed $\alpha \in (0,1)$, $\hat{\theta}_t \to \theta^*$ almost surely.*

Proof. By Lemma S1, there exists an event Ω_H with $\mathbb{P}(\Omega_H) = 1$ on which

$$\liminf_{t\to\infty} \frac{\inf_{\theta\in\Theta} \lambda_{\min}\big(H_t(\theta)\big)}{t^\alpha} \geq \kappa_{W,K} c_X > 0.$$

Consequently, on every sample path in Ω_H, there exists a finite, sample-path-dependent time T such that

$$\inf_{\theta\in\Theta} \lambda_{\min}\big(H_t(\theta)\big) \geq \frac{1}{2}\kappa_{W,K} c_X t^\alpha, \qquad t \geq T.$$

Since $\theta^* \in \Theta$ and $\alpha > 0$, this also implies $\lambda_{\min}(H_t(\theta^*)) \to \infty$ on Ω_H.

Fix any $\lambda_0 > 0$. On every sample path in Ω_H, we may therefore choose a finite, sample-path-dependent time $t_0 \geq T$ such that $\lambda_{\min}(H_t(\theta^*)) \geq \lambda_0$ for every $t \geq t_0$. For any fixed $\delta \in (0,1)$, define

$$E_\delta := \left\{ \begin{matrix} \text{for every } t \geq 1 \text{ such that } \lambda_{\min}(H_t(\theta^*)) \geq \lambda_0, \\ \|\hat{\theta}_t - \theta^*\|_{H_t(\hat{\theta}_t)} \leq e^{3WD_X} \gamma_t^{\lambda_0}(\delta) \end{matrix} \right\}.$$

By Proposition 1, $\mathbb{P}(E_\delta) \geq 1 - \delta$.

Fix a sample path in $E_\delta \cap \Omega_H$. For every $t \geq t_0$, the true information satisfies the threshold required in the definition of E_δ. Moreover, since $\hat{\theta}_t \in \Theta$ and $t_0 \geq T$, the uniform information lower bound applies to $H_t(\hat{\theta}_t)$. Hence

$$\begin{aligned} \|\hat{\theta}_t - \theta^*\|_2 &\leq \frac{\|\hat{\theta}_t - \theta^*\|_{H_t(\hat{\theta}_t)}}{\sqrt{\lambda_{\min}(H_t(\hat{\theta}_t))}} \\ &\leq \frac{\sqrt{2}\, e^{3WD_X} \gamma_t^{\lambda_0}(\delta)}{\sqrt{\kappa_{W,K} c_X}\, t^{\alpha/2}}. \end{aligned}$$

For fixed $\lambda_0 > 0$ and $\delta \in (0,1)$, $\gamma_t^{\lambda_0}(\delta) = O(\log t)$. Because $\alpha > 0$, we have $(\log t)/t^{\alpha/2} \to 0$. Thus every sample path in $E_\delta \cap \Omega_H$ satisfies $\hat{\theta}_t \to \theta^*$.

Define $A := \{\hat{\theta}_t \to \theta^*\}$. The preceding argument shows that $E_\delta \cap \Omega_H \subseteq A$. Since $\mathbb{P}(\Omega_H) = 1$, it follows that

$$\mathbb{P}(A) \geq \mathbb{P}(E_\delta \cap \Omega_H) = \mathbb{P}(E_\delta) \geq 1 - \delta.$$

This holds for every $\delta \in (0,1)$, so $\mathbb{P}(A) = 1$, proving the claimed almost-sure consistency. □

C.2 Design Stability and Task-Aware Tracking

C.2.1 Stage I: Separation and Joint Continuity

Alternative set and separation. Recall that $\Theta := B_2(0, W)$ and $\bar{x}_i := \sum_{j=1}^J p_j x_{j,i}$. Recall that the alternative domain $\mathcal{D}$ is fixed and satisfies $\Theta \subseteq \mathcal{D} \subseteq \mathbb{R}^d$. For $\theta \in \Theta$ with a unique best model, write $i^*(\theta) := \arg\max_{i\in[N]} \bar{x}_i^\top \theta$ and recall the strict alternative set

$$\mathrm{Alt}_{\mathcal{D}}(\theta) := \left\{ \lambda \in \mathcal{D} : \exists\, i \neq i^*(\theta) \text{ such that } \bar{x}_i^\top \lambda > \bar{x}_{i^*(\theta)}^\top \lambda \right\}.$$

The strict inequality avoids ambiguity at decision ties.

Let $\mathcal{W} := \Delta(\mathcal{S})^J$ denote the product allocation simplex. For $w = (w_1, \ldots, w_J) \in \mathcal{W}$, recall $V(w;\theta) := \sum_{j=1}^J p_j \sum_{S\in\mathcal{S}} w_{j,S} \Sigma_{j,S}(\theta)$. For $\theta, \lambda \in \mathbb{R}^d$ and $w \in \mathcal{W}$, define $f(\theta, \lambda, w) := \frac{1}{2}(\theta - \lambda)^\top V(w;\theta)(\theta - \lambda)$. The separation is

$$\psi_{\mathcal{D}}(\theta, w) := \inf_{\lambda \in \mathrm{Alt}_{\mathcal{D}}(\theta)} f(\theta, \lambda, w).$$

The joint-continuity argument below uses local stability of the alternative set and a multiplicative comparison of Fisher information matrices derived from their finite fixed-direction decomposition.

Local stability of the best-model identity. Define the regular parameter set $\mathcal{R}_\Theta := \{\theta \in \Theta : i^*(\theta) \text{ is unique}\}$. Fix $\theta \in \mathcal{R}_\Theta$, write $i^* := i^*(\theta)$ and $d_i := \bar{x}_{i^*} - \bar{x}_i$ for $i \neq i^*$. Since i^* is uniquely optimal, $\Delta_i(\theta) := d_i^\top \theta > 0$ for every $i \neq i^*$. Define the distance to the nearest decision boundary by

$$\rho_\theta := \min_{i \neq i^*} \frac{\Delta_i(\theta)}{\|d_i\|_2} = \min_{i \neq i^*} \frac{d_i^\top \theta}{\|d_i\|_2} > 0.$$

Lemma S6 (Local stability of the best-model identity and alternatives)**.** *Fix $\theta \in \mathcal{R}_\Theta$, and let ρ_θ be defined as above. Then every $\theta' \in \Theta$ satisfying $\|\theta' - \theta\|_2 < \rho_\theta$ has the same unique best model:*

$$i^*(\theta') = i^*(\theta).$$

Consequently, $\mathrm{Alt}_\mathcal{D}(\theta') = \mathrm{Alt}_\mathcal{D}(\theta)$.

Proof. Write $i^* := i^*(\theta)$ and $d_i := \bar{x}_{i^*} - \bar{x}_i$ for $i \neq i^*$. For any $\theta' \in \Theta$ with $\|\theta' - \theta\|_2 < \rho_\theta$,

$$d_i^\top \theta' \geq d_i^\top \theta - \|d_i\|_2 \|\theta' - \theta\|_2 > 0, \qquad i \neq i^*,$$

where the strict inequality follows from the definition of ρ_θ. Hence $\bar{x}_{i^*}^\top \theta' > \bar{x}_i^\top \theta'$ for every $i \neq i^*$, so $i^*(\theta') = i^*(\theta)$ and the best model remains unique.

Since $\mathrm{Alt}_\mathcal{D}(\cdot)$ depends on its reference parameter only through the identity of the unique best model, $\mathrm{Alt}_\mathcal{D}(\theta') = \mathrm{Alt}_\mathcal{D}(\theta)$. □

Define the relative neighborhood $U_\theta := \{\theta' \in \Theta : \|\theta' - \theta\|_2 < \rho_\theta\}$. Lemma S6 implies $U_\theta \subseteq \mathcal{R}_\Theta$ and, for every $\theta' \in U_\theta$, $i^*(\theta') = i^*(\theta)$ and $\mathrm{Alt}_\mathcal{D}(\theta') = \mathrm{Alt}_\mathcal{D}(\theta)$. In particular, $\mathcal{R}_\Theta$ is relatively open in Θ. Thus, locally around every regular parameter, the alternative set is fixed.

Lemma S7 (Joint continuity of the separation)**.** *Let $\mathcal{D}$ be a fixed set satisfying $\Theta \subseteq \mathcal{D} \subseteq \mathbb{R}^d$. For $\theta \in \mathcal{R}_\Theta$, define*

$$\mathrm{Alt}_\mathcal{D}(\theta) := \left\{\lambda \in \mathcal{D} : \exists i \neq i^*(\theta),\ (\bar{x}_{i^*(\theta)} - \bar{x}_i)^\top \lambda < 0\right\} = \mathcal{D} \cap \mathrm{Alt}_\infty(\theta).$$

The separation

$$\psi_\mathcal{D} : \mathcal{R}_\Theta \times \mathcal{W} \longrightarrow \mathbb{R}_{\geq 0}, \qquad \psi_\mathcal{D}(\theta, w) := \frac{1}{2} \inf_{\lambda \in \mathrm{Alt}_\mathcal{D}(\theta)} \|\theta - \lambda\|^2_{V(w;\theta)},$$

is jointly continuous, where $\mathcal{R}_\Theta$ is endowed with the relative topology inherited from Θ.

Proof. Fix $(\theta, w) \in \mathcal{R}_\Theta \times \mathcal{W}$ and a sequence $(\theta_n, w_n) \to (\theta, w)$ in this domain. Write

$$i^* := i^*(\theta), \qquad V_n := V(w_n; \theta_n), \qquad V := V(w; \theta).$$

Since i^* is uniquely optimal,

$$(\bar{x}_{i^*} - \bar{x}_i)^\top \theta > 0, \qquad i \neq i^*.$$

There are finitely many competitors, so $\theta_n \to \theta$ implies that all these inequalities remain strict at θ_n for sufficiently large n. Hence $i^*(\theta_n) = i^*$ eventually, and, since $\mathcal{D}$ is fixed, the alternative set is locally fixed:

$$\mathrm{Alt}_\mathcal{D}(\theta_n) = \mathrm{Alt}_\mathcal{D}(\theta) = A := \mathcal{D} \cap \bigcup_{i \neq i^*} \left\{\lambda \in \mathbb{R}^d : (\bar{x}_{i^*} - \bar{x}_i)^\top \lambda < 0\right\}.$$

This set is nonempty: since $\Theta = B_2(0, W)$, we have $-\theta \in \Theta \subseteq \mathcal{D}$, and for any $i \neq i^*$,

$$(\bar{x}_{i^*} - \bar{x}_i)^\top (-\theta) = -(\bar{x}_{i^*} - \bar{x}_i)^\top \theta < 0.$$

Thus $-\theta \in A$. In particular, the separation is finite and nonnegative. We henceforth consider only sufficiently large n for which the alternative set equals A.

We first establish a multiplicative lower bound for V_n. The single-comparison Fisher matrix admits the representation

$$\begin{aligned}\Sigma_{j,S}(\vartheta) &= \sum_{i \in S} \rho_{j,S,i}(\vartheta) x_{j,i} x_{j,i}^\top - \mu_{j,S}(\vartheta) \mu_{j,S}(\vartheta)^\top \\ &= \sum_{\substack{i,k \in S \\ i<k}} \rho_{j,S,i}(\vartheta) \rho_{j,S,k}(\vartheta) (x_{j,i} - x_{j,k})(x_{j,i} - x_{j,k})^\top.\end{aligned}$$

Consequently, for the finite index set

$$\mathcal{I} := \{(j, S, i, k) : j \in [J],\ S \in \mathcal{S},\ i, k \in S,\ i < k\},$$

we may write

$$V(v; \vartheta) = \sum_{r \in \mathcal{I}} a_r(\vartheta, v) b_r b_r^\top,$$

where, for $r = (j, S, i, k)$,

$$b_r := x_{j,i} - x_{j,k}, \qquad a_r(\vartheta, v) := p_j v_{j,S} \rho_{j,S,i}(\vartheta) \rho_{j,S,k}(\vartheta).$$

Each b_r is fixed, and each coefficient a_r is continuous and nonnegative.

Set $a_{r,n} := a_r(\theta_n, w_n)$ and $a_r := a_r(\theta, w)$, and define

$$c_n := \min\left(\{1\} \cup \left\{\frac{a_{r,n}}{a_r} : r \in \mathcal{I},\ a_r > 0\right\}\right).$$

Then $0 \le c_n \le 1$ and $c_n \to 1$, because $\mathcal{I}$ is finite and $a_{r,n}/a_r \to 1$ whenever $a_r > 0$. Moreover, $a_{r,n} \ge c_n a_r$ for every r: this follows from the definition when $a_r > 0$ and from nonnegativity when $a_r = 0$. Thus

$$V_n - c_n V = \sum_{r \in \mathcal{I}} (a_{r,n} - c_n a_r) b_r b_r^\top \succeq 0.$$

The same finite-sum representation also gives $V_n \to V$.

For a positive semidefinite matrix M, define

$$d_M(z, A) := \inf_{\lambda \in A} \|z - \lambda\|_M, \qquad \|u\|_M := \sqrt{u^\top M u}.$$

These distances are finite because A is nonempty. The inequality $V_n \succeq c_n V$ yields

$$d_{V_n}(\theta_n, A) \ge \sqrt{c_n}\, d_V(\theta_n, A).$$

Furthermore, the triangle inequality for the seminorm $\|\cdot\|_V$ gives, for every $\lambda \in A$,

$$\|\theta_n - \lambda\|_V \ge \|\theta - \lambda\|_V - \|\theta_n - \theta\|_V.$$

Taking infima over $\lambda \in A$ and combining the preceding inequalities, we obtain

$$\begin{aligned}\sqrt{2\psi_{\mathcal{D}}(\theta_n, w_n)} &= d_{V_n}(\theta_n, A) \\ &\ge \sqrt{c_n}\left(d_V(\theta, A) - \|\theta_n - \theta\|_V\right).\end{aligned}$$

Since $c_n \to 1$ and $\|\theta_n - \theta\|_V \to 0$, taking lower limits and using nonnegativity gives

$$\liminf_{n \to \infty} \psi_{\mathcal{D}}(\theta_n, w_n) \ge \frac{1}{2} d_V(\theta, A)^2 = \psi_{\mathcal{D}}(\theta, w).$$

For the reverse inequality, fix $\eta > 0$. By the definition of the infimum, there exists $\lambda_\eta \in A$ such that

$$\frac{1}{2}\|\theta - \lambda_\eta\|_V^2 \le \psi_{\mathcal{D}}(\theta, w) + \eta.$$

The same λ_η is feasible for all sufficiently large n, so

$$\psi_{\mathcal{D}}(\theta_n, w_n) \le \frac{1}{2}\|\theta_n - \lambda_\eta\|_{V_n}^2.$$

Because λ_η is fixed, $\theta_n \to \theta$ and $V_n \to V$ imply

$$\limsup_{n \to \infty} \psi_{\mathcal{D}}(\theta_n, w_n) \le \frac{1}{2}\|\theta - \lambda_\eta\|_V^2 \le \psi_{\mathcal{D}}(\theta, w) + \eta.$$

Letting $\eta \downarrow 0$ and combining this with the lower-limit inequality yields

$$\psi_{\mathcal{D}}(\theta_n, w_n) \longrightarrow \psi_{\mathcal{D}}(\theta, w).$$

Since the point and the convergent sequence were arbitrary, $\psi_{\mathcal{D}}$ is jointly continuous on $\mathcal{R}_\Theta \times \mathcal{W}$. $\square$

C.2.2 Stage II: Stability of the Optimal Design

Optimal value and optimal-design correspondence. Recall that, for $\theta \in \mathcal{R}_\Theta$,

$$\psi^*_{\mathcal{D}}(\theta) := \max_{w \in \mathcal{W}} \psi_{\mathcal{D}}(\theta, w), \qquad C^*_{\mathcal{D}}(\theta) := \arg\max_{w \in \mathcal{W}} \psi_{\mathcal{D}}(\theta, w).$$

Regularity via Berge's maximum theorem. By Lemma S7, $\psi_{\mathcal{D}}$ is jointly continuous on $\mathcal{R}_\Theta \times \mathcal{W}$. Since $\mathcal{W}$ is fixed, nonempty, and compact, the outer optimization satisfies the hypotheses of Berge's maximum theorem.

Lemma S8 (Regularity of the optimal value and optimal-design correspondence)**.** *The following properties hold:*

1. *For every $\theta \in \mathcal{R}_\Theta$, $C^*_{\mathcal{D}}(\theta)$ is nonempty and compact.*
2. *The optimal value $\psi^*_{\mathcal{D}} : \mathcal{R}_\Theta \to \mathbb{R}_{\geq 0}$ is continuous.*
3. *The correspondence $C^*_{\mathcal{D}} : \mathcal{R}_\Theta \rightrightarrows \mathcal{W}$ is upper hemicontinuous with respect to the relative topology on $\mathcal{R}_\Theta$.*

Proof. Fix $\theta \in \mathcal{R}_\Theta$. Since $w \mapsto \psi_{\mathcal{D}}(\theta, w)$ is continuous on the nonempty compact set $\mathcal{W}$, the maximum is attained by the Weierstrass theorem. Hence $C^*_{\mathcal{D}}(\theta)$ is nonempty. Moreover, $C^*_{\mathcal{D}}(\theta) = \{w \in \mathcal{W} : \psi_{\mathcal{D}}(\theta, w) = \psi^*_{\mathcal{D}}(\theta)\}$ is closed in $\mathcal{W}$ and therefore compact. This proves (i).

For (ii), let $\theta_n \to \theta$ in $\mathcal{R}_\Theta$. Joint continuity of $\psi_{\mathcal{D}}$ and compactness of $\mathcal{W}$ imply

$$\delta_n := \sup_{w \in \mathcal{W}} |\psi_{\mathcal{D}}(\theta_n, w) - \psi_{\mathcal{D}}(\theta, w)| \longrightarrow 0.$$

Indeed, otherwise there would exist $\varepsilon > 0$, a subsequence θ_{n_k}, and $w_k \in \mathcal{W}$ such that the displayed difference is at least ε for every k. By compactness, a further subsequence satisfies $w_k \to \bar{w} \in \mathcal{W}$, contradicting joint continuity at $(\theta, \bar{w})$. Using $|\sup_w a(w) - \sup_w b(w)| \leq \sup_w |a(w) - b(w)|$, we obtain $|\psi^*_{\mathcal{D}}(\theta_n) - \psi^*_{\mathcal{D}}(\theta)| \leq \delta_n \to 0$.

For (iii), first consider sequential stability of the maximizers. Let $\theta_n \to \theta$, $w_n \in C^*_{\mathcal{D}}(\theta_n)$, and suppose that $w_{n_k} \to \bar{w}$ along a subsequence. Since $\mathcal{W}$ is closed, $\bar{w} \in \mathcal{W}$. Optimality gives $\psi_{\mathcal{D}}(\theta_{n_k}, w_{n_k}) = \psi^*_{\mathcal{D}}(\theta_{n_k})$. Joint continuity of $\psi_{\mathcal{D}}$ and continuity of $\psi^*_{\mathcal{D}}$ therefore yield $\psi_{\mathcal{D}}(\theta, \bar{w}) = \psi^*_{\mathcal{D}}(\theta)$, so $\bar{w} \in C^*_{\mathcal{D}}(\theta)$.

To obtain upper hemicontinuity, let $O \subseteq \mathcal{W}$ be open with $C^*_{\mathcal{D}}(\theta) \subseteq O$. If no relative neighborhood U of θ satisfied $C^*_{\mathcal{D}}(\theta') \subseteq O$ for every $\theta' \in U$, we could choose $\theta_n \to \theta$ and $w_n \in C^*_{\mathcal{D}}(\theta_n) \setminus O$. Compactness of $\mathcal{W}$ gives a subsequence $w_{n_k} \to \bar{w} \in \mathcal{W} \setminus O$, whereas the sequential stability just proved implies $\bar{w} \in C^*_{\mathcal{D}}(\theta) \subseteq O$, a contradiction. Thus $C^*_{\mathcal{D}}$ is upper hemicontinuous. □

Distance-to-set stability. The following consequence is the form used in the tracking argument below.

Corollary S3 (Distance-to-set stability of optimal designs)**.** *For every sequence $\theta_n \to \theta$ in $\mathcal{R}_\Theta$,*

$$\sup_{w \in C^*_{\mathcal{D}}(\theta_n)} \operatorname{dist}\left(w, C^*_{\mathcal{D}}(\theta)\right) \longrightarrow 0,$$

where $\operatorname{dist}(w, C) := \inf_{v \in C} \|w - v\|_2$.

Proof. Suppose the conclusion fails. Then there exist $\varepsilon > 0$, a sequence $\theta_n \to \theta$, and $w_n \in C^*_{\mathcal{D}}(\theta_n)$ such that, after passing to a subsequence if necessary, $\operatorname{dist}(w_n, C^*_{\mathcal{D}}(\theta)) \geq \varepsilon$ for every n. Compactness of $\mathcal{W}$ yields a further subsequence $w_{n_k} \to \bar{w}$. By the sequential stability established in Lemma S8, $\bar{w} \in C^*_{\mathcal{D}}(\theta)$. Hence

$$\operatorname{dist}\left(w_{n_k}, C^*_{\mathcal{D}}(\theta)\right) \leq \|w_{n_k} - \bar{w}\|_2 \longrightarrow 0,$$

a contradiction. □

Concavity and convexity. Convexity of the optimal-design set follows from concavity of $w \mapsto \psi_{\mathcal{D}}(\theta, w)$.

Lemma S9 (Concavity of the separation and convexity of optimal designs)**.** *For every $\theta \in \mathcal{R}_\Theta$:*

1. *$w \mapsto \psi_{\mathcal{D}}(\theta, w)$ is concave on $\mathcal{W}$;*
2. *$C^*_{\mathcal{D}}(\theta)$ is convex.*

Proof. Fix $\theta \in \mathcal{R}_\Theta$. Let $w^{(1)}, w^{(2)} \in \mathcal{W}$ and $\eta \in [0,1]$, and write $w^\eta := \eta w^{(1)} + (1-\eta)w^{(2)}$. Since $\mathcal{W}$ is convex, $w^\eta \in \mathcal{W}$. Because $V(w;\theta)$ is affine in w, $V(w^\eta;\theta) = \eta V(w^{(1)};\theta) + (1-\eta)V(w^{(2)};\theta)$. Thus, for every fixed $\lambda \in \mathrm{Alt}_{\mathcal{D}}(\theta)$, $f(\theta, \lambda, w^\eta) = \eta f(\theta, \lambda, w^{(1)}) + (1-\eta) f(\theta, \lambda, w^{(2)})$.

Since $\mathrm{Alt}_{\mathcal{D}}(\theta)$ does not depend on w, taking the infimum over the same set gives

$$\psi_{\mathcal{D}}(\theta, w^\eta) \geq \eta \psi_{\mathcal{D}}(\theta, w^{(1)}) + (1-\eta)\psi_{\mathcal{D}}(\theta, w^{(2)}),$$

proving (i).

For (ii), let $w^{(1)}, w^{(2)} \in C^*_{\mathcal{D}}(\theta)$. By concavity, $\psi_{\mathcal{D}}(\theta, w^\eta) \geq \psi^*_{\mathcal{D}}(\theta)$, while the definition of the optimal value gives $\psi_{\mathcal{D}}(\theta, w^\eta) \leq \psi^*_{\mathcal{D}}(\theta)$. Hence equality holds, so $w^\eta \in C^*_{\mathcal{D}}(\theta)$. Therefore $C^*_{\mathcal{D}}(\theta)$ is convex. □

C.2.3 Regularity of the Optimal Design

Proposition S4 (Regularity of the optimal design)**.** *The separation $\psi_{\mathcal{D}}$ is jointly continuous on $\mathcal{R}_\Theta \times \mathcal{W}$. Moreover, for every $\theta \in \mathcal{R}_\Theta$, $C^*_{\mathcal{D}}(\theta)$ is nonempty, compact, and convex, and the correspondence $\theta \mapsto C^*_{\mathcal{D}}(\theta)$ is upper hemicontinuous.*

Proof. Lemma S7 establishes that $(\theta, w) \mapsto \psi_{\mathcal{D}}(\theta, w)$ is jointly continuous on $\mathcal{R}_\Theta \times \mathcal{W}$.

Fix any $\theta \in \mathcal{R}_\Theta$. By Lemma S8, $C^*_{\mathcal{D}}(\theta)$ is nonempty and compact, and the correspondence $\theta \mapsto C^*_{\mathcal{D}}(\theta)$ is upper hemicontinuous on $\mathcal{R}_\Theta$. By Lemma S9, $C^*_{\mathcal{D}}(\theta)$ is also convex.

Since θ was arbitrary, all the asserted properties hold on $\mathcal{R}_\Theta$, which proves the proposition. □

C.2.4 Task-Aware Tracking under Exogenous Arrivals

Task-wise deficit tracking. Recall that $\mathcal{F}_t := \sigma((q_s, S_s, Y_s) : 1 \leq s \leq t)$ denotes the history available after round t. Recall that, throughout the theoretical analysis, the target design

$$w(t) = \big(w_1(t), \ldots, w_J(t)\big) \in \Delta(\mathcal{S})^J$$

is chosen before observing q_{t+1}. Whenever $\hat{\theta}_{t+1} \in \mathcal{R}_\Theta$, we take $w(t) \in C^*_{\mathcal{D}}(\hat{\theta}_{t+1})$; otherwise, the empirical best model is nonunique, and we may choose $w(t)$ according to any fixed, δ-independent rule such that $w(t) \in \Delta(\mathcal{S})^J$ is $\mathcal{F}_t$-measurable.

Recall also the realized task–assortment counts $N_{j,S}(t) := \sum_{s=1}^t \mathbf{1}\{q_s = j,\ S_s = S\}$. Under the indexing used here, write the corresponding pre-action cumulative target as

$$C_{j,S}(t+1) := \sum_{s=1}^{t} \mathbf{1}\{q_{s+1} = j\}\, w_{j,S}(s), \qquad C_j(t+1) := \big(C_{j,S}(t+1)\big)_{S \in \mathcal{S}},$$

with $C_{j,S}(1) := 0$.

After observing $q_{t+1} = j$, on a non-forced round the deficit tracker selects

$$S_{t+1} \in \arg \min_{S \in \mathrm{supp}(C_j(t+1))} \{N_{j,S}(t) - C_{j,S}(t+1)\}.$$

On a forced-exploration round, this choice is overridden by the task-wise round-robin safeguard of Proposition S1.

Averaging under exogenous task arrivals. Define the Cesàro average of the target designs by $\overline{w}(t) := t^{-1} \sum_{s=1}^{t} w(s)$.

Lemma S10 (Averaging under exogenous task arrivals)**.** *Suppose that the task types are drawn independently according to $p = (p_1, \dots, p_J)$ and that $w(t)$ is $\mathcal{F}_t$-measurable for every $t \geq 1$. Then, almost surely, for every $j \in [J]$ and $S \in \mathcal{S}$,*

$$\frac{C_{j,S}(t+1)}{t} - p_j \overline{w}_{j,S}(t) \longrightarrow 0.$$

Proof. Fix $j \in [J]$ and $S \in \mathcal{S}$. By the definition of the pre-action cumulative target, $C_{j,S}(t+1) = \sum_{s=1}^{t} \mathbf{1}\{q_{s+1} = j\}\, w_{j,S}(s)$. Hence

$$M_{j,S}(t) := C_{j,S}(t+1) - p_j \sum_{s=1}^{t} w_{j,S}(s) = \sum_{s=1}^{t} \big(\mathbf{1}\{q_{s+1} = j\} - p_j\big) w_{j,S}(s).$$

Since $w(s)$ is $\mathcal{F}_s$-measurable and q_{s+1} is independent of $\mathcal{F}_s$ with $\mathbb{P}(q_{s+1} = j) = p_j$, we have $\mathbb{E}[(\mathbf{1}\{q_{s+1} = j\} - p_j) w_{j,S}(s) \mid \mathcal{F}_s] = 0$. Moreover, $|(\mathbf{1}\{q_{s+1} = j\} - p_j) w_{j,S}(s)| \leq 1$. Thus $\{M_{j,S}(t)\}_{t \geq 1}$ is a martingale with bounded increments with respect to the shifted filtration $\{\mathcal{F}_{t+1}\}_{t \geq 0}$.

By Azuma–Hoeffding, for every $\varepsilon > 0$,

$$\mathbb{P}(|M_{j,S}(t)| \geq \varepsilon t) \leq 2 \exp\left(-\frac{\varepsilon^2 t}{2}\right).$$

Hence $\sum_{t=1}^{\infty} \mathbb{P}(|M_{j,S}(t)| \geq \varepsilon t) \leq 2 \sum_{t=1}^{\infty} \exp(-\varepsilon^2 t/2) < \infty$. By the Borel–Cantelli lemma, $\mathbb{P}(|M_{j,S}(t)| \geq \varepsilon t \text{ infinitely often}) = 0$. Equivalently, with probability one, there exists a finite T_ε such that $|M_{j,S}(t)|/t < \varepsilon$ for all $t \geq T_\varepsilon$.

Applying this argument to the countable sequence $\varepsilon_m = 1/m$, $m \geq 1$, and intersecting the corresponding probability-one events yields $M_{j,S}(t)/t \to 0$ almost surely. Therefore,

$$\frac{C_{j,S}(t+1)}{t} - p_j \overline{w}_{j,S}(t) \longrightarrow 0 \qquad \text{a.s.}$$

Since J and $\mathcal{S}$ are finite, the preceding convergence holds almost surely simultaneously for all $j \in [J]$ and $S \in \mathcal{S}$. □

Sublinear tracking error under forced exploration. Let $m := |\mathcal{S}|$. For each $j \in [J]$ and $t \geq 1$, let $F_j(t)$ denote the number of forced-exploration overrides in the task-j block among rounds $2, \dots, t$, and set $F(t) := \sum_{j=1}^{J} F_j(t)$. The initialization at round 1 is not counted as an override.

Lemma S11 (Tracking discrepancy under forced exploration)**.** *Under the task-wise deficit-tracking rule above, for every $t \geq 1$, $j \in [J]$, and $S \in \mathcal{S}$,*

$$N_{j,S}(t) - C_{j,S}(t) \leq F_j(t) + 1.$$

Consequently,

$$\max_{S \in \mathcal{S}} |N_{j,S}(t) - C_{j,S}(t)| \leq m\big(F_j(t) + 1\big).$$

Under the forced-exploration safeguard of Proposition S1 with fixed $\alpha \in (0, 1)$,

$$\max_{j \in [J],\, S \in \mathcal{S}} |N_{j,S}(t) - C_{j,S}(t)| = O(t^\alpha) = o(t) \qquad \textit{a.s.}$$

Proof. For each $j \in [J]$ and $S \in \mathcal{S}$, define the post-action discrepancy $D_{j,S}(t) := N_{j,S}(t) - C_{j,S}(t)$.

Step 1: Task-wise deficit dynamics. We first prove by induction that

$$D_{j,S}(t) \le F_j(t) + 1, \qquad \forall\, t \ge 1,\ j \in [J],\ S \in \mathcal{S}. \tag{19}$$

At $t = 1$, since $C_{j,S}(1) = 0$ and $F_j(1) = 0$, $D_{j,S}(1) = N_{j,S}(1) = \mathbf{1}\{q_1 = j,\ S_1 = S\} \le 1 = F_j(1) + 1$. Hence (19) holds at the initial time.

Suppose now that it holds after round t, and consider round $t+1$. If $q_{t+1} \neq j$, then the task-j counts, targets, and forced-round count are unchanged: $D_{j,S}(t+1) = D_{j,S}(t)$ and $F_j(t+1) = F_j(t)$, so the induction claim is preserved.

It therefore remains to consider the case $q_{t+1} = j$. Before the assortment is selected, define $\widetilde{D}_{j,S}(t+1) := N_{j,S}(t) - C_{j,S}(t+1)$. Since $C_{j,S}(t+1) = C_{j,S}(t) + w_{j,S}(t)$ on such a round,

$$\widetilde{D}_{j,S}(t+1) = D_{j,S}(t) - w_{j,S}(t) \le D_{j,S}(t). \tag{20}$$

We first note that the minimum pre-action discrepancy over the cumulative support is nonpositive. Indeed, $\sum_{S\in\mathcal{S}} \widetilde{D}_{j,S}(t+1) = \sum_{r=1}^{t} \mathbf{1}\{q_r = j\} - \sum_{s=1}^{t} \mathbf{1}\{q_{s+1} = j\} = \mathbf{1}\{q_1 = j\} - \mathbf{1}\{q_{t+1} = j\} = \mathbf{1}\{q_1 = j\} - 1 \le 0$. Moreover, if $S \notin \mathrm{supp}(C_j(t+1))$, then $C_{j,S}(t+1) = 0$, and hence $\widetilde{D}_{j,S}(t+1) = N_{j,S}(t) \ge 0$. It follows that $\sum_{S\in\mathrm{supp}(C_j(t+1))} \widetilde{D}_{j,S}(t+1) \le 0$, and therefore

$$\min_{S\in\mathrm{supp}(C_j(t+1))} \widetilde{D}_{j,S}(t+1) \le 0. \tag{21}$$

We now distinguish the two possible actions at round $t+1$.

Case 1: non-forced tracking. By the tracking rule and (21), $\widetilde{D}_{j,S_{t+1}}(t+1) \le 0$. After playing S_{t+1}, $D_{j,S_{t+1}}(t+1) = \widetilde{D}_{j,S_{t+1}}(t+1) + 1 \le 1$. For every $S \neq S_{t+1}$, $D_{j,S}(t+1) = \widetilde{D}_{j,S}(t+1) \le D_{j,S}(t) \le F_j(t) + 1$. Since the round is not forced, $F_j(t+1) = F_j(t)$, and hence $D_{j,S}(t+1) \le F_j(t+1) + 1$ for every $S \in \mathcal{S}$.

Case 2: forced-exploration override. The selected assortment need no longer minimize the deficit. Nevertheless, using (20), $D_{j,S_{t+1}}(t+1) = D_{j,S_{t+1}}(t) - w_{j,S_{t+1}}(t) + 1 \le D_{j,S_{t+1}}(t) + 1 \le F_j(t) + 2$. Since this round contributes one additional task-j forced override, $F_j(t+1) = F_j(t) + 1$, and therefore $D_{j,S_{t+1}}(t+1) \le F_j(t+1) + 1$. For every $S \neq S_{t+1}$, $D_{j,S}(t+1) = D_{j,S}(t) - w_{j,S}(t) \le D_{j,S}(t) \le F_j(t) + 1 \le F_j(t+1) + 1$. Thus (19) is preserved in both cases, completing the induction.

To obtain the corresponding lower bound, observe that for every $t \ge 1$, $\sum_{S\in\mathcal{S}} D_{j,S}(t) = \sum_{r=1}^{t} \mathbf{1}\{q_r = j\} - \sum_{s=1}^{t-1} \mathbf{1}\{q_{s+1} = j\} = \mathbf{1}\{q_1 = j\}$. Hence, for any $S \in \mathcal{S}$, $D_{j,S}(t) = \mathbf{1}\{q_1 = j\} - \sum_{S'\neq S} D_{j,S'}(t) \ge -(m-1)\big(F_j(t) + 1\big)$, where the last inequality follows from (19). Combining the upper and lower bounds gives

$$\max_{S\in\mathcal{S}} |N_{j,S}(t) - C_{j,S}(t)| \le m\big(F_j(t) + 1\big). \tag{22}$$

Step 2: Sublinear tracking error. By Proposition S1, there exists an event Ω_{tr} of probability one on which $\limsup_{t\to\infty} F(t)/t^{\alpha} \le 1$. Here $F(t)$ counts forced overrides among rounds $2, \dots, t$, as in that proposition.

Fix a sample path in Ω_{tr}. By the definition of the upper limit, there exists a finite, sample-path-dependent time T such that $F(t) \le 2t^{\alpha}$ for every $t \ge T$. In particular, $F(t) = O(t^{\alpha})$ on this path.

For every $j \in [J]$, we have $F_j(t) \le F(t)$. Combining this inequality with the deterministic discrepancy bound established in Step 1 gives

$$\max_{j\in[J],\, S\in\mathcal{S}} |N_{j,S}(t) - C_{j,S}(t)| \le m\big(F(t) + 1\big) = O(t^{\alpha}).$$

Since $\alpha < 1$, division by t bounds the normalized discrepancy by a quantity of order $t^{\alpha-1}$, which tends to zero. Thus the discrepancy is also $o(t)$ on this sample path. Since the path in Ω_{tr} was arbitrary and $\mathbb{P}(\Omega_{\mathrm{tr}}) = 1$, we conclude that

$$\max_{j\in[J],\, S\in\mathcal{S}} |N_{j,S}(t) - C_{j,S}(t)| = O(t^{\alpha}) = o(t) \qquad \text{a.s.}$$

□

C.2.5 Proof of Proposition 2: Task-Aware Allocation Convergence

Proof. Let $C^* := C^*_{\mathcal{D}}(\theta^*)$, and recall that $\overline{w}(t) := t^{-1}\sum_{s=1}^{t} w(s)$. By Corollary S2, there exists an event Ω_{con} with $\mathbb{P}(\Omega_{\mathrm{con}}) = 1$ on which $\hat{\theta}_t \to \theta^*$. By Lemma S10, there exists an event Ω_{arr} with $\mathbb{P}(\Omega_{\mathrm{arr}}) = 1$ on which, simultaneously for all $j \in [J]$ and $S \in \mathcal{S}$,

$$\frac{C_{j,S}(t+1)}{t} - p_j \overline{w}_{j,S}(t) \longrightarrow 0.$$

By Lemma S11, there exists an event Ω_{tr} with $\mathbb{P}(\Omega_{\mathrm{tr}}) = 1$ on which

$$\max_{j\in[J],\, S\in\mathcal{S}} |N_{j,S}(t) - C_{j,S}(t)| = o(t).$$

Set $\Omega_0 := \Omega_{\mathrm{con}} \cap \Omega_{\mathrm{arr}} \cap \Omega_{\mathrm{tr}}$. Then $\mathbb{P}(\Omega_0) = 1$. We work throughout the remainder of the proof on Ω_0.

Step 1: Convergence of the averaged target design. By Lemma S6, $\mathcal{R}_\Theta$ contains a relative neighborhood of θ^*; hence $\hat{\theta}_{t+1} \in \mathcal{R}_\Theta$ eventually on Ω_0. Consequently, for all sufficiently large t, the empirical best model is unique, and $w(t) \in C^*_{\mathcal{D}}(\hat{\theta}_{t+1})$ by the target-design rule above. Since $\hat{\theta}_{t+1} \to \theta^*$, Corollary S3 therefore yields

$$\mathrm{dist}(w(t), C^*) \longrightarrow 0. \tag{23}$$

For every $t \geq 1$, compactness of C^* from Lemma S8 allows us to choose $v(t) \in C^*$ such that $\|w(t) - v(t)\|_2 = \mathrm{dist}(w(t), C^*)$. Define $\overline{v}(t) := t^{-1}\sum_{s=1}^{t} v(s)$. Since C^* is convex by Lemma S9, $\overline{v}(t) \in C^*$. Consequently,

$$\mathrm{dist}(\overline{w}(t), C^*) \leq \|\overline{w}(t) - \overline{v}(t)\|_2 \leq \frac{1}{t}\sum_{s=1}^{t} \|w(s) - v(s)\|_2 = \frac{1}{t}\sum_{s=1}^{t} \mathrm{dist}(w(s), C^*)\,.$$

The last expression converges to zero by (23) and Cesàro convergence. Hence

$$\mathrm{dist}(\overline{w}(t), C^*) \longrightarrow 0. \tag{24}$$

Step 2: Alignment of realized counts with the averaged target. Lemma S11 gives

$$\max_{j\in[J],\, S\in\mathcal{S}} |N_{j,S}(t+1) - C_{j,S}(t+1)| = o(t).$$

Combining this with the task-arrival averaging on Ω_0, we obtain, for every $j \in [J]$ and $S \in \mathcal{S}$,

$$\frac{N_{j,S}(t+1)}{t} - p_j \overline{w}_{j,S}(t) = \frac{N_{j,S}(t+1) - C_{j,S}(t+1)}{t} + \left[\frac{C_{j,S}(t+1)}{t} - p_j \overline{w}_{j,S}(t)\right] \longrightarrow 0.$$

Thus,

$$\frac{N_{j,S}(t+1)}{t} - p_j \overline{w}_{j,S}(t) \longrightarrow 0, \qquad \forall\, j \in [J],\ S \in \mathcal{S}. \tag{25}$$

Summing (25) over $S \in \mathcal{S}$ and using $\sum_{S\in\mathcal{S}} \overline{w}_{j,S}(t) = 1$, we obtain

$$\frac{N_j(t+1)}{t} \longrightarrow p_j, \qquad j \in [J]. \tag{26}$$

Since $p_j > 0$, for every $j \in [J]$ there is a finite, sample-path-dependent time after which $N_j(t+1) > 0$. As J is finite, all empirical conditional allocations $\widehat{w}_j(t+1)$ are therefore simultaneously well defined for all sufficiently large t on Ω_0.

For such t, fix $j \in [J]$ and $S \in \mathcal{S}$. We may write

$$\begin{aligned}\widehat{w}_{j,S}(t+1) - \overline{w}_{j,S}(t) &= \frac{N_{j,S}(t+1)/t}{N_j(t+1)/t} - \overline{w}_{j,S}(t) \\ &= \frac{[N_{j,S}(t+1)/t - p_j\overline{w}_{j,S}(t)] - \overline{w}_{j,S}(t)\,[N_j(t+1)/t - p_j]}{N_j(t+1)/t}.\end{aligned}$$

By (25) and (26), both bracketed terms in the numerator converge to zero. Moreover, $0 \leq \overline{w}_{j,S}(t) \leq 1$ and $N_j(t+1)/t \to p_j > 0$. It follows that $\widehat{w}_{j,S}(t+1) - \overline{w}_{j,S}(t) \to 0$. Since J and $\mathcal{S}$ are finite,

$$\|\widehat{w}(t+1) - \overline{w}(t)\|_2 \longrightarrow 0. \tag{27}$$

Step 3: Convergence to the optimal-design set. By the triangle inequality for the distance to a set,

$$\operatorname{dist}(\widehat{w}(t+1), C^*) \le \|\widehat{w}(t+1) - \overline{w}(t)\|_2 + \operatorname{dist}(\overline{w}(t), C^*)\,.$$

The first term converges to zero by (27), while the second converges to zero by (24). Therefore, $\operatorname{dist}(\widehat{w}(t+1), C^*) \to 0$ on Ω_0. Since $\mathbb{P}(\Omega_0) = 1$, and shifting the time index by one does not affect the limit,

$$\operatorname{dist}(\widehat{w}(t), C^*_{\mathcal{D}}(\theta^*)) \longrightarrow 0 \qquad \text{a.s.}$$

□

C.3 Separation Growth and Fixed-Confidence Sample-Complexity Upper Bound

C.3.1 Proof of Proposition 3: Asymptotic Growth of Separation

Proof. Let $C^* := C^*_{\mathcal{D}}(\theta^*)$. By Corollary S2, Proposition 2, and (26), there exists an event Ω_0 with $\mathbb{P}_{\theta^*}(\Omega_0) = 1$ on which

$$\hat{\theta}_t \longrightarrow \theta^*, \qquad \operatorname{dist}(\widehat{w}(t), C^*) \longrightarrow 0, \qquad \frac{N_j(t)}{t} \longrightarrow p_j \quad \forall\, j \in [J],$$

where the last convergence follows from (26) and $(t-1)/t \to 1$. We work throughout the proof on Ω_0.

For $t \ge 2$, define the normalized realized Fisher information $V_t^{\mathrm{emp}}(\theta) := H_t(\theta)/(t-1)$. Since $H_t(\theta)$ contains the observations from rounds $1, \dots, t-1$, for all sufficiently large t, when every $N_j(t-1) > 0$, we may write

$$\begin{aligned} V_t^{\mathrm{emp}}(\theta) &= \sum_{j=1}^{J} \sum_{S \in \mathcal{S}} \frac{N_{j,S}(t-1)}{t-1} \Sigma_{j,S}(\theta) \\ &= \sum_{j=1}^{J} \frac{N_j(t-1)}{t-1} \sum_{S \in \mathcal{S}} \widehat{w}_{j,S}(t-1) \Sigma_{j,S}(\theta). \end{aligned}$$

On the other hand, $V(\widehat{w}(t-1); \vartheta) = \sum_{j=1}^{J} p_j \sum_{S \in \mathcal{S}} \widehat{w}_{j,S}(t-1) \Sigma_{j,S}(\vartheta)$. Define

$$\varepsilon_t := \max_{j \in [J]} \left| \frac{N_j(t-1)}{(t-1)p_j} - 1 \right|.$$

Since $N_j(t-1)/(t-1) \to p_j > 0$ for every j on Ω_0, and J is finite, we have $\varepsilon_t \to 0$. Thus, for all sufficiently large t, $\varepsilon_t < 1$ and $(1-\varepsilon_t)p_j \le N_j(t-1)/(t-1) \le (1+\varepsilon_t)p_j$ for every $j \in [J]$. For each j, the matrix $\sum_{S \in \mathcal{S}} \widehat{w}_{j,S}(t-1) \Sigma_{j,S}(\vartheta)$ is positive semidefinite, since its coefficients are nonnegative and each $\Sigma_{j,S}(\vartheta)$ is positive semidefinite. Comparing the coefficients in the two representations therefore gives, simultaneously for all $\vartheta \in \Theta$,

$$(1-\varepsilon_t) V(\widehat{w}(t-1); \vartheta) \preceq V_t^{\mathrm{emp}}(\vartheta) \preceq (1+\varepsilon_t) V(\widehat{w}(t-1); \vartheta).$$

By Lemma S6, $\mathcal{R}_\Theta$ contains a relative neighborhood of θ^* in Θ. Since $\hat{\theta}_t \in \Theta$ and $\hat{\theta}_t \to \theta^*$, we have $\hat{\theta}_t \in \mathcal{R}_\Theta$ for all sufficiently large t. For such t, the alternative set $A_t := \mathrm{Alt}_{\mathcal{D}}(\hat{\theta}_t)$ is nonempty: indeed, $-\hat{\theta}_t \in \Theta \subseteq \mathcal{D}$ and reverses each strict inequality defining the unique best model at $\hat{\theta}_t$, so $-\hat{\theta}_t \in A_t$.

Set $\vartheta = \hat{\theta}_t$ in the matrix comparison above. For every $\lambda \in A_t$, applying the matrix comparison to $\hat{\theta}_t - \lambda$ and multiplying by $1/2$ yields

$$\begin{aligned} \frac{1-\varepsilon_t}{2} \|\hat{\theta}_t - \lambda\|^2_{V(\widehat{w}(t-1); \hat{\theta}_t)} &\le \frac{1}{2} \|\hat{\theta}_t - \lambda\|^2_{V_t^{\mathrm{emp}}(\hat{\theta}_t)} \\ &\le \frac{1+\varepsilon_t}{2} \|\hat{\theta}_t - \lambda\|^2_{V(\widehat{w}(t-1); \hat{\theta}_t)}. \end{aligned}$$

Taking infima over the same set A_t, using $V_t^{\mathrm{emp}}(\hat{\theta}_t) = H_t(\hat{\theta}_t)/(t-1)$ and $1 - \varepsilon_t > 0$, yields

$$\begin{aligned} (1-\varepsilon_t) \psi_{\mathcal{D}}(\hat{\theta}_t, \widehat{w}(t-1)) &\le \frac{\mathrm{Sep}_t^{\mathcal{D}}}{t-1} \\ &\le (1+\varepsilon_t) \psi_{\mathcal{D}}(\hat{\theta}_t, \widehat{w}(t-1)). \end{aligned} \tag{28}$$

It remains to identify the limit of the population separation. Since $C^* = C^*_{\mathcal{D}}(\theta^*)$ is nonempty and compact, for each sufficiently large t choose $v(t-1) \in C^*$ such that

$$\|\widehat{w}(t-1) - v(t-1)\|_2 = \operatorname{dist}(\widehat{w}(t-1), C^*) \longrightarrow 0.$$

We claim that

$$\psi_{\mathcal{D}}(\hat{\theta}_t, \widehat{w}(t-1)) - \psi_{\mathcal{D}}(\theta^*, v(t-1)) \longrightarrow 0.$$

Indeed, if this convergence failed, there would exist $\eta > 0$ and a subsequence (t_k) along which the absolute value of the displayed difference is at least η. By compactness of C^*, passing to a further subsequence if necessary, we may assume that $v(t_k - 1) \to \bar{w} \in C^*$. The distance convergence above then gives $\widehat{w}(t_k - 1) \to \bar{w}$. Since $\hat{\theta}_{t_k} \to \theta^*$ and $\hat{\theta}_{t_k} \in \mathcal{R}_\Theta$ eventually, joint continuity from Lemma S7 implies that both $\psi_{\mathcal{D}}(\hat{\theta}_{t_k}, \widehat{w}(t_k - 1))$ and $\psi_{\mathcal{D}}(\theta^*, v(t_k - 1))$ converge to $\psi_{\mathcal{D}}(\theta^*, \bar{w})$. Their difference therefore converges to zero, contradicting its absolute value being at least η.

Because $v(t-1) \in C^*_{\mathcal{D}}(\theta^*)$, we have $\psi_{\mathcal{D}}(\theta^*, v(t-1)) = \psi^*_{\mathcal{D}}(\theta^*)$. Consequently,

$$\psi_{\mathcal{D}}(\hat{\theta}_t, \widehat{w}(t-1)) \longrightarrow \psi^*_{\mathcal{D}}(\theta^*).$$

Together with $\varepsilon_t \to 0$, this shows that both the lower and upper bounds in (28) converge to the same finite value $\psi^*_{\mathcal{D}}(\theta^*)$ on Ω_0. The squeeze theorem and $\mathbb{P}_{\theta^*}(\Omega_0) = 1$ therefore yield

$$\frac{\mathrm{Sep}^{\mathcal{D}}_t}{t-1} \longrightarrow \psi^*_{\mathcal{D}}(\theta^*), \qquad \text{a.s.} \tag{29}$$

□

C.3.2 Fixed-confidence correctness

Proposition S5 (Fixed-confidence correctness)**.** *The stopping rule* (5) *is δ-correct over Θ:*

$$\mathbb{P}_{\theta^*}\left(\tau_\delta < \infty,\, \hat{i}_{\tau_\delta} \notin \arg\max_{i \in [N]} \bar{x}_i^\top \theta^*\right) \le \delta, \qquad \forall\, \theta^* \in \Theta.$$

Proof. Fix $\delta \in (0,1)$ and $\theta^* \in \Theta$. Define the event

$$\mathcal{E}_\delta := \left\{ \|\hat{\theta}_t - \theta^*\|_{H_t(\hat{\theta}_t)} \le e^{3WD_X} \gamma_t^{\lambda_\delta}(\delta) \text{ for every } t \text{ such that } \lambda_{\min}(H_t(\theta^*)) \ge \lambda_\delta \right\}.$$

By Proposition 1, $\mathbb{P}_{\theta^*}(\mathcal{E}_\delta) \ge 1 - \delta$.

Consider any sample path in $\mathcal{E}_\delta$ on which $\tau_\delta = t < \infty$. By the information condition in (5), $\lambda_{\min}(U_{t-1}) \ge \kappa_{W,K}^{-1} \lambda_\delta$. Lemma S1 gives $H_t(\theta^*) \succeq \kappa_{W,K} U_{t-1}$, and therefore $\lambda_{\min}(H_t(\theta^*)) \ge \kappa_{W,K} \lambda_{\min}(U_{t-1}) \ge \lambda_\delta$. Hence, on $\mathcal{E}_\delta$,

$$\|\hat{\theta}_t - \theta^*\|_{H_t(\hat{\theta}_t)} \le e^{3WD_X} \gamma_t^{\lambda_\delta}(\delta). \tag{30}$$

Suppose, toward a contradiction, that the recommendation at time t is incorrect: $\hat{i}_t \notin \arg\max_{i \in [N]} \bar{x}_i^\top \theta^*$. By the stopping rule, $\hat{i}_t = i^*(\hat{\theta}_t)$ is unique. Choose any $i^\dagger \in \arg\max_{i \in [N]} \bar{x}_i^\top \theta^*$. Since $\hat{i}_t$ is not a true maximizer, $i^\dagger \ne \hat{i}_t$ and $\bar{x}_{i^\dagger}^\top \theta^* > \bar{x}_{\hat{i}_t}^\top \theta^*$. Moreover, $\theta^* \in \Theta \subseteq \mathcal{D}$. By the definition of the alternative set, it follows that $\theta^* \in \mathrm{Alt}_{\mathcal{D}}(\hat{\theta}_t)$.

Using θ^* as an admissible point in the infimum defining $\mathrm{Sep}^{\mathcal{D}}_t$, we obtain $\mathrm{Sep}^{\mathcal{D}}_t \le \frac{1}{2} \|\hat{\theta}_t - \theta^*\|^2_{H_t(\hat{\theta}_t)}$. Combining this with (30) gives

$$\mathrm{Sep}^{\mathcal{D}}_t \le \frac{1}{2} e^{6WD_X} \left(\gamma_t^{\lambda_\delta}(\delta)\right)^2 = \beta_t^{\lambda_\delta}(\delta).$$

This contradicts the stopping condition $\mathrm{Sep}^{\mathcal{D}}_t > \beta_t^{\lambda_\delta}(\delta)$. Therefore, every finite stopping time on $\mathcal{E}_\delta$ yields the correct recommendation.

Consequently,

$$\left\{ \tau_\delta < \infty,\, \hat{i}_{\tau_\delta} \notin \arg\max_{i \in [N]} \bar{x}_i^\top \theta^* \right\} \subseteq \mathcal{E}_\delta^c.$$

Hence

$$\mathbb{P}_{\theta^*}\left(\tau_\delta < \infty,\, \hat{i}_{\tau_\delta} \notin \arg\max_{i \in [N]} \bar{x}_i^\top \theta^*\right) \le \mathbb{P}_{\theta^*}(\mathcal{E}_\delta^c) \le \delta.$$

□

C.3.3 Proof of Theorem 3: Sample Complexity Upper Bound

Proof. Fix $\alpha \in (0,1)$, independent of δ, as in the forced-exploration safeguard. For notational convenience, let

$$\ell_\delta := \log(1/\delta), \qquad p_\alpha := \frac{2}{1+\alpha}, \qquad r_\alpha := \frac{2\alpha}{1+\alpha}.$$

Then $p_\alpha + r_\alpha = 2$, $\alpha p_\alpha = r_\alpha$, and $r_\alpha < p_\alpha$. Under the specialization preceding the theorem, $\lambda_\delta = \ell_\delta^{r_\alpha}$.

Because the sampling rule, including the fixed exponent α, does not depend on δ, we may regard all stopping rules $\{\tau_\delta\}_{\delta\in(0,1)}$ as defined on the same infinite sampling trajectory obtained by continuing the sampling rule indefinitely. By Corollary S2 and Propositions 3 and S1, there exists an event Ω_0 with $\mathbb{P}_{\theta^*}(\Omega_0) = 1$ on which

$$\hat{\theta}_t \longrightarrow \theta^*, \qquad \frac{\mathrm{Sep}_t^{\mathcal{D}}}{t-1} \longrightarrow \psi_{\mathcal{D}}^*(\theta^*),$$
$$\liminf_{t\to\infty} \frac{\lambda_{\min}(U_t)}{t^\alpha} \geq c_X.$$

We work throughout the proof on a fixed sample path in Ω_0.

Define

$$C_\alpha(\theta^*) := \max\left\{8e^{6WD_X} T_{\mathcal{D}}^*(\theta^*), \left(\frac{1}{\kappa_{W,K} c_X}\right)^{1/\alpha}\right\},$$

and fix any constant $C > C_\alpha(\theta^*)$. For $\delta \in (0,1)$, consider the deterministic candidate time

$$t_\delta := \lceil C\ell_\delta^{p_\alpha} \rceil + 1.$$

As $\delta \downarrow 0$,

$$\frac{t_\delta - 1}{\ell_\delta^{p_\alpha}} \longrightarrow C, \qquad t_\delta \longrightarrow \infty. \tag{31}$$

We verify that, for all sufficiently small δ, the three conditions in (5) hold at $t = t_\delta$.

Empirical-best uniqueness. Since $\theta^* \in \mathcal{R}_\Theta$, Lemma S6 yields a relative neighborhood of θ^* on which the best-model identity is constant and unique. Because $\hat{\theta}_t \to \theta^*$ on Ω_0, there exists a finite, sample-path-dependent time T_{best} such that

$$i^*(\hat{\theta}_t) = i^*(\theta^*) \quad \text{and is unique for every } t \geq T_{\mathrm{best}}.$$

By (31), $t_\delta \to \infty$ as $\delta \downarrow 0$. Hence, for all sufficiently small δ, $t_\delta \geq T_{\mathrm{best}}$, and the first stopping condition holds at t_δ.

Information gate. On the fixed sample path in Ω_0, we have $\liminf_{n\to\infty} \lambda_{\min}(U_n)/n^\alpha \geq c_X$. Fix any $\eta \in (0, c_X)$. By the definition of the lower limit, there exists a finite, sample-path-dependent time T_η such that $\lambda_{\min}(U_n) \geq (c_X - \eta)n^\alpha$ for every $n \geq T_\eta$.

By (31), $t_\delta - 1 \to \infty$ as $\delta \downarrow 0$. Thus, for all sufficiently small δ, $t_\delta - 1 \geq T_\eta$, and the preceding bound applies to $U_{t_\delta - 1}$. Using $\lambda_\delta = \ell_\delta^{r_\alpha}$ and $r_\alpha = \alpha p_\alpha$, we obtain

$$\frac{\lambda_{\min}(U_{t_\delta-1})}{\lambda_\delta} \geq (c_X - \eta)\frac{(t_\delta - 1)^\alpha}{\ell_\delta^{r_\alpha}} = (c_X - \eta)\left(\frac{t_\delta - 1}{\ell_\delta^{p_\alpha}}\right)^\alpha.$$

Since $(t_\delta - 1)/\ell_\delta^{p_\alpha} \to C$ by (31), taking lower limits gives $\liminf_{\delta\downarrow 0} \lambda_{\min}(U_{t_\delta-1})/\lambda_\delta \geq (c_X - \eta)C^\alpha$. Because $\eta \in (0, c_X)$ was arbitrary, letting $\eta \downarrow 0$ yields

$$\liminf_{\delta\downarrow 0} \frac{\lambda_{\min}(U_{t_\delta-1})}{\lambda_\delta} \geq c_X C^\alpha.$$

The choice $C > C_\alpha(\theta^*)$ implies $C > (\kappa_{W,K} c_X)^{-1/\alpha}$, and hence $c_X C^\alpha > \kappa_{W,K}^{-1}$. The lower limit is therefore strictly larger than $\kappa_{W,K}^{-1}$, so, for all sufficiently small δ,

$$\lambda_{\min}(U_{t_\delta-1}) \geq \kappa_{W,K}^{-1}\lambda_\delta.$$

Thus the second stopping condition holds at t_δ.

Separation gate. We first identify the asymptotic size of the confidence threshold at the candidate time. Write

$$A_\delta := \log\left[\frac{2^{3d/2}}{\delta}\left(1+\frac{2t_\delta K}{\lambda_\delta d}\right)^{d/2}\right].$$

Since $t_\delta = O(\ell_\delta^{p_\alpha})$ and $\lambda_\delta = \ell_\delta^{r_\alpha}$,

$$\frac{t_\delta}{\lambda_\delta} = O\big(\ell_\delta^{p_\alpha - r_\alpha}\big).$$

Therefore

$$\log\left(1+\frac{2t_\delta K}{\lambda_\delta d}\right) = O(\log \ell_\delta) = o(\ell_\delta),$$

and hence

$$\frac{A_\delta}{\ell_\delta} = 1 + \frac{3d\log 2}{2\ell_\delta} + \frac{d}{2\ell_\delta}\log\left(1+\frac{2t_\delta K}{\lambda_\delta d}\right) \longrightarrow 1. \tag{32}$$

Recalling

$$A_\delta := \log\left[\frac{2^{3d/2}}{\delta}\left(1+\frac{2t_\delta K}{\lambda_\delta d}\right)^{d/2}\right], \qquad \lambda_\delta = \ell_\delta^{r_\alpha},$$

the definition of $\gamma_t^\lambda(\delta)$ gives

$$\begin{aligned}\gamma_{t_\delta}^{\lambda_\delta}(\delta) &= \frac{\sqrt{\lambda_\delta}}{2} + \frac{4}{\sqrt{\lambda_\delta}}\log\left[\frac{2^{3d/2}}{\delta}\left(1+\frac{2t_\delta K}{\lambda_\delta d}\right)^{d/2}\right] \\ &= \frac{\sqrt{\lambda_\delta}}{2} + \frac{4A_\delta}{\sqrt{\lambda_\delta}} \\ &= \frac{1}{2}\ell_\delta^{r_\alpha/2} + 4A_\delta \ell_\delta^{-r_\alpha/2}.\end{aligned}$$

Hence

$$\frac{\gamma_{t_\delta}^{\lambda_\delta}(\delta)}{\ell_\delta^{p_\alpha/2}} = \frac{1}{2}\ell_\delta^{(r_\alpha - p_\alpha)/2} + 4\frac{A_\delta}{\ell_\delta^{(p_\alpha+r_\alpha)/2}}.$$

Since $r_\alpha < p_\alpha$, $p_\alpha + r_\alpha = 2$, and (32) holds, we obtain

$$\frac{\gamma_{t_\delta}^{\lambda_\delta}(\delta)}{\ell_\delta^{p_\alpha/2}} \longrightarrow 4.$$

Thus,

$$\frac{\beta_{t_\delta}^{\lambda_\delta}(\delta)}{\ell_\delta^{p_\alpha}} = \frac{1}{2}e^{6WD_X}\left(\frac{\gamma_{t_\delta}^{\lambda_\delta}(\delta)}{\ell_\delta^{p_\alpha/2}}\right)^2 \longrightarrow 8e^{6WD_X}. \tag{33}$$

On the other hand, Proposition 3 and (31) give, on Ω_0,

$$\frac{\mathrm{Sep}_{t_\delta}^{\mathcal{D}}}{\ell_\delta^{p_\alpha}} = \frac{\mathrm{Sep}_{t_\delta}^{\mathcal{D}}}{t_\delta - 1}\frac{t_\delta - 1}{\ell_\delta^{p_\alpha}} \longrightarrow C\psi_{\mathcal{D}}^*(\theta^*).$$

Because $C > C_\alpha(\theta^*) \geq 8e^{6WD_X}T_{\mathcal{D}}^*(\theta^*)$ and $T_{\mathcal{D}}^*(\theta^*) = 1/\psi_{\mathcal{D}}^*(\theta^*)$,

$$C\psi_{\mathcal{D}}^*(\theta^*) > 8e^{6WD_X}.$$

Combining this strict inequality with (33), we conclude that, for all sufficiently small δ,

$$\mathrm{Sep}_{t_\delta}^{\mathcal{D}} > \beta_{t_\delta}^{\lambda_\delta}(\delta).$$

Hence the third stopping condition also holds at t_δ.

Conclusion. On the fixed sample path in Ω_0, all three stopping conditions are therefore satisfied at t_δ for every sufficiently small δ. By the definition of τ_δ,

$$\tau_\delta \le t_\delta \qquad \text{for all sufficiently small } \delta.$$

Using (31),

$$\limsup_{\delta \downarrow 0} \frac{\tau_\delta}{\ell_\delta^{p_\alpha}} \le C.$$

The preceding argument holds for every $C > C_\alpha(\theta^*)$. Applying it, for example, to the countable sequence $C_m := C_\alpha(\theta^*) + 1/m$, $m \ge 1$, gives

$$\limsup_{\delta \downarrow 0} \frac{\tau_\delta}{\ell_\delta^{p_\alpha}} \le C_\alpha(\theta^*) \qquad \text{on } \Omega_0.$$

Since $\mathbb{P}_{\theta^*}(\Omega_0) = 1$ and $p_\alpha = 2/(1+\alpha)$,

$$\mathbb{P}_{\theta^*}\left(\limsup_{\delta \downarrow 0} \frac{\tau_\delta}{\log^{2/(1+\alpha)}(1/\delta)} \le \max\left\{ 8e^{6WD_X} T^*_{\mathcal{D}}(\theta^*), \left(\frac{1}{\kappa_{W,K} c_X} \right)^{1/\alpha} \right\} \right) = 1.$$

□

C.3.4 Near-Optimal Confidence Dependence

Corollary S4 (Near-optimal confidence dependence)**.** *For any fixed $\varepsilon \in (0,1)$, choose $\alpha = \frac{1-\varepsilon}{1+\varepsilon}$. Then, under the conditions of Theorem 3, $\tau_\delta = O_{\mathrm{a.s.}}\big(\log^{1+\varepsilon}(1/\delta)\big)$ as $\delta \downarrow 0$.*

Proof. Fix any $\varepsilon \in (0,1)$. By the choice $\alpha = (1-\varepsilon)/(1+\varepsilon)$ in the statement, we have $\alpha \in (0,1)$, so Theorem 3 applies. Moreover, $1+\alpha = 2/(1+\varepsilon)$, and hence $2/(1+\alpha) = 1+\varepsilon$, while $1/\alpha = (1+\varepsilon)/(1-\varepsilon)$.

Therefore, Theorem 3 yields

$$\mathbb{P}_{\theta^*}\left(\limsup_{\delta \downarrow 0} \frac{\tau_\delta}{\log^{1+\varepsilon}(1/\delta)} \le C_\varepsilon(\theta^*) \right) = 1,$$

where $C_\varepsilon(\theta^*) := \max\big\{ 8e^{6WD_X} T^*_{\mathcal{D}}(\theta^*), (\kappa_{W,K} c_X)^{-(1+\varepsilon)/(1-\varepsilon)} \big\} < \infty$.

Let Ω_ε denote the probability-one event in the preceding display, and fix any sample path $\omega \in \Omega_\varepsilon$. By the definition of the limit superior, for any $\eta > 0$ there exists $\delta_0 = \delta_0(\omega, \eta) > 0$ such that, for every $0 < \delta < \delta_0$, $\tau_\delta(\omega)/\log^{1+\varepsilon}(1/\delta) \le C_\varepsilon(\theta^*) + \eta$. Equivalently, $\tau_\delta(\omega) \le \big(C_\varepsilon(\theta^*) + \eta\big) \log^{1+\varepsilon}(1/\delta)$ for all sufficiently small δ.

Since this holds on an event of probability one,

$$\tau_\delta = O_{\mathrm{a.s.}}\big(\log^{1+\varepsilon}(1/\delta)\big),$$

which proves the corollary. □

D Implementation Details for the Arena Experiments

D.1 Environment construction

We supplement Section 7.1 with the main preprocessing details:

(i) Response embeddings. Final responses are split into non-overlapping chunks of at most 254 content tokens. Their 384-dimensional MiniLM embeddings are ℓ_2-normalized, pooled using content-token weights, and normalized again.

(ii) Task-dependent features. In the original 384-dimensional embedding space, we apply centered, full-dimensional PCA without whitening or truncation, average features within task–LLM cells, center within each task, and express the resulting features in the 28-dimensional span of within-task differences.

(iii) Preference model and task distribution. An unregularized shared logistic MLE fitted to all 9,895 retained comparisons defines the feedback probabilities. The code/non-code task probabilities are $(3327, 6568)/9895$.

The released code includes this frozen 15-LLM environment and instructions for rebuilding it from the external data and encoder.

D.2 Algorithm implementation and computational cost

We use $T = 1000$ and 100 replications with matched random seeds across methods.

(i) Estimation. MULTI-FW and RANDOM share a 28-dimensional contextual estimator; MIS1H-FW fits a 14-dimensional identifiable one-hot surrogate. These methods refit the full-history constrained MLE every 10 observations, warm-starting from the previous estimate and imposing an ℓ_2-norm bound of 50. We use SLSQP with analytic derivatives, tolerance 10^{-9}, and a nominal 200-iteration budget, retaining finite feasible iterates at the iteration limit. Estimates remain fixed between updates.

(ii) Allocation. Both FW methods refresh their allocations every 50 rounds, starting at the first round; targets remain fixed between refreshes and are implemented through deficit tracking. Each refresh uses 200 FW iterations, warm-starting from the previous allocation. The step sizes are $2/(j + 2)$ for the initial allocation update and $2/(j + 3)$ for subsequent updates, with j indexing the inner FW iterations from zero. We add a 10^{-6} diagonal ridge to the Fisher matrices used for allocation and plotted separation computations.

(iii) Exploration. Estimator-based methods use 10 initial forced-exploration rounds and subsequently force exploration when $\lambda_{\min}(U_{t-1}) < 10^{-4}\sqrt{t-1}$.

(iv) MISTTB. MISTTB uses $(\epsilon, \delta) = (0.25, 0.20)$, giving batches of 321 comparisons. At $T = 1000$, the final 37 comparisons contribute to evaluation but do not update the incumbent because their batch is incomplete.

(v) Evaluation. Correctness is recorded every round; true-world separation is recomputed from the accumulated history every 50 rounds.

(vi) Computational cost. Table 2 reports the mean recorded runner time, including built-in diagnostics and terminal auxiliary fits, but excluding environment construction, shared initialization, post-run evaluation, and output generation.

Table 2: Mean runner wall-clock time per replication, averaged over the original 100 replications with $T = 1000$.

Method	Time (seconds)
MULTI-FW	126.10
RANDOM	109.75
MIS1H-FW	36.13
MISTTB	1.11