Learning for Irreversible Subset Expansion: Objective Degeneracy and Thompson Sampling under Information Dilution

Yifan Hong[1,2] Huiwen Jia[1] Chen Wang[2]

Abstract. We study an online irreversible subset-expansion problem in which a decision-maker sequentially adds alternatives while learning from interaction feedback. Motivated by facility expansion and product rollouts, we model Poisson arrivals with an unknown rate and multinomial logit choice with unknown linear utilities. The first challenge is that, under bounded-parameter assumptions, terminal market-share regret vanishes as the set grows, even for policies that systematically select low-attraction alternatives. Thus, vanishing regret can reflect saturation of the evaluation scale rather than successful learning. We characterize this objective-induced degeneracy for a separable class of monotone submodular objectives. Total latent attraction preserves terminal subset rankings under market share and provides a nondegenerate intrinsic-quality criterion under a candidate-pool separation condition. Learning under this nondegenerate objective remains challenging: the expanding active set dilutes information, precluding a constant information-ratio bound uniform over time and candidate sets. Our main methodological contribution is a decomposition for Thompson sampling that bounds the squared instantaneous greedy gap by the product of an action-uncertainty term and a parameter-uncertainty term. Using the unnormalized centered design of the selected set, we control these terms through an elliptical potential bound and a mutual information budget, respectively. This yields a Bayesian regret bound for exact Thompson sampling that matches our minimax lower bound in subset size and feature dimension, up to logarithmic factors. Numerical experiments illustrate the results.

Key words: online learning, multi-armed bandit, multinomial logit choice model, Thompson sampling, objective degeneracy

1. Introduction

Many market-facing operational decisions require a firm or planner to build a set of active alternatives sequentially from a larger pool of candidates. For example, a retailer expands a store network (Lin et al. 2024), a logistics provider opens locker or pickup locations (Lin et al. 2020), a platform selects creators or partners to promote content (Mallipeddi et al. 2022, Han et al. 2023), and a seller rolls out products across markets (Elberse and Eliashberg 2003). In such settings, expansion decisions are often persistent and *irreversible*. Opening a facility, launching a product, or committing to a contractual partner may involve fixed costs, contractual obligations, operational frictions, or reputational costs, so that a selected alternative remains active even if later information reveals that it was not ideal. This form of irreversibility is a central feature of many operational investment and capacity-expansion decisions (Luss 1982, Bensoussan and Chevalier-Roignant 2019).

At the same time, the decision-maker typically begins with *incomplete information* about candidate quality and must learn from noisy market feedback while the active set is being constructed

[1] Department of Industrial Engineering and Operations Research, University of California, Berkeley.

[2] Department of Industrial Engineering, Tsinghua University.

(Wen et al. 2017, Zhang et al. 2022). This creates a committed-exploration problem: early decisions determine both the information available for future decisions and the quality of the final constructed subset. Recent work on online learning with irreversible decisions highlights this tension between experimentation and persistent commitment (Jacquillat and Li 2026).

This learning problem differs from a standard multi-armed bandit problem because feedback is generated by the entire *active set*, comprising both the newly selected alternative and all previously selected alternatives, rather than by the new alternative alone. In a choice-based market, opening a new store, adding a product, or selecting a partner changes the choice environment faced by customers. Sales or engagement for any alternative therefore reflects both its own latent quality and its substitution relationship with the rest of the active set. This active-set dependence is the same economic force that motivates random-utility choice models in facility location, revenue management, and assortment optimization (Talluri and van Ryzin 2004, Rusmevichientong et al. 2010, Agrawal et al. 2019). Together, irreversibility and active-set feedback make exploration especially costly. Adding a marginal but informative alternative may improve learning, but it also commits the decision-maker to an asset that remains active throughout the construction process.

We study irreversible subset expansion under choice-based market feedback. In each period of the construction phase, the decision-maker adds one candidate to the active set, and selected candidates cannot subsequently be removed. Customers arrive according to a Poisson process and, conditional on the current active set, choose among the active candidates and an outside option according to a multinomial logit (MNL) model with unknown linear utilities. The learner observes the resulting choice-based feedback and seeks to maximize long-run performance, measured by the quality of the final constructed subset. We focus on settings in which the completed set operates for much longer than the construction phase lasts. With the construction phase fixed, its relative contribution to cumulative rewards becomes negligible as the subsequent operating horizon grows. This temporal separation motivates evaluating learning by terminal subset quality, whereas standard bandit models typically evaluate cumulative rewards earned during learning. We refer to this class of problems as *expansion bandits*.

Irreversibility and the terminal subset objective pose a new challenge for expansion bandits: a meaningful regret notion should remain discriminative across different subset sizes to distinguish between good and bad algorithms regardless of the underlying objective. Market share and expected revenue are natural metrics for evaluating demand capture and cash generation in a single operating period. However, they do not separate the effects of subset size from subset quality. As the subset

size grows, the objective can saturate, causing terminal regret to vanish even for poorly constructed subsets. This observation motivates the first question of the paper: which terminal objectives remain informative for evaluating irreversible subset construction? Under an informative objective, learning faces a further challenge: the expanding active set dilutes information about newly selected alternatives. We then ask how difficult it is to learn from active-set choice feedback and whether adaptive policies can achieve near-optimal regret despite this information dilution.

1.1. Our Contributions

We introduce expansion bandits, a sequential subset-construction bandit model in which a decision-maker selects alternatives from a large candidate set over time. Each candidate has observable features, customer arrivals follow a Poisson process with an unknown arrival rate, and choices are generated by an MNL model with an unknown linear utility parameter. The model captures three operational features that jointly distinguish our setting from standard bandit and assortment models: selections are irreversible, noisy feedback is generated by the entire active set, and the value of a new candidate depends on previously selected alternatives. This formulation formalizes the committed-exploration problem faced in facility expansion, content selection, product rollout, and related operational settings. Our contributions are as follows.

Separating the economic objective from the learning-performance criterion. Our first contribution is to show that the evaluation of an irreversibly constructed set is not settled by choosing an economically meaningful objective. Even after the economic objective is fixed, the choice of terminal regret criterion remains consequential because regret depends on the cardinal scale of the objective. Market share and total attraction mass illustrate this distinction: they rank every constructed set identically; the former is a fixed strictly increasing transformation of the latter, so they differ only in how a shortfall is recorded numerically, not in what they prefer. However, the worst-case terminal regret under market share converges to zero for every policy, while under attraction mass it grows linearly in the subset size. Thus, directly using an economically meaningful objective as the terminal learning criterion can yield an uninformative measure of performance. We refer to this loss of discriminating power as objective degeneracy. In our subset-expansion setting, it arises when terminal regret is measured on the saturating scale of a choice-based objective. We diagnose this loss as follows: for any objective of the form $F_g(S) = g(\sum_{i\in S} w_i)$ in a natural separable class, regret is the attraction-mass regret multiplied by a factor $g'(\Theta(n))$ that depends on the objective alone, so all such criteria are the same quantity read on different scales. Classifying objectives by

the growth of that factor yields a complete characterization: a terminal criterion retains discriminating power in absolute terms exactly when g is asymptotically linear, is bounded exactly when $g'(y) = O(1/y)$, and otherwise grows at a rate we track explicitly.

An informative and nondegenerate terminal objective. We study total latent attraction mass as an intrinsic portfolio-quality measure. Under the MNL model, this quantity is exactly the odds of an inside choice relative to the outside option. This interpretation separates stock from flow: market share and revenue evaluate realized demand and cash flow in a particular operating period, whereas attraction mass evaluates the demand-generating potential embodied in the completed irreversible portfolio. Because this potential is not compressed by the MNL denominator, terminal regret under the intrinsic-quality criterion remains sensitive to systematic selection mistakes as the portfolio grows, and learning remains statistically nontrivial.

Statistical limits and learning algorithms. Under the intrinsic-quality objective, we establish a minimax lower bound that captures the statistical difficulty of learning from active-set choice feedback. The lower bound shows that regret must scale at least on the order of $d\sqrt{n/\underline{\lambda}}$ in the subset size n, feature dimension d, and arrival rate lower bound $\underline{\lambda}$, up to constants and the technical conditions specified later. As a frequentist benchmark, we analyze an explore-then-exploit policy with a fixed-set estimator (ETE-FSE), which exploits the independence-of-irrelevant-alternatives (IIA) structure of the MNL model to update the preference estimate using repeated observations on an early selected subset. Although IIA preserves relative choice probabilities within this reference set, its share of arrivals decreases as the active set expands. Our estimation-error analysis accounts for this dilution and yields a frequentist regret bound with $O(n^{2/3})$ dependence on subset size under the stated feature and exploration conditions. Our main algorithmic result is for Thompson sampling (TS), which updates its posterior using feedback from the entire active set and selects greedily with respect to a sampled preference parameter. For exact Thompson sampling, we establish an $\widetilde{O}(d\sqrt{n})$ Bayesian regret bound that matches the minimax lower bound in subset size and feature dimension, up to logarithmic factors, under stated regularity conditions.

An action–parameter uncertainty decomposition for Thompson sampling. We show that the standard information-theoretic analysis based on a uniform constant bound on the information ratio cannot be applied directly. Specifically, for each period $t \geq 2$, we construct a two-dimensional instance whose information ratio at period t is $\Omega(t)$, ruling out a constant bound that holds uniformly over time and candidate sets. Our main methodological contribution is an action–parameter uncertainty decomposition, formalized in Lemma 2, which bounds the squared instantaneous greedy gap

by the product of an action-uncertainty term and a parameter-uncertainty term. The two terms use inverse and direct weights of the same regularized, *unnormalized* centered design of the selected set. Probability matching links action uncertainty to growth of this design, allowing its cumulative contribution to be controlled by an elliptical potential bound. The parameter-uncertainty term is controlled through one-step lifted information gain and a covering-based mutual information budget. The decomposition yields the near-optimal Bayesian regret bound without requiring a uniform per-period information-ratio bound. The probability-matching and matrix Cauchy–Schwarz steps do not use the Poisson–MNL likelihood, and thus the decomposition can be generalized in other settings with an analogous greedy-gap structure and design-growth and information bounds.

1.2. Paper Structure

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 introduces the irreversible subset-expansion model, the Poisson–MNL feedback process, terminal objectives, and regret criteria. Section 4 studies the degeneracy of market-share and expected-revenue terminal regret and motivates the intrinsic-quality objective. Section 5 establishes a minimax lower bound under the intrinsic-quality objective. Section 6 presents the explore-then-exploit policy, Thompson sampling, and their regret analysis. Section 7 reports numerical experiments comparing the proposed algorithms and illustrating their behavior. Section 8 concludes.

2. Related Work

Our work lies at the intersection of four streams of literature: online learning with choice-based feedback, irreversible bandits, choice-based facility location, and online sequential infrastructure decisions. Online assortment-learning models capture endogenous choice feedback, but allow the offered set to be revised across periods and evaluate performance through cumulative reward. Irreversible bandit models capture commitment, but typically observe rewards or marginal gains associated with individual actions rather than demand jointly generated by an expanding active set. Choice-based facility-location models optimize a final set under discrete-choice demand, but generally take the choice primitives as known or estimate them before optimization. Online infrastructure models study sequential investment or facility-opening decisions under arriving demand, but usually treat demand as exogenous rather than shaped by customer choice over the currently active alternatives. Our setting links these elements by learning preferences from active-set-dependent

demand while constructing a subset irreversibly and evaluating the resulting final set. This formulation also creates a distinct objective-design issue: market share and revenue are natural one-period criteria, yet they may lose discriminating power when used as terminal regret measures.

We do not survey the broad literature on bandits in detail. For foundational results on stochastic bandits and UCB-type algorithms, we refer interested readers to Lai and Robbins (1985), Auer et al. (2002); for linear contextual bandits, see Li et al. (2010), Chu et al. (2011), Abbasi-Yadkori et al. (2011), Chen et al. (2024); for posterior sampling, see Russo and Van Roy (2014). In the technical analysis, we build in particular on the information-theoretic analysis of Thompson sampling by Russo and Van Roy (2016) and the lifted information-ratio framework of Neu et al. (2022). Related work studies information-directed sampling (Russo and Van Roy 2018) and rate-distortion bounds for Thompson sampling (Dong and Van Roy 2018). Resent work in Hamidi and Bayati (2023) establish a posterior-covariance elliptical potential bound for linear bandits under general prior and noise distributions. Our analysis instead uses the unnormalized centered design of the expanding selected set to separate action uncertainty from parameter uncertainty, combining a geometric potential bound with a mutual information budget to address information dilution under Poisson–MNL feedback. We cite additional probabilistic results locally where they are used.

Online learning with choice-based feedback. The closest models are online assortment optimization and MNL bandits, where a decision-maker learns unknown choice parameters from customer choices on repeated assortments (Agrawal et al. 2019, 2026, Oh and Iyengar 2019, Chen et al. 2020, Han et al. 2026). Dueling and choice bandits also use preference-based feedback, but typically focus on identifying or competing with the best individual arm (Sui et al. 2018, Agarwal et al. 2020). Our setting differs in the decision structure. Existing assortment-bandit models choose a new assortment in each period and evaluate cumulative per-period reward. By contrast, our learner incrementally constructs one persistent active set: previously selected alternatives cannot be removed, and each selection continues to affect both future observations and final subset quality.

Irreversible and sequential subset construction. Another related stream studies bandit problems with irreversible or sequential subset decisions. Chen et al. (2017) study interactive submodular bandits, where the learner observes noisy marginal-gain feedback, and Korkut and Li (2021) study disposable linear bandits, where selected arms become unavailable. Chen (2021) studies continuum-armed bandits with monotone action sequences. These models capture aspects of sequential construction or irreversibility, but their feedback is directly tied to reward. In our model, feedback consists of choice counts generated by the entire active set, and the relationship between

observations and terminal subset quality is mediated by an unknown preference parameter. This active-set choice feedback makes learning path-dependent and prevents observed demand from being interpreted as standalone arm quality. Unlike the pure-exploration and best-arm identification settings in Bubeck et al. (2011), Russo (2020), we evaluate the entire set irreversibly constructed during learning.

Choice-based facility location. Choice-based facility-location models incorporate discrete-choice demand into location decisions, including MNL and random-utility formulations (Lin et al. 2022, Bravo et al. 2026, Lin et al. 2024, Qi et al. 2024, Legault and Frejinger 2025, Guo et al. 2025). These papers typically assume that choice primitives are known or estimated before optimization, and then solve a static or sequential location problem. Our problem instead integrates preference learning with irreversible sequential selection. Moreover, while market-share and revenue objectives are natural in static choice-based facility location, we show that they can be degenerate when used as terminal regret criteria for an irreversibly constructed active set.

Online sequential infrastructure decisions. Online facility-location models study persistent decisions under sequential demand, balancing facility-opening costs and service costs (Meyerson 2001). Recent prediction-augmented variants use forecasts of offline optimal assignments to improve online decisions (Almanza et al. 2021, Jiang et al. 2022). These models are related through persistence, but demand is typically exogenous and assigned to facilities. The work of Cao et al. (2026) is closer in combining sequential location decisions with demand learning, but locations can change across periods. Our setting models persistent assets such as stores, lockers, products, or partnerships: selected alternatives remain active, demand is generated by customer choice over the active set, and early decisions affect both future information and terminal subset quality.

3. Model

We first introduce a general irreversible subset-construction framework in Section 3.1. We then specialize the feedback process to Poisson arrivals and MNL choice with unknown linear utilities in Section 3.2. Section 3.3 defines the terminal objectives considered in the paper, and Section 3.4 introduces the full-information oracle benchmark and regret criteria.

3.1. General Subset Expansion, Observations, and Policies

Fix integers $N, d \geq 1$ and $1 \leq n \leq N$. Let $\mathcal{N} = \{1, \ldots, N\}$ be the candidate set, where each arm $i \in \mathcal{N}$ has a known feature vector $x_i \in \mathbb{R}^d$. The learner sequentially constructs an active set of n

distinct arms by selecting one new arm in each period. In the main no-delay setting, the construction horizon equals the target subset size n.

Let $S_t \subseteq \mathcal{N}$ denote the active set after period t, with $S_0 = \varnothing$. For a positive integer t, let $[t] := \{1, \dots, t\}$. In the main setting, $|S_t| = t$ for all $t \in [n]$. At period $t \in [n]$, let $\mathcal{A}_t := \mathcal{N} \setminus S_{t-1}$ denote the available arms. The learner selects $a_t \in \mathcal{A}_t$ and updates the active set according to $S_t = S_{t-1} \cup \{a_t\}$. The decision is irreversible: once an arm is selected, it remains active for all subsequent periods. After the update in period t, the learner observes arm-level feedback from the current active set, $\boldsymbol{y}_t = (y_{it})_{i \in S_t}$. Conditional on S_t, the observation at arm $i \in S_t$ has mean $\mathbb{E}[y_{it} \mid S_t] = \mu_i(S_t)$, and we write $\boldsymbol{\mu}(S) = (\mu_i(S))_{i \in S}$ for the vector of expected observations under active set S. The dependence of $\mu_i(S)$ on the full active set captures the active-set feedback feature of the model: the observation associated with an arm may change when other arms are added.

Performance is evaluated through a real-valued terminal set function $F : 2^{\mathcal{N}} \to \mathbb{R}$. The objective $F(S)$ need not coincide with the sum of the arm-level observations or their expectations. Thus, the learner observes active-set-dependent arm-level feedback while being evaluated by the quality of the constructed subset. This separation between feedback and evaluation is central to our setting.

Let $\mathbf{H}_t := \big((a_1, \boldsymbol{y}_1), \dots, (a_t, \boldsymbol{y}_t)\big)$ denote the observed history through period t, and let $\mathcal{H}_t := \sigma(\mathbf{H}_t)$ be the corresponding sigma-field, with $\mathcal{H}_0$ trivial. A policy $\pi = (\pi_t)_{t=1}^n$ is adapted to $(\mathcal{H}_t)_{t=0}^n$ if, for each $t \in [n]$, the action rule π_t is $\mathcal{H}_{t-1}$-measurable. Given $\mathbf{H}_{t-1}$, the distribution $\pi_t(\cdot \mid \mathbf{H}_{t-1})$ is supported on the available set $\mathcal{N} \setminus S_{t-1}$. Under policy π, write a_t^π, $\boldsymbol{y}_t^\pi$, and S_t^π for the induced action, observation, and active-set processes. The action satisfies $a_t^\pi \mid \mathcal{H}_{t-1}^\pi \sim \pi_t(\cdot \mid \mathbf{H}_{t-1}^\pi)$, while the active set evolves as $S_0^\pi = \varnothing$ and $S_t^\pi = S_{t-1}^\pi \cup \{a_t^\pi\}$. Let Π_n denote the class of such policies. When a single policy is under discussion, we suppress the superscript π.

3.2. Poisson–MNL Feedback Model

We now specify the active-set-dependent feedback model. In each period t, feedback opportunities arrive according to $N_t \sim \text{Poisson}(\lambda)$, independently across periods, where $\lambda > 0$ is the unknown arrival rate. Each opportunity is exposed to the current active set S_t and generates a choice of either an active alternative or the outside alternative 0. For opportunity $k = 1, \dots, N_t$, the latent utility of alternative $i \in S_t \cup \{0\}$ is $U_{kit} = V_i(\theta) + \epsilon_{kit}$, where the noise terms ϵ_{kit} are i.i.d. Gumbel and

$V_i(\theta) = x_i^\top \theta$, with unknown preference parameter $\theta \in \mathbb{R}^d$. The utility $V_0(\theta)$ is normalized to 0. The choice outcome is $c_{kt} \in \arg\max_{i \in S_t \cup \{0\}} U_{kit}$. Under the MNL model, the choice probabilities are

$$\begin{cases} p_i(S_t;\theta) = \mathbb{P}(c_{kt} = i \mid S_t) = \dfrac{\exp(x_i^\top \theta)}{1 + \sum_{j \in S_t} \exp(x_j^\top \theta)}, \ i \in S_t, \text{ and} \\ p_0(S_t;\theta) = \mathbb{P}(c_{kt} = 0 \mid S_t) = \dfrac{1}{1 + \sum_{j \in S_t} \exp(x_j^\top \theta)}. \end{cases}$$

The learner observes the inside-choice counts $\boldsymbol{y}_t = (y_{it})_{i \in S_t}$, where $y_{it} := \sum_{k=1}^{N_t} \mathbb{I}\{c_{kt} = i\}$, $i \in S_t$. Neither the arrival count N_t nor the outside-option count $y_{0t} := \sum_{k=1}^{N_t} \mathbb{I}\{c_{kt} = 0\}$ is observed. By Poisson thinning, the full count vector $\{y_{it}\}_{i \in S_t \cup \{0\}}$ has mutually independent components with $y_{it} \sim \text{Poisson}\big(\lambda p_i(S_t;\theta)\big)$, $i \in S_t \cup \{0\}$. Thus the expected observed count for arm $i \in S_t$ is $\mu_i(S_t;\theta) = \lambda p_i(S_t;\theta)$. Define the inside-choice probability $p_{\text{in}}(S;\theta) := \sum_{i \in S} p_i(S;\theta) = 1 - p_0(S;\theta)$, and the inside count $M_t := \sum_{i \in S_t} y_{it}$. We distinguish the full MNL probabilities $p_i(S;\theta)$ from the conditional ones over a subset $A \subseteq S$, and define

$$q_i(A;\theta) := \frac{\exp(x_i^\top \theta)}{\sum_{j \in A} \exp(x_j^\top \theta)}, \qquad i \in A,$$

and write $q(A;\theta) := (q_i(A;\theta))_{i \in A}$. The probabilities $p_i(S;\theta)$ determine the Poisson intensities of the observed counts. The conditional probabilities $q_i(A;\theta)$ are used when conditioning on the event that an observed inside choice falls in a fixed reference set A, and are independent of the alternatives in $S \setminus A$ by the MNL IIA property.

Remark 1 *Although N_t and y_{0t} are unobserved, the inside total $M_t = \sum_{i \in S_t} y_{it}$ is observed, and $M_t \mid S_t, \lambda, \theta \sim \text{Poisson}\big(\lambda p_{\text{in}}(S_t;\theta)\big)$. Conditional on this total, $\boldsymbol{y}_t \mid M_t, S_t, \theta \sim \text{Multinomial}\big(M_t, q(S_t;\theta)\big)$, which does not involve λ. Thus, a procedure concerned only with θ may use the conditional composition of the inside counts without estimating λ. The exact Bayesian update studied here instead retains both factors of the observed-data likelihood and therefore updates the joint distribution of (λ, θ). For a nonempty active set S and an observed inside-count vector $\boldsymbol{y} = (y_i)_{i \in S}$, let $m = \sum_{i \in S} y_i$. The full observed-data likelihood is*

$$\begin{aligned} L_S(\lambda, \theta; \boldsymbol{y}) &= \prod_{i \in S} \frac{\exp\{-\lambda p_i(S;\theta)\}[\lambda p_i(S;\theta)]^{y_i}}{y_i!} \\ &= \frac{\exp\{-\lambda p_{\text{in}}(S;\theta)\}[\lambda p_{\text{in}}(S;\theta)]^m}{m!} \frac{m!}{\prod_{i \in S} y_i!} \prod_{i \in S} q_i(S;\theta)^{y_i}. \end{aligned} \tag{1}$$

3.3. Objectives under the Poisson–MNL Model

We evaluate policies by the terminal reward of the active set constructed at the end of the expansion process. The feedback process and the reward objective need not coincide: the learner observes arm-level counts generated by the Poisson–MNL feedback model, while performance is evaluated through a set-level objective. We consider three terminal reward objectives.

1. A *market-share* objective (see, e.g., Ben-Porat and Tennenholtz 2019, Qi et al. 2024, Legault and Frejinger 2025, Zhu et al. 2026), which is the expected number of feedback opportunities captured by the alternatives in S per period:

$$F^{MS}(S;\lambda,\theta) = \lambda \sum_{i\in S} p_i(S;\theta) = \lambda\big(1 - p_0(S;\theta)\big). \tag{2}$$

2. An *expected-revenue* objective (see, e.g., Talluri and van Ryzin 2004, Rusmevichientong and Topaloglu 2012, Cao et al. 2023), where each arm i has a known per-outcome revenue r_i:

$$F^{rev}(S;\lambda,\theta) = \lambda \sum_{i\in S} r_i\, p_i(S;\theta). \tag{3}$$

This objective is economically meaningful as a per-period operational payoff, but unlike the market-share objective in (2), F^{rev} is generally neither monotone nor submodular because of substitution and cannibalization effects under the MNL model.

3. An *intrinsic portfolio-quality* objective, based on the additive attraction primitive used in MNL and attraction-based choice models (see, e.g., Gallego et al. 2015, El Housni et al. 2026):

$$F^{IQ}(S;\lambda,\theta) = \sum_{i\in S} \exp\left(V_i(\theta)\right). \tag{4}$$

This objective measures the total latent attraction mass of the selected alternatives. It is not normalized by the total attraction of the active set and therefore does not internalize within-set substitution.

In the proposition below, we show that F^{MS} and F^{IQ} are not two preference orders, but one preference order expressed on two different numerical scales.

Proposition 1 (Ordinal Equivalence of F^{MS} and F^{IQ}) *Let $\phi_\lambda(y) := \lambda y/(1+y)$ for $y \geq 0$. Then ϕ_λ is strictly increasing on $[0,\infty)$, and*

$$F^{MS}(S;\lambda,\theta) = \phi_\lambda\big(F^{IQ}(S;\theta)\big) \qquad \text{for every } S \subseteq \mathcal{N}.$$

Consequently, the two objectives induce the same preference order on subsets: for $S, S' \subseteq \mathcal{N}$,

$$F^{MS}(S;\lambda,\theta) \leq F^{MS}(S';\lambda,\theta) \iff F^{IQ}(S;\theta) \leq F^{IQ}(S';\theta).$$

Proposition 2 (Optimality of Top-n Selection for F^{MS} and F^{IQ}) *Order the arms such that $V_{i_1}(\theta) \geq V_{i_2}(\theta) \geq \cdots \geq V_{i_N}(\theta)$. Then the top-n set $S^{\star,n} = \{i_1, \ldots, i_n\}$ is optimal for F^{MS} and F^{IQ}.*

Condition 1 *The function $F(\cdot)$ is monotone and submodular on $2^{\mathcal{N}}$.*

Remark 2 *For any objective F satisfying Condition 1, the greedy algorithm achieves a $(1 - \frac{1}{e})$-approximation guarantee (Nemhauser et al. 1978).*

3.4. Regret and Full-Information Oracle Benchmark

We consider an incomplete-information setting in which the learner does not know θ a priori. For a given terminal objective F, the full-information oracle selects the best size-n subset under the true parameter. Let $S_F^{\star,n} \in \arg\max_{S \subseteq \mathcal{N}, |S|=n} F(S; \lambda, \theta)$ denote this oracle subset, where F can be F^{MS}, F^{rev}, F^{IQ}, or a more general objective. For a policy π, recall that S_n^π is the random active set produced after n selections. The terminal regret under objective F is defined as

$$\mathrm{Regret}_n^F(\pi; \lambda, \theta) = F(S_F^{\star,n}; \lambda, \theta) - \mathbb{E}_\pi\left[F(S_n^\pi; \lambda, \theta)\right]. \tag{5}$$

When (λ, θ) is sampled from a known prior G_1, the Bayesian regret is

$$\mathrm{BayesRegret}_n^F(\pi) = \mathbb{E}_{(\lambda,\theta)\sim G_1}\left[\mathrm{Regret}_n^F(\pi; \lambda, \theta)\right]. \tag{6}$$

4. Objective Degeneracy and Spread

The terminal objective determines whether regret is an informative measure of policy performance. In our irreversible subset-construction setting, this choice is more subtle than in repeated assortment or one-period choice models. As the subset size grows, saturation and cannibalization can make high-quality and poorly constructed subsets appear nearly indistinguishable under terminal regret.

We formalize this issue by defining *objective-induced degeneracy* in Section 4.2. The key quantity is the spread of an objective over feasible active sets of a fixed size. If this spread is sublinear, then even the worst terminal regret is sublinear, so the regret criterion cannot meaningfully distinguish learning policies at the scale of the construction problem. Our analysis does not seek to select the economic objective, which is taken as exogenously given, but instead asks whether using that objective directly as the terminal learning criterion records learning losses on an informative scale. We first identify exactly what the flow criterion's scale factor is for the three objective functions, F^{MS}, F^{rev}, and F^{IQ} in Section 4.3. We show that the standard choice-based objectives F^{MS} and F^{rev} are degenerate in this sense, while the intrinsic-quality objective F^{IQ} remains sensitive to the latent quality of selected alternatives. Thus, we focus on F^{IQ} in the following sections. We then show that the same mechanism operates across a general class of objectives in Section 4.4.

4.1. Key Assumptions

Throughout this section, asymptotic statements are taken along a sequence of instances indexed by n, where n is the target subset size and the candidate pool $\mathcal{N}_n$ satisfies $|\mathcal{N}_n| \geq n$. The candidate pool grows with the subset size. To simplify notation, we write $\mathcal{N}$ when the dependence on n is clear. All definitions and bounds in this section are stated under true parameters (λ, θ). Let $\Theta \subseteq \mathbb{R}^d$ denote the preference parameter space, and assume that $\theta \in \Theta$. Assume $\lambda \in (0, \infty)$. In Bayesian statements, the prior G_1 is supported on $(0, \infty) \times \Theta$.

Assumption 1 (Bounded Features and Parameters) *The feature vectors and parameter space satisfy* $\|x_i\|_2 \leq 1$ *for all* $i \in \mathcal{N}$*, and* $\|\theta'\|_2 \leq 1$ *for all* $\theta' \in \Theta$*, and* $\lambda' \in [\underline{\lambda}, \overline{\lambda}]$ *with* $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$*. In particular, for the true* $\theta \in \Theta$*, the systematic utility satisfies* $V_i(\theta) = x_i^\top \theta \in [-1, 1]$ *for all* $i \in \mathcal{N}$*.*

Assumption 2 (Bounded Revenues) *There exists a constant* $\overline{r} < \infty$ *such that* $0 \leq r_i \leq \overline{r},\ i \in \mathcal{N}$*.*

4.2. Objective-Induced Degeneracy and Spread

The first key quantity is the spread of the objective over feasible active sets of size n, which measures the largest possible objective gap between high- and low-quality constructed subsets.

Definition 1 (Spread) *For an objective* F*, the size-*n *spread at parameter* (λ, θ) *is*

$$\bar{R}_n^F(\lambda, \theta) := \max_{S \subseteq \mathcal{N}_n, |S|=n} F(S; \lambda, \theta) \ - \min_{S \subseteq \mathcal{N}_n, |S|=n} F(S; \lambda, \theta).$$

Lemma 1 (Spread Bounds Worst-Case Regret) *For any objective* F*, parameter* θ*, and policy* π*,*

$$0 \ \leq \ \mathrm{Regret}_n^F(\pi; \lambda, \theta) \ \leq \ \bar{R}_n^F(\lambda, \theta), \ \textit{and consequently,} \ \sup_{\pi \in \Pi} \mathrm{Regret}_n^F(\pi; \lambda, \theta) \ \leq \ \bar{R}_n^F(\lambda, \theta).$$

Proof. By definition, $S_F^{\star,n}$ maximizes $F(S; \lambda, \theta)$ over all feasible active sets of size n. Since S_n^π is also a feasible size-n active set, we have

$$\min_{S \subseteq \mathcal{N}_n,\ |S|=n} F(S; \lambda, \theta) \ \leq \ F(S_n^\pi; \lambda, \theta) \ \leq \ F(S_F^{\star,n}; \lambda, \theta)$$

almost surely. Taking expectations and subtracting from $F(S_F^{\star,n}; \lambda, \theta)$ gives the result. □

By Lemma 1, spread reflects the scale on which regret under an objective function F can be read. We refer to this scale as the *resolution* of F, and say that F has *full resolution* when its spread grows linearly in n, so that one unit of objective value is available per irreversible decision. We then define degeneracy as a property of a terminal objective, classifying objectives by the growth rate of their spread.

Definition 2 (Objective Degeneracy) *Fix a sequence of instances indexed by n. Using the spread* $\bar{R}_n^F(\lambda,\theta)$ *in Definition 1, we say that*

- *the terminal objective* F *is* ***degenerate*** *if*

$$\lim_{n\to\infty} \frac{\bar{R}_n^F(\lambda,\theta)}{n} = 0;$$

- *a degenerate objective* F *is* ***strongly degenerate*** *if there exists a constant* $C < \infty$ *such that*

$$\limsup_{n\to\infty} \bar{R}_n^F(\lambda,\theta) \le C;$$

- *a degenerate objective* F *is* ***weakly degenerate*** *if it is degenerate but not strongly degenerate; equivalently,*

$$\lim_{n\to\infty} \frac{\bar{R}_n^F(\lambda,\theta)}{n} = 0 \quad \textit{and} \quad \limsup_{n\to\infty} \bar{R}_n^F(\lambda,\theta) = \infty.$$

A degenerate objective is uninformative because the terminal regret of every policy is sublinear in n under that objective. Thus, degeneracy reflects a lack of discriminating power in the terminal objective, not the effectiveness of a specific algorithm.

4.3. Degeneracy and Nondegeneracy of the Three Poisson–MNL Objectives

We now apply the spread-based definition to the three Poisson–MNL objectives introduced in Section 3.3. The market-share and revenue objectives are economically meaningful as one-period operational quantities, but their terminal regret can be ineffective in irreversible subset construction. We show that F^{MS} and F^{rev} are degenerate, whereas F^{IQ} is nondegenerate under a natural separation condition on the candidate pool. The intrinsic-quality objective avoids degeneracy because its spread grows linearly with the subset size. For this reason, we use regret under F^{IQ} as the main performance criterion in the learning-theoretic analysis in Sections 5 and 6.

4.3.1. Degeneracy of F^{MS} The market-share objective saturates quickly as the cardinality of the active set grows. Consequently, the classical terminal regret criterion fails to characterize the effectiveness of a learning algorithm. Specifically, under Assumption 1, for any candidate $i \in \mathcal{N}$, $V_i(\theta) = x_i^\top \theta \in [-1,1]$, and hence $e^{-1} \le \exp(V_i(\theta)) \le e$. For any active set S with $|S| = n$, this implies

$$\lambda\left(1 - \frac{1}{1+ne^{-1}}\right) \;\le\; F^{MS}(S;\lambda,\theta) \;\le\; \lambda\left(1 - \frac{1}{1+ne}\right).$$

Therefore,

$$\bar{R}_n^{F^{MS}}(\lambda,\theta) \;\le\; \lambda\left(\frac{1}{1+ne^{-1}} - \frac{1}{1+ne}\right) \;=\; \lambda\frac{e-1/e}{n+1/n+e+1/e}.$$

Thus, the spread of the market-share objective is $O(1/n)$. In particular, F^{MS} is strongly degenerate; in fact, the worst-case regret of any policy converges to zero.

Proposition 3 *Under Assumption 1, the market-share objective satisfies*

$$\sup_{\pi\in\Pi} \mathrm{Regret}_n^{F^{MS}}(\pi;\lambda,\theta) \;\le\; \bar{R}_n^{F^{MS}}(\lambda,\theta) \;\le\; \overline{\lambda}\left(\frac{1}{1+ne^{-1}} - \frac{1}{1+ne}\right) \;=\; O\left(\frac{1}{n}\right).$$

Consequently, F^{MS} is strongly degenerate, with spread converging to zero.

4.3.2. Degeneracy of F^{rev} We next show that the expected-revenue objective (3) is also strongly degenerate. Writing $w_i := \exp(V_i(\theta))$ for simplicity, we have

$$F^{rev}(S;\lambda,\theta) = \lambda \frac{\sum_{i\in S} r_i w_i}{1+\sum_{j\in S} w_j}.$$

Under Assumptions 1 and 2, we have $e^{-1} \le w_i \le e$ and $0 \le r_i \le \overline{r}$. Thus, for any set $S \subseteq \mathcal{N}$, $|S| = n$,

$$0 \;\le\; F^{rev}(S;\lambda,\theta) \;\le\; \lambda \frac{n\overline{r}e}{1+ne^{-1}} \;\le\; \lambda \overline{r} e^2.$$

As a result, the spread of F^{rev} is uniformly bounded:

$$\bar{R}_n^{F^{rev}}(\lambda,\theta) \;\le\; \overline{\lambda}\overline{r}e^2.$$

Therefore, the gap between the best and worst feasible size-n active sets does not grow with n. In the terminology of Definition 2, the expected-revenue objective is strongly degenerate.

Proposition 4 *Under Assumption 1 and Assumption 2, the expected-revenue objective satisfies*

$$\sup_{\pi\in\Pi} \mathrm{Regret}_n^{F^{rev}}(\pi;\lambda,\theta) \;\le\; \bar{R}_n^{F^{rev}}(\lambda,\theta) \;\le\; \overline{\lambda}\overline{r}e^2.$$

Consequently, F^{rev} is strongly degenerate.

The natural candidate for a nondegenerate revenue criterion is the revenue attraction $R(S;\theta) := \sum_{i\in S} r_i w_i(\theta)$, which is modular and of full resolution. Its maximizer, however, is the top-n set by $r_i w_i(\theta)$, which is not the F^{rev}-maximizer, so adopting it changes the decision problem rather than only the scale. Quantifying that discrepancy and resolving the dilution are left as open questions.

Remark 3 *This degeneracy problem arises in our subset-construction problem but not in standard online assortment optimization. Previous work on online assortment optimization also adopts the expected-revenue objective to evaluate learning algorithms (Agrawal et al. 2019, 2026). The key distinction lies in the nature of the decision. In online assortment optimization, the learner chooses a new assortment in each period, whereas in our setting the learner incrementally constructs one*

active set over the entire horizon. Under the MNL model, the expected revenue of any assortment is uniformly bounded, and so the per-period regret is also bounded. Since regret in online assortment optimization is incurred period by period and accumulated over time, this boundedness does not make the regret criterion vacuous. In contrast, our long-term regret is evaluated only on the terminal constructed active set. Consequently, uniformly bounded objectives lead to strong degeneracy.

4.3.3. Nondegeneracy of F^{IQ} The intrinsic-quality objective behaves differently. Let $w_i(\theta) := \exp(V_i(\theta))$. Then $F^{IQ}(S;\lambda,\theta) = \sum_{i\in S} w_i(\theta)$ is a modular function. For a fixed size n, let $\overline{S}_n^{IQ}$ and $\underline{S}_n^{IQ}$ denote the size-n active sets with the largest and smallest values of F^{IQ}, respectively. Define

$$\Delta_n^{IQ}(\theta) := \sum_{i\in\overline{S}_n^{IQ}} w_i(\theta) - \sum_{i\in\underline{S}_n^{IQ}} w_i(\theta).$$

Assumption 3 (Linear Attraction Separation) *There exists a constant $\gamma > 0$ such that, along the sequence of instances, $\Delta_n^{IQ}(\theta) \geq \gamma n$.*

Theorem 1 *Under Assumption 3, $\bar{R}_n^{F^{IQ}}(\lambda,\theta) \geq \gamma n$ and F^{IQ} is nondegenerate.*

4.4. Degeneracy for a General Class of Objectives

The contrast between F^{MS} and F^{IQ} suggests a general mechanism behind objective-induced degeneracy. Both objectives are functions of the total latent attraction of the selected set, but they apply different outer transformations to this total. Market share uses a saturating transformation, whereas F^{IQ} uses a linear one. We now formalize this mechanism for a general separable-weight objective class. This class contains F^{MS} and F^{IQ} as special cases, and its assumptions are general analogues of the bounded-weight and separation conditions used above. The class in Condition 2 is monotone and submodular, and therefore satisfies Condition 1 (Lovász 1983).

Condition 2 (Separable-Weight Objective Class) *The objective has the form $F_g(S) = g\left(\sum_{i\in S} w_i\right)$, where $w_i \geq 0$ are candidate weights, and $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is nondecreasing, concave, and differentiable on $(0,\infty)$.*

Remark 4 *In the MNL specialization, the weights are $w_i = \exp(x_i^\top \theta)$. The market-share objective corresponds to $g(y) = \lambda y/(1+y)$, while the intrinsic-quality objective corresponds to $g(y) = y$. The revenue objective with heterogeneous revenues is not generally in this separable-weight class; its degeneracy was handled separately in Section 4.3.2.*

For a fixed cardinality n, let $C_n := \{S \subseteq \mathcal{N} : \ |S| = n\}$. Define the largest and smallest feasible weight sums by

$$W_n^+ := \max_{S \in C_n} \sum_{i \in S} w_i, \qquad W_n^- := \min_{S \in C_n} \sum_{i \in S} w_i, \qquad \Delta_n := W_n^+ - W_n^- \geq 0.$$

For this subsection, write $\bar{R}_n := \bar{R}_n^{F_g}(\lambda, \theta)$ for the spread defined in Definition 1. Since g is nondecreasing, the spread of F_g is

$$\bar{R}_n = g(W_n^+) - g(W_n^-).$$

This representation separates the contribution of the candidate pool, captured by Δ_n, from the contribution of the objective function, captured by the slope of g.

Assumption 4 (General Bounded Weights) *There exist constants* $0 < w_{\min} \leq w_{\max} < \infty$ *such that* $w_{\min} \leq w_i \leq w_{\max}$ *for all* $i \in \mathcal{N}$.

Under MNL, Assumption 1 implies Assumption 4 with $w_{\min} = e^{-1}$ and $w_{\max} = e$. Thus, Assumption 4 is the general bounded-weight analogue of the bounded-utility condition used for the Poisson–MNL objectives. Under Assumption 4, $W_n^-, W_n^+ \in [nw_{\min}, nw_{\max}]$ and $\Delta_n \leq n(w_{\max} - w_{\min})$.

Assumption 5 (General Linear Separation Between Extremes) *There exists a constant* $\gamma > 0$ *such that, for all* n, $\Delta_n = W_n^+ - W_n^- \geq \gamma n$.

Assumption 5 is the general candidate-pool separation condition. When $w_i = \exp(V_i(\theta))$, it reduces to the linear attraction separation condition used for F^{IQ}. It requires the best and worst size-n active sets to differ by a constant average weight gap, and it fails when the candidate pool becomes asymptotically homogeneous.

Proposition 5 (Slope Window for Spread) *Under Condition 2 and Assumption 4,*

$$\Delta_n \, g'(nw_{\max}) \leq \bar{R}_n \leq \Delta_n \, g'(nw_{\min}). \tag{7}$$

The detailed proof is in Appendix B.1. Proposition 5 shows that the spread of F_g is controlled by two terms: the weight separation Δ_n, which is determined by the candidate pool, and the slope of g on the relevant scale $[nw_{\min}, nw_{\max}]$, which is determined by the objective function.

Theorem 2 (Spread Scaling) *Under Condition 2, Assumption 4, and Assumption 5,*

$$n\gamma \, g'(nw_{\max}) \leq \bar{R}_n \leq n(w_{\max} - w_{\min}) \, g'(nw_{\min}), \qquad n = 1, 2, \ldots. \tag{8}$$

Proof. The upper bound in (8) follows from Proposition 5 and $\Delta_n \le n(w_{\max} - w_{\min})$. The lower bound follows from (7) and Assumption 5. □

Theorem 2 shows that, under bounded weights and linear separation, the spread scales as $ng'(\Theta(n))$. Thus, objective-induced degeneracy is governed by the asymptotic slope of g. A nonvanishing slope yields linear spread and nondegenerate terminal regret; a slope of order $O(1/y)$ yields uniformly bounded spread and strong degeneracy; and intermediate slope decay yields weak degeneracy. We summarize these regimes in Figure 1.

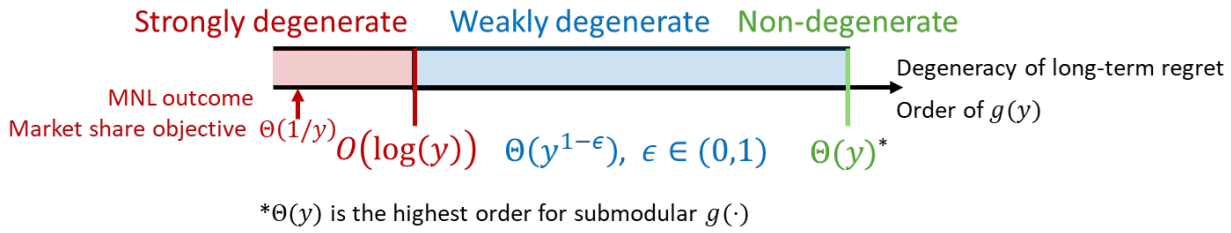


Figure 1 Degeneracy regimes for terminal regret under separable-weight objectives $F_g(S) = g(\sum_{i\in S} w_i)$. Under bounded weights and linear separation, Theorem 2 shows that objective spread scales as $ng'(\Theta(n))$. A nonvanishing asymptotic slope yields nondegeneracy, $g'(y) = O(1/y)$ yields strong degeneracy, and intermediate slope decay yields weak degeneracy. Formal regime characterizations are given in Appendix B.2.1–B.2.3.

Remark 5 *For F^{MS}, $g(y) = \lambda y/(1+y)$ and $g'(y) = \lambda/(1+y)^2 = \Theta(y^{-2})$. Therefore, under Assumption 5, Theorem 2 gives $\bar{R}_n = \Theta(n^{-1})$, consistent with Proposition 3. This explains the role of F^{IQ}: it preserves the latent attraction scale by taking $g(y) = y$, and therefore avoids the saturation effects that make normalized choice-based objectives degenerate as terminal regret criteria.*

We use F^{IQ} as the terminal objective throughout the rest of the paper. Accordingly, "regret" and "Bayesian regret" refer to the definitions in Section 3.4 evaluated under F^{IQ}.

5. Regret Lower Bound Under Intrinsic-Quality Objective

We now study the statistical difficulty of learning under the intrinsic-quality objective F^{IQ}. After the degeneracy analysis in Section 4, F^{IQ} serves as the nondegenerate terminal objective for evaluating the constructed subset. Even under this objective, the learner observes only noisy choice-based feedback generated by the active set, rather than direct observations of the latent attraction weights. The following lower bound shows that this indirect feedback creates an unavoidable learning cost.

Theorem 3 (Minimax Lower Bound for Intrinsic-Quality Regret) *Consider the Poisson–MNL feedback model and the objective F^{IQ} in (4). Fix integers $n, d \geq 1$. There exists a candidate pool $\mathcal{N}_{n,d}$ of $N = n\,2^d$ arms and a parameter space Θ such that Assumption 1 holds and*

$$\inf_{\pi \in \Pi_n} \sup_{\lambda \in [\underline{\lambda}, \overline{\lambda}]} \sup_{\theta \in \Theta} \mathrm{Regret}_n^{F^{\mathrm{IQ}}}(\pi; \lambda, \theta) \;\geq\; \frac{1}{32e^2} \min\left\{ n,\; d\sqrt{\frac{n}{\underline{\lambda}}} \right\}.$$

The proof is provided in Section C.1. It uses a hypercube construction for the unknown preference parameter and a Bretagnolle–Huber argument to lower bound the probability of selecting arms with incorrect coordinate signs. The key event bound is stated in Lemma 3 in Appendix C. The theorem shows that any policy can incur regret of order at least $d\sqrt{n/\underline{\lambda}}$ in the regime where $d \leq \sqrt{\underline{\lambda} n}$. Thus, regret necessarily increases with the feature dimension d and the square root of the terminal subset size n, while a larger arrival rate λ reduces the learning cost by providing more choice observations per period. The $1/\sqrt{\underline{\lambda}}$ dependence is consistent with the effective-sample-size interpretation common in demand-learning and MNL-bandit models (Wang et al. 2014, Ferreira et al. 2018, Agrawal et al. 2019, Oh and Iyengar 2021).

6. Algorithms under Intrinsic-Quality Objective

We analyze two learning policies under F^{IQ}. We first introduce an explore-then-exploit policy with a fixed-set estimator (ETE-FSE), which exploits repeated observations from an early active set and serves as a transparent frequentist benchmark. ETE-FSE attains an $O(n^{2/3})$ regret bound. Our main algorithmic result is for Thompson sampling, which uses posterior uncertainty to guide irreversible selections. For exact posterior Thompson sampling, we establish a Bayesian regret bound with $\widetilde{O}(\sqrt{n})$ dependence on the terminal subset size under stated regularity conditions. This matches the minimax lower bound of Theorem 3 in both its n- and its d-dependence, up to logarithmic factors.

6.1. Explore-then-Exploit with Fixed-Set Estimator (ETE-FSE)

To obtain a transparent benchmark, we use a fixed-set estimator based on the first m selected alternatives. By the IIA property of the MNL model, relative choice probabilities within S_m are unaffected by alternatives added later. ETE-FSE first explores uniformly for m periods, then selects greedily using the current fixed-set estimate. The estimator uses only observations on S_m from periods m onward and is updated after every period $t \geq m$, including during exploitation. The formal estimator and pseudocode are given in Appendix D.1.

Under bounded features and parameters and the i.i.d. isotropic-feature condition, Theorem 6 yields, with probability at least $1-\delta$ over the candidate features,

$$\mathrm{Regret}_n^{F^{IQ}}\left(\pi^{\text{ETE-FSE}};\lambda,\theta\right) \le C'\left(1+\frac{d}{\sqrt{\underline{\lambda}}}\right)n^{2/3},$$

for a constant $C'>0$, when $m=\lceil n^{2/3}\rceil$ satisfies its stated exploration-length requirement. The analysis relates greedy-exploitation regret to estimation error. The appendix establishes the m/d scaling of the centered exploration design's smallest eigenvalue, the fixed-set estimator's mean-square error bound, and the Bayesian regret counterpart; see Propositions 7 and 8 and Corollary 7.

The fixed-set construction stabilizes conditional choice probabilities, but does not eliminate information dilution: as the active set expands, a smaller share of arrivals selects S_m. This leaves a tradeoff between irreversible exploration and estimation error on the fixed design. We next turn to Thompson sampling, which uses feedback from the full active set and controls decision errors through a novel uncertainty decomposition based on the evolving selected-set geometry.

6.2. Thompson Sampling (TS)

We develop a Thompson sampling (TS) algorithm in Algorithm 1. TS maintains a posterior distribution of the parameter (λ,θ). In round t, TS first draws a sample $(\widetilde{\lambda}_t,\widetilde{\theta}_t)$ from the joint posterior G_t, and then acts greedily with respect to the sampled $\widetilde{\theta}_t$. Drawing only the θ-marginal is decision-equivalent, but the posterior update must remain joint and use the full likelihood (1). Given the observations up to the previous period $\mathcal{H}_{t-1}$, the sample and the true parameter are identically distributed, and consequently, the sampled and true optimal actions are identically distributed.

After introducing the notation in Section 6.2.1, we show how the expanding active set precludes a uniform constant information-ratio bound in Section 6.2.2. We then develop an action–parameter uncertainty decomposition in Section 6.2.3 and use separate geometric and information bounds to establish the Bayesian regret guarantee for exact Thompson sampling.

6.2.1. Notation We collect the notation used throughout the analysis. All model primitives retain the definitions given in the preceding sections.

Conditional laws and information. For each period t, write $\mathbb{P}_t(\cdot):=\mathbb{P}(\cdot\mid\mathcal{H}_{t-1})$, $\mathbb{E}_t[\cdot]:=\mathbb{E}[\cdot\mid\mathcal{H}_{t-1}]$, $\mathrm{Cov}_t(\cdot):=\mathrm{Cov}(\cdot\mid\mathcal{H}_{t-1})$. For random elements U and V, let $\mathbb{P}_t^U$ denote the conditional law of U under $\mathbb{P}_t$. Define the mutual information $I_t(U;V):=D_{\mathrm{KL}}\left(\mathbb{P}_t^{(U,V)}\,\middle\|\,\mathbb{P}_t^U\otimes\mathbb{P}_t^V\right)$, where D_{KL} denotes Kullback–Leibler divergence. The quantity $I_t(U;V)$ is $\mathcal{H}_{t-1}$-measurable. Since $\mathcal{H}_{t-1}=\sigma(\mathbf{H}_{t-1})$, its expectation satisfies $\mathbb{E}[I_t(U;V)]=I(U;V\mid\mathbf{H}_{t-1})$, where $I(\cdot;\cdot)$ denotes mutual information.

Algorithm 1 Thompson Sampling

Input: prior distribution G_1.

Initialize: $S_0 = \varnothing$, $\mathcal{H}_0 = \varnothing$, and prior G_1 on $[\underline{\lambda}, \overline{\lambda}] \times \Theta$.

for $t = 1, 2, \ldots, n$ **do**

Sample $(\widetilde{\lambda}_t, \widetilde{\theta}_t) \sim G_t$.

Select the next arm greedily w.r.t. F^{IQ} and $\widetilde{\theta}_t$:

$$a_t \in \operatorname*{arg\,max}_{i \in \mathcal{N} \setminus S_{t-1}} F^{IQ}(S_{t-1} \cup \{i\}; \widetilde{\theta}_t), \qquad S_t \leftarrow S_{t-1} \cup \{a_t\}.$$

Observe counts $y_{i,t}$ for all $i \in S_t$.

Update the posterior distribution $G_{t+1}(\lambda, \theta) \propto G_t(\lambda, \theta) L_{S_t}(\lambda, \theta; \boldsymbol{y}_{S_t})$.

end for

Policy-related quantities. Let $a_t^\star$ denote a greedy action under the true preference parameter θ from the current available set $\mathcal{A}_t = \mathcal{N} \setminus S_{t-1}$, with ties resolved according to the same convention as the Thompson action. Thompson sampling gives posterior action probabilities $\pi_{a,t} := \mathbb{P}_t(a_t = a) = \mathbb{P}_t(a_t^\star = a)$ for $a \in \mathcal{A}_t$. Let $\Sigma_t := \mathrm{Cov}_t(\theta)$. For $\pi_{a,t} > 0$, define the conditional posterior mean $m_t(a) := \mathbb{E}_t[\theta \mid a_t^\star = a]$. When $\pi_{a,t} = 0$, set $m_t(a) := \mathbb{E}_t[\theta]$; this convention does not affect any sum weighted by $\pi_{a,t}$. Define the instantaneous greedy gap and the one-step lifted information gain as

$$\Delta_t := \mathbb{E}_t\left[e^{x_{a_t^\star}^\top \theta} - e^{x_{a_t}^\top \theta}\right], \qquad \mathit{I}_t := I_t\big((\lambda, \theta); (a_t, \boldsymbol{Y}_t)\big).$$

The gap Δ_t is nonnegative. Moreover, under exact Thompson sampling, a_t is conditionally independent of (λ, θ); hence $I_t((\lambda, \theta); a_t) = 0$. The chain rule therefore gives $\mathit{I}_t = I_t\big((\lambda, \theta); \boldsymbol{Y}_t \mid a_t\big)$.

Centered-design geometry. For every nonempty set S, define its mean feature and normalized centered design matrix by

$$\bar{x}_S := \frac{1}{|S|} \sum_{i \in S} x_i, \qquad B_S := \frac{1}{|S|} \sum_{i \in S} (x_i - \bar{x}_S)(x_i - \bar{x}_S)^\top.$$

Along the selected path, write $\bar{x}_t := \bar{x}_{S_t}$, $B_t := B_{S_t}$, $\widetilde{B}_t := tB_t$. Because $|S_t| = t$, $\widetilde{B}_t$ is the unnormalized centered design matrix. In particular, $\widetilde{B}_1 = 0$. Throughout, I_d denotes the $d \times d$ identity matrix and $\mathrm{tr}(\cdot)$ denotes the matrix trace. For symmetric matrices A and B, $A \succeq B$ means that $A - B$ is positive semidefinite. For a positive-semidefinite matrix M, write $\|z\|_M^2 := z^\top M z$.

For $t \geq 2$, define

$$\ell_t := \mathrm{tr}\Big(\mathrm{Cov}_t(x_{a_t^\star})(4I_d + \widetilde{B}_{t-1})^{-1}\Big).$$

This quantity is the posterior elliptical leverage of the period-t greedy action. It measures posterior disagreement about the action while discounting directions well-explored by the selected design. Here the design records selected-feature geometry, not a Fisher information matrix accumulated from historical count observations.

For an active set of size $s \geq 1$, define $b_s := 1 - \exp\left(-\frac{\lambda s}{e+s}\right)$, and $\kappa_0 := \frac{1-\exp[-1/(2e^2)]}{2}$.

6.2.2. Information Dilution and the Information Ratio A natural question is whether the elegant information-theoretic argument of Russo and Van Roy (2016) can yield a near-optimal regret guarantee in the subset-expansion setting. Following the lifted formulation of Neu et al. (2022), define the information ratio as

$$\Gamma_t := \Delta_t^2 / \mathcal{I}_t.$$

If there exists $\overline{\Gamma} > 0$ such that $\Gamma_t \leq \overline{\Gamma}$ almost surely for every $2 \leq t \leq n$, then Cauchy–Schwarz gives

$$\mathrm{BayesRegret}_n^{F^{IQ}}(\pi^{\mathrm{TS}}) \leq e - e^{-1} + \sqrt{(n-1)\overline{\Gamma}\, I((\lambda,\theta);\mathbf{H}_n)}.$$

Russo and Van Roy (2016) provide uniform upper bounds on the information ratio for classical problems, such as independent finite-arm and linear bandits.

However, without additional assumptions on the instance class, such as structural properties of the candidate set, no constant bound $\overline{\Gamma}$ can hold uniformly over time and candidate sets in subset expansion with MNL feedback. Specifically, for each prescribed period $t \geq 2$, Proposition 6 establishes an $\Omega(t)$ bound on Γ_t for the corresponding instance in Example 1. The growing active set drives the example: MNL normalizes the informative entrant by the attraction of the expanding active set, whereas the greedy gap remains independent of the subset size.

Example 1 (Information dilution by the expanding active set) *Fix $d = 2$ and $t \geq 2$, and choose $0 < r < 3/8$. Let the prior on λ be degenerate at a fixed $\lambda > 0$, and let the preference prior assign probability $1/2$ to each of $\theta_+ = (1/2, r)$ and $\theta_- = (1/2, -r)$. The candidate set contains*

- *comparison candidates i_+ and i_-, with $x_{i_+} = (0,1)$ and $x_{i_-} = (0,-1)$*
- *$t-1$ baseline candidates with $x_{v_j} = (c_j, 0)$, $1 \geq c_1 > \cdots > c_{t-1} \geq \frac{3}{4}$, and*
- *$N - t - 1$ dominated candidates with $x_i = (-\eta_i, 0)$, $\eta_i \in (0,1)$.*

Under either prior atom, baseline candidate v_j has utility $c_j/2 \geq 3/8$; dominated candidates have negative utility, whereas the comparison utilities are r and $-r$. Exact Thompson sampling therefore selects $v_1, \ldots, v_{t-1}$ in this exact order with probability 1 during the first $t-1$ periods. Every active

feature in this prefix has zero second coordinate, so the likelihood of every feedback realization is identical under θ_+ and θ_-. The posterior remains uniform through period $t-1$.

At period t, candidate i_+ is greedy under θ_+, whereas candidate i_- is greedy under θ_-. Thompson sampling will not select a dominated candidate. The true preference parameter and the Thompson draw are conditionally independent and uniform over the two atoms. They disagree with probability $1/2$, and a disagreement produces an instantaneous gap $e^r - e^{-r}$. Hence $\Delta_t = (e^r - e^{-r})/2$. Meanwhile, the expanding active set dilutes the information. The total attraction of the $s = t-1$ baselines is $W_s = \Omega(s)$. The probability that an arrival chooses the comparison candidate is therefore $O(1/s)$. Conditional on not choosing the comparison candidate, the choice is sign-invariant. Even the augmented feedback provides only $O(1/s)$ information about the preference direction. Proposition 6 quantifies how the expanding active set dilutes the information available at this decision, with proof provided in Appendix F.

Proposition 6 (Lower bound for the lifted information ratio) *For Example 1,*

$$\Gamma_t \geq \frac{(t-1)e^{3/8}(e^r - e^{-r})}{2\lambda r}, \qquad 0 < r < \frac{3}{8}. \tag{9}$$

Consequently, no bound of the form $\Gamma_t \leq \overline{\Gamma}$ can hold uniformly over time and candidate sets.

This example rules out a constant bound on the lifted information ratio that is uniform over time and candidate sets. We instead separate action uncertainty from parameter uncertainty and control the two factors separately over time.

6.2.3. Action–Parameter Uncertainty Decomposition We first use Lemma 7 to upper bound the Bayesian regret with the expected sum of instantaneous greedy gaps:

$$\mathrm{BayesRegret}_n^{FIQ}(\pi^{\mathrm{TS}}) \leq \sum_{t=1}^{n} \mathbb{E}[\Delta_t]. \tag{10}$$

Thus, the terminal-regret analysis reduces to controlling the instantaneous greedy gaps.

Lemma 2 (Action–parameter uncertainty decomposition) *Under Assumption 1, exact Thompson sampling satisfies, for every $t \geq 2$,*

$$\Delta_t^2 \leq e^2 \ell_t \,\mathrm{tr}\Big((4I_d + \widetilde{B}_{t-1})\,\mathrm{Cov}_t(m_t(a_t^\star))\Big). \tag{11}$$

In particular,

$$\begin{aligned} \Delta_t^2 &\leq e^2 \ell_t \,\mathrm{tr}\Big((4I_d + \widetilde{B}_{t-1})\Sigma_t\Big) \\ &\leq e^2 \ell_t \left(4 + \frac{t\mathcal{I}_t}{\kappa_0 b_t}\right). \end{aligned} \tag{12}$$

The proof is presented in Appendix E.2.

Remark 6 *The two factors use inverse and direct weights of the same unnormalized centered design. The action factor measures posterior disagreement about which alternative to add, whereas* $\mathrm{Cov}_t(m_t(a_t^\star))$ *captures variation in posterior mean preferences across greedy actions. The decomposition does not assert independence between these uncertainties. Probability matching links the action factor to growth of the selected-set design, while the information inequality controls the design-weighted parameter factor; see Lemmas 5 and 6. Information dilution remains present through the factor* t*; the analysis avoids requiring a uniform per-period information-ratio bound by controlling the two factors separately over time.*

The probability-matching and covariance arguments are related to the linear Thompson-sampling analysis of Hamidi and Bayati (2023). Here, the unnormalized centered design permits separate geometric and information controls despite dilution in active-set feedback. The probability-matching and matrix Cauchy–Schwarz steps do not use the Poisson–MNL likelihood, whereas the conversion of parameter uncertainty into information gain depends on the observation model. An extension to another setting would require an analogous greedy-gap factorization, a design update admitting a cumulative leverage bound, and compatible one-step and total information bounds. Thus, the decomposition suggests an analysis strategy beyond the present model, rather than establishing a regret guarantee for other feedback models.

We next establish the cumulative bounds needed to translate Lemma 2 into a regret guarantee. By Cauchy–Schwarz, it suffices to control $\sum_{t=2}^{n}\mathbb{E}[\ell_t]$ and $\sum_{t=2}^{n}\mathbb{E}[4+t\mathcal{I}_t/(\kappa_0 b_t)]$ separately, using selected-set geometry and a mutual information budget, respectively.

Action uncertainty. Probability matching links the posterior action covariance to the selected features. The rank-one update of the unnormalized centered design in Lemma 5 then permits the elliptical-potential control in Lemma 8:

$$\sum_{t=2}^{n}\mathbb{E}[\ell_t] \le \frac{d}{\log(3/2)}\log\left(1+\frac{n}{4d}\right). \tag{13}$$

Parameter uncertainty. The mutual information chain rule gives

$$\sum_{t=2}^{n}\mathbb{E}[I_t] \le I((\lambda,\theta);\mathbf{H}_n). \tag{14}$$

Lemma 9 bounds the right-hand side through a joint-parameter covering argument.

6.2.4. Bayesian Regret Bound The decomposition separates the greedy-gap analysis into a cumulative leverage bound and an information budget. Combining these controls yields the following Bayesian regret guarantee for exact Thompson sampling.

Theorem 4 (Bayesian regret of Thompson sampling) *Let* $Q_n := \max\{1, \sqrt{en\bar{\lambda}}\}$. *Under Assumption 1, for* $2 \le n \le N$,

$$\mathrm{BayesRegret}_n^{F^{IQ}}(\pi^{\mathrm{TS}}) \le \min\left\{ n(e-e^{-1}),\ e-e^{-1} + e\left(\frac{d}{\log(3/2)}\log\left(1+\frac{n}{4d}\right)\right.\right. \\ \left.\left. \times\left[4(n-1) + \frac{n}{\kappa_0 b_n}\left\{ d\log(1+8Q_n) + \log\left(1+2Q_n\log\frac{\bar{\lambda}}{\underline{\lambda}}\right) + \frac{1}{2}\right\}\right]\right)^{1/2}\right\}. \tag{15}$$

Corollary 1 *For* $n \ge 2$, *the bound in Theorem 4 is*

$$\mathrm{BayesRegret}_n^{F^{IQ}}(\pi^{\mathrm{TS}}) = \widetilde{O}\left(d\sqrt{\frac{n}{\min\{\underline{\lambda}, 1\}}}\right), \tag{16}$$

where the tilde suppresses only the logarithms in (15).

Corollary 1 demonstrates that the Bayes regret of Thompson sampling matches the minimax lower bound in subset size and feature dimension, up to logarithmic factors. The proof of Corollary 1 is presented in Section E.3.

Proof of Theorem 4. By (10), it suffices to bound the sum of the expected instantaneous greedy gaps. Under Assumption 1, systematic utility lies in $[-1, 1]$, so the first-period greedy gap is at most $e - e^{-1}$. For $t \ge 2$, Lemma 2 bounds each instantaneous greedy gap by the action and information factors. Applying Cauchy–Schwarz first on each history space and then over time gives

$$\begin{aligned}
\sum_{t=2}^{n}\mathbb{E}[\Delta_t] &\le e\sum_{t=2}^{n}\mathbb{E}\left[\sqrt{\ell_t\left(4+\frac{t\mathcal{I}_t}{\kappa_0 b_t}\right)}\right] && \text{(Lemma 2)}\\
&\le e\sum_{t=2}^{n}\sqrt{\mathbb{E}[\ell_t]\cdot\mathbb{E}\left[4+\frac{t\mathcal{I}_t}{\kappa_0 b_t}\right]} && \text{(Cauchy–Schwarz)}\\
&\le e\left(\sum_{t=2}^{n}\mathbb{E}[\ell_t]\right)^{1/2}\left\{\sum_{t=2}^{n}\mathbb{E}\left[4+\frac{t\mathcal{I}_t}{\kappa_0 b_t}\right]\right\}^{1/2}. && \text{(Cauchy–Schwarz)}
\end{aligned}$$

The first cumulative factor is controlled by (13). For the second, t/b_t is nondecreasing for $t > 0$, so $t/b_t \le n/b_n$ for $t \le n$. Applying (14) controls the resulting total information term. Squaring the resulting inequality gives

$$\left(\sum_{t=2}^{n}\mathbb{E}[\Delta_t]\right)^2 \le \frac{e^2 d}{\log(3/2)}\log\left(1+\frac{n}{4d}\right)\left[4(n-1)+\frac{n}{\kappa_0 b_n}I((\lambda,\theta);\mathbf{H}_n)\right].$$

Then, Lemma 9 controls the total information term via an ϵ-net that balances the cardinality of a parameter cover against the KL error within a quantization cell. This leads to the second entry of Equation (15). Bounding every greedy gap by $e - e^{-1}$ gives the first entry. The supporting lemmas and the proof of Lemma 2 are presented in Appendices E.1 and E.2. □

7. Numerical Experiments

In this section we validate the theoretical results via simulations on synthetic data. The prior is independent, with $\lambda \sim \mathrm{Unif}[\underline{\lambda}, \bar{\lambda}]$ and $\theta \sim \mathcal{N}(0, I_d)$ conditioned on $\Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le 1\}$.

7.1. Practical Approximation of the Exact Posterior

The Thompson sampling policy in Algorithm 1 requires a draw from the exact joint posterior of (λ, θ), which is generally unavailable in closed form. For the numerical experiments, we approximate this posterior using a warm-started Gibbs sampler, referred to as Gibbs TS. Given the history $\mathcal{H}_{t-1}$, the posterior takes the form

$$G_t(\lambda, \theta) \propto \lambda^{M_{<t}} \mathcal{L}^{\theta}_{<t}(\theta; \lambda) \exp\left(-\frac{\|\theta\|_2^2}{2}\right) \mathbb{I}\{\underline{\lambda} \le \lambda \le \bar{\lambda}\} \mathbb{I}\{\theta \in \Theta\},$$

where $M_{<t}$ is the cumulative inside count and $\mathcal{L}^{\theta}_{<t}$ collects the likelihood terms involving θ.

The Gibbs updates exploit the structure of the Poisson–MNL likelihood. Conditional on θ, the posterior of λ is a truncated Gamma distribution and can be sampled directly. Conditional on λ, the posterior of θ is non-standard; we update it using elliptical slice sampling (Murray et al. 2010), which is well suited to the Gaussian prior and requires little tuning. At each decision period, we warm-start the chain from the previous posterior draw and perform K alternating Gibbs sweeps, each consisting of one λ-update and one elliptical-slice update for θ. The final draw is then used in Algorithm 1. We use $K = 64$ Gibbs sweeps per decision in all experiments. Detailed conditional distributions and sampling steps are provided in Appendix G.

7.2. Learning Dynamics under Poisson–MNL Feedback

We first illustrate how Thompson sampling learns the arrival rate and preference parameters from Poisson–MNL feedback. We consider a fixed design with $N = 1000$ alternatives, a horizon of $n = 50$ periods, and two-dimensional features ($d = 2$). The feature vectors are sampled once from a truncated standard Gaussian distribution and then held fixed throughout the experiment. One parameter pair $(\lambda^\star, \boldsymbol{\theta}^\star)$ is drawn from the prior and fixed across all replications. This design isolates

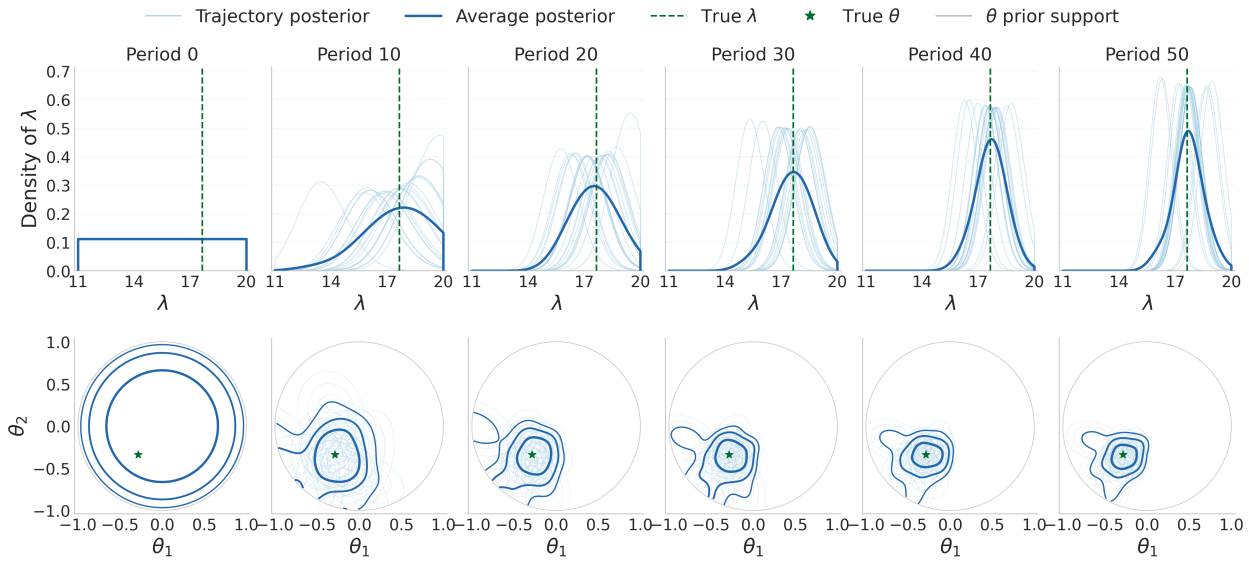


Figure 2 Posterior learning across repeated trajectories under Poisson–MNL feedback. Columns correspond to periods $0, 10, \ldots, 50$. The upper row displays the marginal posterior density of λ; the lower row displays the 50%, 80%, and 95% highest-density regions of $\boldsymbol{\theta}$. Light curves and contours represent the individual trajectories, while dark curves and contours are pointwise averages. The dashed vertical line and star mark $\lambda^\star$ and $\boldsymbol{\theta}^\star$, respectively.

learning uncertainty from instance heterogeneity: differences across trajectories arise only from Thompson-sampling randomization and stochastic feedback.

Figure 2 reports the posterior every 10 decision periods. At period 0, the displayed distributions coincide with the priors. As feedback accumulates, the posteriors can follow different paths, reflecting the interaction between adaptive decisions and realized demand. Nevertheless, their average becomes gradually concentrated around the fixed parameter pair, demonstrating that Poisson–MNL feedback permits the policy to learn both the market size and the preference coefficients within the decision horizon. Meanwhile, the posterior remains sufficiently diffuse to support continued exploration. Figure 3 further shows that the decisions made by Gibbs TS approach optimality as learning progresses.

7.3. Algorithm Comparison

In this section, we aim to validate the sublinear regret growth rate and compare different learning algorithms empirically. To produce statistically meaningful results, we draw a fresh candidate set of arm features for each outer replication. Conditional on this candidate set, we draw multiple pairs of parameters from the prior and average the resulting regret to estimate the Bayesian regret. Curves report means across outer replications, and shaded bands report one standard error.

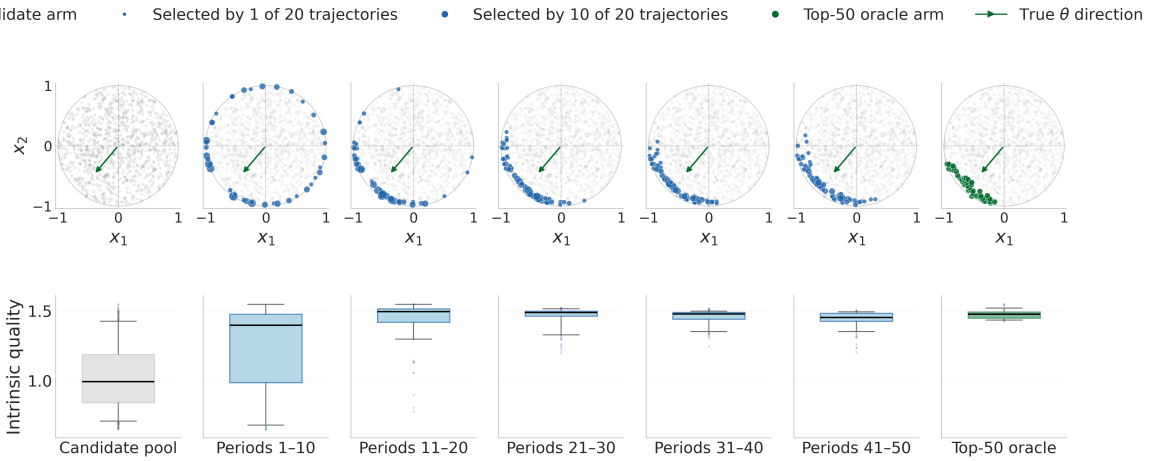


Figure 3 Decisions made by GIBBS TS under the Poisson–MNL feedback. Each column shows the arms selected during the 10 periods across the 20 trajectories as in Figure 2. Gray dots denote candidates, and blue markers denote selected arms. Marker size indicates the number of trajectories selecting an arm. Green circles identify the top intrinsic-quality arms, and arrows indicate $\theta^\star$. The lower panel shows boxplots of selected-arm quality.

We compare four policies. (i) ETE-FSE is implemented by Algorithm 2 in Appendix D.1, with the number of exploration periods set to $m = \min\left\{n, \max\left\{d+1, \lceil n^{2/3}\rceil\right\}\right\}$. We use no ℓ_2 regularization and constrain the estimate to $\|\theta\|_2 \le 1$. (ii) GIBBS TS is a computational Thompson-sampling policy described in Section 7.1. (iii) ETE is a standard explore-then-exploit benchmark: it first selects m arms uniformly at random and then greedily selects the remaining arms using a regularized MLE. We set $\lambda_{\text{reg}} = 0.1$ and constrain the estimate to $\|\theta\|_2 \le 1$. (iv) RANDOM selects arms uniformly from the remaining candidate set.

Across the results, two patterns are consistent. First, RANDOM accumulates regret much faster than the learning policies, confirming the value of using choice feedback to guide irreversible selections. Second, among the adaptive policies, GIBBS TS is generally the strongest performer, especially when feedback is sufficiently informative or the candidate pool is large, and the terminal subset size n is not too small relative to the feature dimension d.

7.3.1. Effect of subset size n. We investigate how the Bayesian regret scales with n to validate our main theoretical result in Theorem 4. In this simulation, we fix the prior $\lambda \sim \text{Uniform}(11, 20)$ and draw θ from a truncated standard Gaussian distribution with $d = 5$. To avoid the confounding effect of N, we scale N with n by a fixed factor. Figure 4 (a) shows how regret grows as n increases. Both ETE-FSE and TS exhibit sublinear regret growth, validating the theoretical results. It also shows that GIBBS TS achieves better performance compared with the ETE family as n grows.

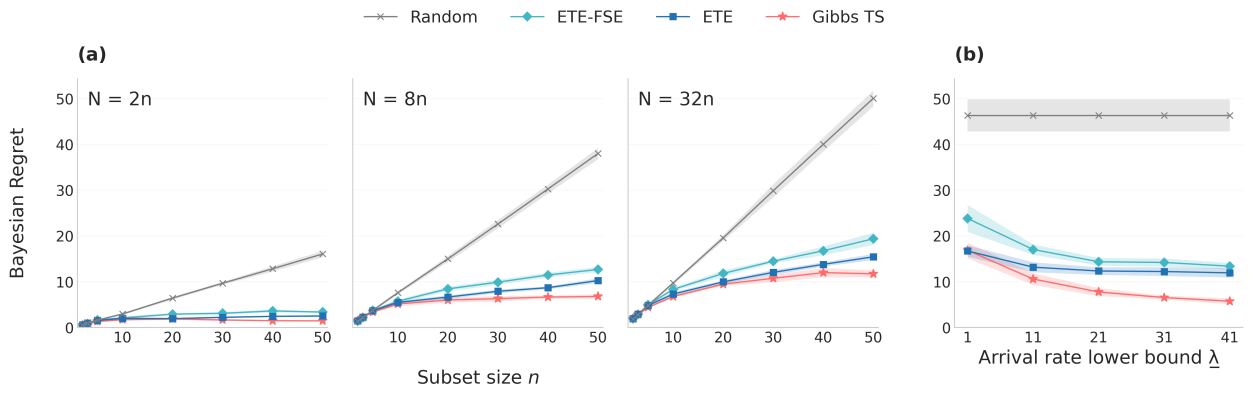


Figure 4 Sensitivity of terminal regret to (a) the subset size and (b) arrival rate.

7.3.2. Effect of arrival rate λ. We investigate the effect of the arrival rate by varying the prior of λ with other parameters $N = 1000$, $n = 50$, $d = 5$ held fixed. For different values of $\underline{\lambda}$, we keep $\overline{\lambda} - \underline{\lambda} = 9$. Figure 4 (b) shows that more feedback improves learning: as $\underline{\lambda}$ increases, the Bayesian regret of all learning algorithms decreases. Notably, Gibbs TS benefits substantially from the increased feedback rate and achieves considerably lower Bayesian regret than the ETE algorithms. This complements our theoretical guarantees by demonstrating the effectiveness of Thompson sampling in the high-$\underline{\lambda}$ regime, suggesting its potential for applications with high demand volumes.

7.3.3. Effect of candidate set size N. We investigate the effect of candidate set size by varying N with other parameters $n = 50$, $d = 5$ and the prior $\lambda \sim \text{Uniform}(11, 20)$ held fixed. Figure 5 (a) shows that a larger candidate pool makes policy guidance more important. When $N = 100$, the horizon is long relative to the candidate pool, and all adaptive policies achieve low Bayesian regret. In this small-pool regime, even random selection can include some high-quality arms. When N takes higher values, the search space is much larger and the performance gap between learning algorithms and the random baseline widens. Among all policies, Gibbs TS performs the best.

7.3.4. Effect of feature dimension d. We investigate how regret changes with the feature dimension d. We fix the prior $\lambda \sim \text{Uniform}(11, 20)$ and $n = 50$, and set $N = n \cdot 2^d$ to avoid vanishing reward variance. Figure 5 (b) illustrates the statistical cost of learning a higher-dimensional preference parameter. In the low- to moderate-dimensional regime, Gibbs TS achieves substantially lower Bayesian regret than the competing algorithms. As d increases, the regret of all three adaptive algorithms grows approximately linearly, consistent with the linear dependence on d in the lower bound of Theorem 3. For larger dimensions, however, this advantage diminishes: at $d = 10$ and 15, Gibbs TS is outperformed by ETE. One explanation is that Gibbs TS continues to explore uncertainty

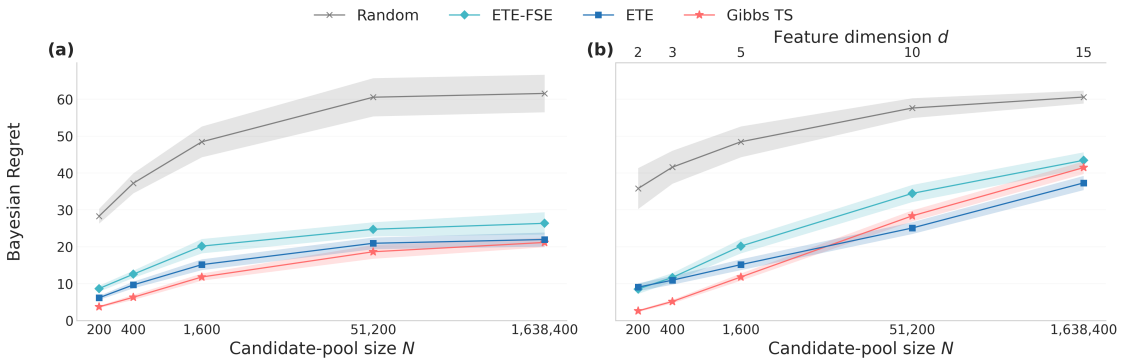


Figure 5 Sensitivity of terminal regret to (a) the candidate set size under $d = 5$ and (b) the feature dimension.

across multiple directions of the parameter space. When d is large relative to the selection budget n, such exploration becomes increasingly costly because exploratory selections are irreversible. In contrast, ETE commits to exploitation after its initial exploration phase, which can yield better empirical performance in this high-dimensional regime.

7.4. Objectives and Regret Criteria

We show that the choice of learning-performance criterion matters even when market share remains the underlying economic objective. We use the prior $\lambda \sim \text{Uniform}(11, 20)$ and draw θ from a truncated standard Gaussian distribution, with $d = 5$ and $N = 32n$. Because F^{MS} and F^{IQ} induce the same ranking of terminal sets and share the same full-information oracle, evaluating learning through F^{IQ}-regret does not change the construction target; it only measures the resulting loss on a different scale. Figure 6(a) shows that market-share regret approaches zero for all policies as n increases, so the saturating market-share scale increasingly obscures differences in construction quality. In contrast, Figure 6(b) shows that the intrinsic-quality criterion retains a linearly expanding scale, consistent with full resolution, and therefore continues to distinguish well-constructed from poorly constructed terminal sets. Thus, when market share is the economic objective, intrinsic-quality regret provides a more informative criterion for evaluating the learning performance of the construction policy.

8. Conclusion

We study online irreversible subset expansion under choice-based market feedback, where each selection affects both future information and terminal subset quality. Under the stated boundedness assumptions, market-share regret vanishes with subset size, while expected-revenue regret remains uniformly bounded, so small regret need not indicate successful learning. For market share, total

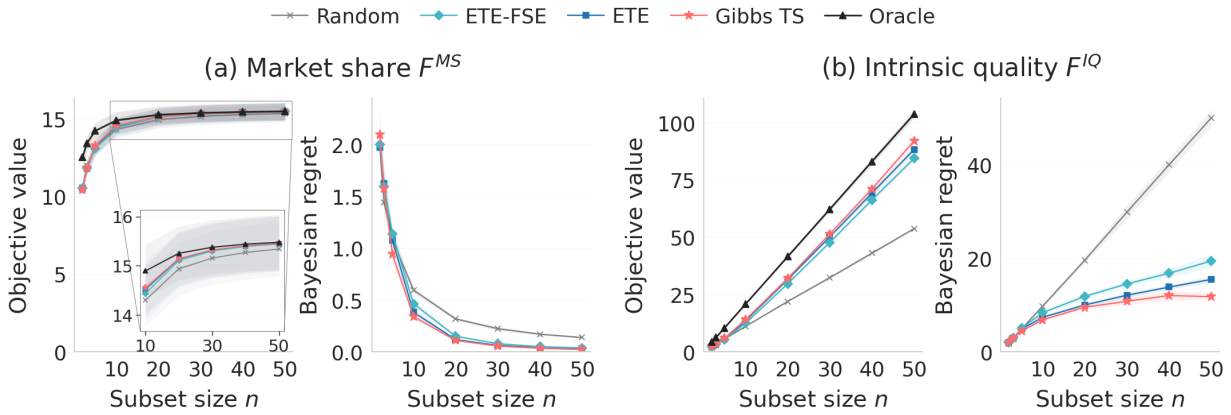


Figure 6 Terminal objective value and Bayesian regret under different objectives: (a) market share, and (b) intrinsic quality.

attraction mass preserves terminal-set rankings and the full-information oracle while providing full resolution under a natural separation condition. On this scale, we establish a minimax lower bound and analyze ETE-FSE as a frequentist benchmark based on fixed-reference-set estimation. Our main methodological contribution is an action–parameter uncertainty decomposition for Thompson sampling. Using the selected set's unnormalized centered design, we bound the squared greedy gap by the product of action- and parameter-uncertainty terms, controlled through an elliptical potential bound and a mutual information budget, respectively. This addresses information dilution without requiring a constant information-ratio bound uniform over time and candidate sets. Under the stated assumptions, the resulting Bayesian regret guarantee for exact Thompson sampling matches the lower bound in subset size and feature dimension, up to logarithmic factors.

Two directions merit further study. First, because expected revenue depends jointly on total attraction and revenue-weighted attraction, the simple re-cardinalization available for market share does not apply; constructing a full-resolution learning criterion that remains aligned with the expected-revenue objective is an important open question. Second, allowing the learner to delay an addition would make expansion timing endogenous, creating a joint learning, selection, and timing problem in which additional feedback must be balanced against the opportunity cost of waiting, modeled through a finite operating horizon, discounting, or an explicit delay cost.

References

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, volume 24, 2312–2320.

Agarwal A, Johnson N, Agarwal S (2020) Choice bandits. *Advances in Neural Information Processing Systems*, volume 33, 18399–18410.

Agrawal S, Avadhanula V, Goyal V, Zeevi A (2019) MNL-bandit: A dynamic learning approach to assortment selection. *Operations Research* 67(5):1453–1485.

Agrawal S, Avadhanula V, Goyal V, Zeevi A (2026) Thompson sampling for the multinomial logit bandit. *Mathematics of Operations Research* 51(1):568–590.

Almanza M, Chierichetti F, Lattanzi S, Panconesi A, Re G (2021) Online facility location with multiple advice. *Advances in Neural Information Processing Systems*, volume 34, 4661–4673.

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47(2–3):235–256, URL http://dx.doi.org/10.1023/A:1013689704352.

Ben-Porat O, Tennenholtz M (2019) Multiunit facility location games. *Mathematics of Operations Research* 44(3):865–889.

Bensoussan A, Chevalier-Roignant B (2019) Sequential capacity expansion options. *Operations research* 67(1):33–57.

Bravo F, Gandhi A, Hu J, Long EF (2026) Closer to home: A structural estimate-then-optimize approach to improve access to healthcare services. *Management Science* 72(3):2345–2363.

Bubeck S, Munos R, Stoltz G (2011) Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science* 412(19):1832–1852.

Cao J, Qi W, Zhang Y (2026) Online facility location: Running stores on wheels with spatial demand learning. *Manufacturing & Service Operations Management* 28(3):935–955.

Cao Y, Rusmevichientong P, Topaloglu H (2023) Revenue management under a mixture of independent demand and multinomial logit models. *Operations Research* 71(2):603–625.

Chen L, Krause A, Karbasi A (2017) Interactive submodular bandit. *Advances in Neural Information Processing Systems*, volume 30, 141–152.

Chen N (2021) Multi-armed bandit requiring monotone arm sequences. *Advances in Neural Information Processing Systems* 34:16093–16103.

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

Chen Y, Wang Y, Fang EX, Wang Z, Li R (2024) Nearly dimension-independent sparse linear bandit over small action spaces via best subset selection. *Journal of the American Statistical Association* 119(545):246–258.

Chu W, Li L, Reyzin L, Schapire R (2011) Contextual bandits with linear payoff functions. *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, volume 15 of *Proceedings of Machine Learning Research*, 208–214 (PMLR).

Dong S, Van Roy B (2018) An information-theoretic analysis for Thompson sampling with many actions. *Advances in Neural Information Processing Systems*, volume 31, 4157–4165 (Curran Associates, Inc.).

El Housni O, Ibn Brahim M, Segev D (2026) Maximum load assortment optimization: Approximation algorithms and adaptivity gaps. *Operations Research* 74(1):408–429.

Elberse A, Eliashberg J (2003) Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures. *Marketing science* 22(3):329–354.

Ferreira KJ, Simchi-Levi D, Wang H (2018) Online network revenue management using Thompson sampling. *Operations research* 66(6):1586–1602.

Gallego G, Ratliff R, Shebalov S (2015) A general attraction model and sales-based linear program for network revenue management under customer choice. *Operations Research* 63(1):212–232.

Guo H, Miao S, Qi W, Cui J, Dong Z (2025) Designing walkable neighborhoods: A data-driven optimization framework. *Available at SSRN 5356595* URL http://dx.doi.org/10.2139/ssrn.5356595, revised September 16, 2026.

Hamidi N, Bayati M (2023) Technical note—the elliptical potential lemma for general distributions with an application to linear Thompson sampling. *Operations Research* 71(4):1434–1439, URL http://dx.doi.org/10.1287/opre.2022.2274.

Han X, Wang L, Fan W (2023) Cost-effective social media influencer marketing. *INFORMS Journal on Computing* 35(1):138–157.

Han Y, Blanchet J, Zhou Z (2026) Improved confidence regions and optimal algorithms for online and offline linear mnl bandits. *Advances in Neural Information Processing Systems* 38:15208–15251.

Jacquillat A, Li ML (2026) Learning to cover: online learning and optimization with irreversible decisions. *Management Science* URL http://dx.doi.org/10.1287/mnsc.2025.01729, articles in Advance.

Jiang SHC, Liu E, Lyu Y, Tang ZG, Zhang Y (2022) Online facility location with predictions. *International Conference on Learning Representations*, URL https://openreview.net/forum?id=DSQHjibtgKR.

Korkut M, Li A (2021) Disposable linear bandits for online recommendations. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4172–4180.

Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6(1):4–22.

Legault R, Frejinger E (2025) A model-free approach for solving choice-based competitive facility location problems using simulation and submodularity. *INFORMS Journal on Computing* 37(3):603–622.

Li L, Chu W, Langford J, Schapire RE (2010) A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th international conference on World wide web*, 661–670.

Lin YH, Tian Q, Zhao Y (2022) Locating facilities under competition and market expansion: Formulation, optimization, and implications. *Production and Operations Management* 31(7):3021–3042.

Lin YH, Tian Q, Zhao Y (2024) Unified framework for choice-based facility location problem. *INFORMS Journal on Computing* 36(6):1436–1458.

Lin YH, Wang Y, He D, Lee LH (2020) Last-mile delivery: Optimal locker location under multinomial logit choice model. *Transportation research part E: Logistics and transportation review* 142:102059.

Lovász L (1983) Submodular functions and convexity. *Mathematical Programming The State of the Art: Bonn 1982*, 235–257 (Springer).

Luss H (1982) Operations research and capacity expansion problems: A survey. *Operations research* 30(5):907–947.

Mallipeddi RR, Kumar S, Sriskandarajah C, Zhu Y (2022) A framework for analyzing influencer marketing in social networks: selection and scheduling of influencers. *Management Science* 68(1):75–104.

Meyerson A (2001) Online facility location. *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, 426–431 (IEEE).

Murray I, Adams R, MacKay D (2010) Elliptical slice sampling. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, volume 9 of *Proceedings of Machine Learning Research*, 541–548 (PMLR).

Nemhauser GL, Wolsey LA, Fisher ML (1978) An analysis of approximations for maximizing submodular set functions—I. *Mathematical programming* 14(1):265–294.

Neu G, Olkhovskaia I, Papini M, Schwartz L (2022) Lifting the information ratio: An information-theoretic analysis of Thompson sampling for contextual bandits. *Advances in Neural Information Processing Systems*, volume 35, 9486–9498.

Oh Mh, Iyengar G (2019) Thompson sampling for multinomial logit contextual bandits. *Advances in Neural Information Processing Systems*, volume 32, 3151–3161.

Oh Mh, Iyengar G (2021) Multinomial logit contextual bandits: Provable optimality and practicality. *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 9205–9213.

Qi M, Jiang R, Shen S (2024) Sequential competitive facility location: Exact and approximate algorithms. *Operations Research* 72(1):300–316.

Rusmevichientong P, Shen ZJM, Shmoys DB (2010) Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research* 58(6):1666–1680.

Rusmevichientong P, Topaloglu H (2012) Robust assortment optimization in revenue management under the multinomial logit choice model. *Operations research* 60(4):865–882.

Russo D (2020) Simple Bayesian algorithms for best-arm identification. *Operations Research* 68(6):1625–1647, URL http://dx.doi.org/10.1287/opre.2019.1911.

Russo D, Van Roy B (2014) Learning to optimize via posterior sampling. *Mathematics of Operations Research* 39(4):1221–1243, URL http://dx.doi.org/10.1287/moor.2014.0650.

Russo D, Van Roy B (2016) An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research* 17(68):1–30.

Russo D, Van Roy B (2018) Learning to optimize via information-directed sampling. *Operations Research* 66(1):230–252, URL http://dx.doi.org/10.1287/opre.2017.1663.

Sui Y, Zoghi M, Hofmann K, Yue Y (2018) Advancements in dueling bandits. *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 5502–5510.

Talluri K, van Ryzin G (2004) Revenue management under a general discrete choice model of consumer behavior. *Management science* 50(1):15–33.

Wang Z, Deng S, Ye Y (2014) Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research* 62(2):318–331.

Wen Z, Kveton B, Valko M, Vaswani S (2017) Online influence maximization under independent cascade model with semi-bandit feedback. *Advances in Neural Information Processing Systems*, volume 30, 3022–3032.

Zhang Z, Chen W, Sun X, Zhang J (2022) Online influence maximization with node-level feedback using standard offline oracles. *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 9153–9161.

Zhu W, Rusmevichientong P, Topaloglu H (2026) A unified framework to impose market share constraints for selected product classes: Randomized and deterministic assortments under the multinomial logit model. *Manufacturing & Service Operations Management* 28(1):172–192.

Appendix A: Supporting Material for Section 3

A.1. Proof for Proposition 1

Proof. The map ϕ_λ is continuous and strictly increasing on $[0,\infty)$, since $\phi'_\lambda(y) = \lambda/(1+y)^2 > 0$. Write $W(S;\theta) := \sum_{i\in S} w_i(\theta)$ and $w_i(\theta) := \exp\big(V_i(\theta)\big)$. Under the Poisson–MNL model, $p_0(S;\theta) = 1/(1+W(S;\theta))$, and therefore

$$F^{MS}(S;\lambda,\theta) = \lambda\big(1-p_0(S;\theta)\big) = \lambda\Big(1-\tfrac{1}{1+W(S;\theta)}\Big) = \frac{\lambda W(S;\theta)}{1+W(S;\theta)} = \phi_\lambda\big(F^{IQ}(S;\theta)\big).$$

The stated equivalence follows from the strict monotonicity of ϕ_λ. □

A.2. Proof for Proposition 2

Proof. First, note the equivalence of these maximization problems

$$\max_{S:|S|=n} \lambda(1-p_0(S)) \;\Leftrightarrow\; \max_{S:|S|=n} 1 - \frac{1}{1+\sum_{i\in S}\exp(V_i)} \;\Leftrightarrow\; \max_{S:|S|=n} \sum_{i\in S}\exp(V_i)$$

Define weights $w_i := e^{V_i} > 0$. The problem maximizes a *modular* objective $\sum_{i\in S} w_i$ under a cardinality constraint. Let $S^\star$ be any maximizer. If there exist $i \in S^\star$ and $j \notin S^\star$ with $w_j > w_i$, swapping improves the value: $\sum_{\ell\in(S^\star\setminus\{i\})\cup\{j\}} w_\ell = \sum_{\ell\in S^\star} w_\ell - w_i + w_j > \sum_{\ell\in S^\star} w_\ell$, contradicting optimality of $S^\star$. Therefore every maximizer must contain the n largest values of V_i. By construction, the greedy rule selects in decreasing order of V_i, so $S^{\text{greedy},n}$ is optimal. □

A.3. Examples of F^{rev} Violating Monotonicity and Submodularity

We consider facility location as an application for these examples.

Monotonicity Adding a new facility may reduce total expected revenue if it attracts demand away from high-revenue incumbent facilities. This failure of monotonicity can occur even under our boundedness assumption $V_i(\theta) \in [-1,1]$, for which $w_i \in [e^{-1}, e]$. For example, let $(r_1, w_1) = (1, e)$, $(r_2, w_2) = (0, e)$. Then $F^{\text{rev}}(\{1\}) = \lambda\frac{e}{1+e}$, $F^{\text{rev}}(\{1,2\}) = \lambda\frac{e}{1+2e}$, and therefore $F^{\text{rev}}(\{1,2\}) < F^{\text{rev}}(\{1\})$, so opening facility 2 strictly decreases the objective.

Submodularity Moreover, the marginal value of a facility can increase after enlarging the current set, so the diminishing-returns property need not hold either. To see this, consider three facilities i, k, h with $(r_i, w_i) = (1,1)$, $(r_k, w_k) = \left(1, \frac{1}{2}\right)$, $(r_h, w_h) = (0, e^{-1})$. Let $S = \{i\}$ and $T = \{i, h\}$, so that $S \subset T$. Then the marginal contribution of facility k is

$$F^{\text{rev}}(S\cup\{k\}) - F^{\text{rev}}(S) = \lambda\left(\frac{1+\frac{1}{2}}{1+1+\frac{1}{2}} - \frac{1}{1+1}\right) = \frac{\lambda}{10},$$

$$F^{\mathrm{rev}}(T \cup \{k\}) - F^{\mathrm{rev}}(T) = \lambda \left(\frac{1 + \frac{1}{2}}{1 + 1 + e^{-1} + \frac{1}{2}} - \frac{1}{1 + e^{-1} + 1} \right) > \frac{\lambda}{10}.$$

Thus the marginal gain from adding facility k is *larger* for the superset T than for the subset S, which violates the diminishing-returns property. Hence F^{rev} is not submodular in general.

Appendix B: Supporting Material for Section 4

B.1. Proof of Proposition 5

Proof. By concavity, g lies below its tangents and has nonincreasing derivative. For any $x > y \geq 0$, $g'(x)(x-y) \leq g(x) - g(y) \leq g'(y)(x-y)$. Apply this with $x = W_n^+$ and $y = W_n^-$ to get

$$g'(W_n^+)\Delta_n \leq \bar{R}_n \leq g'(W_n^-)\Delta_n.$$

Since $W_n^+ \leq n w_{\max}$, $W_n^- \geq n w_{\min}$, and g' is nonincreasing, $g'(W_n^+) \geq g'(n w_{\max})$, $g'(W_n^-) \leq g'(n w_{\min})$, which yields (7). □

B.2. Corollaries of Theorem 2

We collect several direct implications of Theorem 2. These results characterize when objectives of the form $F_g(S) = g(\sum_{i \in S} w_i)$ are nondegenerate, strongly degenerate, or weakly degenerate.

B.2.1. Nondegeneracy

Corollary 2 (Characterization of Spread Nondegeneracy) *Under the assumptions of Theorem 2, there exists $c_0 > 0$ such that*

$$\liminf_{n \to \infty} \frac{\bar{R}_n}{n} \geq c_0$$

if and only if g is asymptotically linear:

$$g(y) = c_1 y + \phi(y), \qquad c_1 > 0, \quad \phi \text{ concave, and } \phi(y) = o(y).$$

Proof. We first show the "only if" part. By Theorem 2 and bounded weights, for all n, $\bar{R}_n \leq n(w_{\max} - w_{\min})\, g'(n w_{\min})$. If $\bar{R}_n \geq c_0 n$ for all large n, then for all large n, $g'(n w_{\min}) \geq \frac{c_0}{w_{\max} - w_{\min}} =: c_1 > 0$. Since g is concave, g' is nonincreasing on $(0, \infty)$. For any $y \in [n w_{\min}, (n+1) w_{\min})$, we have $g'(y) \geq g'((n+1) w_{\min})$. Therefore, for all sufficiently large y, $g'(y) \geq c_1$. Thus the limit $m := \lim_{y \to \infty} g'(y)$ exists and satisfies $m \geq c_1 > 0$.

Define $\phi(y) := g(y) - my$. Since my is linear and g is concave, ϕ is concave. Fix $\varepsilon > 0$. By definition of m, there exists Y such that for all $y \geq Y$, $m \leq g'(y) \leq m + \varepsilon$. Integrating over $[Y, y]$ gives $m(y - Y) \leq g(y) - g(Y) \leq (m + \varepsilon)(y - Y)$, hence

$$0 \leq \phi(y) - \phi(Y) = g(y) - my - (g(Y) - mY) \leq \varepsilon(y - Y).$$

Dividing by y and letting $y \to \infty$ yields $\limsup_{y\to\infty} \frac{\phi(y)}{y} \le \varepsilon$. Since ε is arbitrary and $\phi(y) \ge \phi(Y)$ for all $y \ge Y$, we conclude that $\phi(y) = o(y)$. Taking $c_1 := m$ completes the "only if" part.

For the "if" part, suppose $g(y) = c_1 y + \phi(y)$ with $c_1 > 0$ and $\phi(y) = o(y)$. Since g is concave and differentiable, this implies $g'(y) \to c_1$. Then Theorem 2 gives $\bar{R}_n \ge n\gamma g'(nw_{\max}) = \Theta(n)$, and hence $\liminf_{n\to\infty} \bar{R}_n/n > 0$. □

Corollary 3 (Necessary Condition for Nondegeneracy) *Within the class in Condition 2, nondegenerate terminal regret is possible only if g is asymptotically linear:*

$$g(y) = c_1 y + \phi(y), \qquad c_1 > 0, \quad \phi \text{ concave, and } \phi(y) = o(y).$$

Proof. This follows immediately from Corollary 2. □

B.2.2. Strong Degeneracy

Corollary 4 (Characterization of Strong Degeneracy) *Under the assumptions of Theorem 2, the spread is strongly degenerate, i.e., $\bar{R}_n = O(1)$, if and only if*

$$g'(y) = O(1/y).$$

In this case, $g(y) = O(\log y)$ up to an additive constant.

Proof. **"Only if."** By Theorem 2, for every $n \in \mathbb{N}^+$, $\gamma n\, g'(nw_{\max}) \le \bar{R}_n$. If $\bar{R}_n = O(1)$, then there exists $C < \infty$ such that $\limsup_{n\to\infty} \bar{R}_n \le C$. Therefore, $\limsup_{n\to\infty} n\, g'(nw_{\max}) \le \frac{C}{\gamma}$. Equivalently, along the diverging sequence $y_n := nw_{\max}$, $\limsup_{n\to\infty} y_n\, g'(y_n) \le \frac{Cw_{\max}}{\gamma}$. Since g is concave, g' is nonincreasing. For any $y \ge w_{\max}$, let $n := \lfloor y/w_{\max} \rfloor$, so that $nw_{\max} \le y < (n+1)w_{\max}$. Then $g'(y) \le g'(nw_{\max})$, and hence

$$y\, g'(y) \le y\, g'(nw_{\max}) \le (n+1)w_{\max}\, g'(nw_{\max}) = w_{\max}\left(1 + \frac{1}{n}\right)\big(n\, g'(nw_{\max})\big).$$

Taking $\limsup_{y\to\infty}$ gives $\limsup_{y\to\infty} y\, g'(y) \le \frac{Cw_{\max}}{\gamma}$. Thus $g'(y) = O(1/y)$.

"If." Suppose $g'(y) = O(1/y)$. Then there exist constants $C < \infty$ and $y_0 > 0$ such that $g'(y) \le C/y$ for all $y \ge y_0$. Choose n_0 such that $n_0 w_{\min} \ge y_0$. For all $n \ge n_0$, the upper bound in Theorem 2 gives

$$\bar{R}_n \le n(w_{\max} - w_{\min})\, g'(nw_{\min}) \le n(w_{\max} - w_{\min})\frac{C}{nw_{\min}} = \frac{C(w_{\max} - w_{\min})}{w_{\min}}.$$

Thus $\bar{R}_n = O(1)$, proving strong degeneracy.

Finally, if $g'(y) = O(1/y)$, then integrating gives $g(y) = O(\log y)$ up to an additive constant. □

B.2.3. Weak Degeneracy

Corollary 5 (Characterization of Weak Degeneracy) *Under the assumptions of Theorem 2, the spread is weakly degenerate if and only if*

$$g'(y) \to 0 \qquad \text{and} \qquad g'(y) \neq O(1/y).$$

Equivalently, $g'(y) \to 0$ and $yg'(y)$ is unbounded.

Proof. We first show the "only if" part. If the spread is weakly degenerate, then it is degenerate, so $\bar{R}_n/n \to 0$. By the lower bound in Theorem 2, $\gamma g'(nw_{\max}) \leq \frac{\bar{R}_n}{n}$. Therefore $g'(nw_{\max}) \to 0$. Since g' is nonincreasing, this implies $g'(y) \to 0$. Moreover, weak degeneracy means that the spread is not strongly degenerate. By Corollary 4, this implies $g'(y) \neq O(1/y)$.

Conversely, suppose $g'(y) \to 0$ and $g'(y) \neq O(1/y)$. The upper bound in Theorem 2 gives $\frac{\bar{R}_n}{n} \leq (w_{\max} - w_{\min})\, g'(nw_{\min}) \to 0$, so the spread is degenerate. Since $g'(y) \neq O(1/y)$, Corollary 4 implies that the spread is not strongly degenerate. Therefore the spread is weakly degenerate. □

B.2.4. Polynomial Strong Degeneracy

Corollary 6 (Polynomial Strong Degeneracy) *If $g'(y) = \Theta(y^{-(1+\varepsilon)})$ for some $\varepsilon > 0$, then g has a finite limit g_∞ and*

$$g_\infty - g(y) = \Theta(y^{-\varepsilon}).$$

Moreover,

$$\bar{R}_n = \Theta(n^{-\varepsilon}).$$

Proof. If $g'(y) = \Theta(y^{-(1+\varepsilon)})$ for some $\varepsilon > 0$, then there exist constants $0 < c_1 < c_2 < \infty$ and $y_1 > 0$ such that for all $y \geq y_1$, $c_1 y^{-(1+\varepsilon)} \leq g'(y) \leq c_2 y^{-(1+\varepsilon)}$. Since g' is integrable at infinity, g has a finite limit g_∞, and

$$g_\infty - g(y) = \int_y^\infty g'(u)\, du = \Theta(y^{-\varepsilon}).$$

Let n be large enough so that $nw_{\max} \geq nw_{\min} \geq y_1$. Applying Theorem 2 yields

$$\gamma n\, c_1 (nw_{\max})^{-(1+\varepsilon)} \leq \bar{R}_n \leq n(w_{\max} - w_{\min})\, c_2 (nw_{\min})^{-(1+\varepsilon)}.$$

Simplifying gives $\bar{R}_n = \Theta(n^{-\varepsilon})$. □

Appendix C: Supporting Material for Section 5

Lemma 3 (Bretagnolle–Huber event bound) *Let $\theta, \theta' \in \Theta$ satisfy $\theta_j = \theta'_j$ for all $j \neq i$ and $\theta_i = -\theta'_i$. Assume the constructed action set satisfies $x_{a_t,i} \in \{\pm 1/\sqrt{d}\}$ for all selected arms. Without loss of generality, take $\theta_i = +\Delta/\sqrt{d}$ and $\theta'_i = -\Delta/\sqrt{d}$, and define the event*

$$E_i \triangleq \Big\{ \sum_{t=1}^{n} \mathbb{I}[x_{a_t,i} = -1/\sqrt{d}] > n/2 \Big\}.$$

Let $p_{\theta i} \triangleq \mathbb{P}^{\pi}_{\theta}(E_i)$ and $p_{\theta' i} \triangleq \mathbb{P}^{\pi}_{\theta'}(E_i^c)$ be the probabilities induced by π. Then

$$p_{\theta i} + p_{\theta' i} \geq \frac{1}{2} \exp\Big(- D_{\mathrm{KL}}(\mathbb{P}^{\pi}_{\theta} \| \mathbb{P}^{\pi}_{\theta'}) \Big). \tag{17}$$

Moreover, the KL divergence admits the bound

$$D_{\mathrm{KL}}(\mathbb{P}^{\pi}_{\theta} \| \mathbb{P}^{\pi}_{\theta'}) \leq \frac{\lambda n}{2} e^{4\Delta/d} \big(e^{4\Delta/d} - 1\big)^2 \leq 8 \lambda n \frac{\Delta^2}{d^2} e^{12\Delta/d}. \tag{18}$$

Consequently,

$$p_{\theta i} + p_{\theta' i} \geq \frac{1}{2} \exp\Big(- 8 \lambda n \frac{\Delta^2}{d^2} e^{12\Delta/d} \Big).$$

Proof. For any two measures P, Q on the same space and any event E, the Bretagnolle–Huber inequality gives

$$P(E) + Q(E^c) \geq \frac{1}{2} \exp\big(- D_{\mathrm{KL}}(P \| Q) \big).$$

Applying this with $P = \mathbb{P}^{\pi}_{\theta}$, $Q = \mathbb{P}^{\pi}_{\theta'}$, and $E = E_i$ yields (17).

Because π_t is $\mathcal{H}_{t-1}$-measurable, and the conditional probability is identical under both θ and θ' given the history, the KL divergence decomposes as

$$D_{\mathrm{KL}}(\mathbb{P}^{\pi}_{\theta} \| \mathbb{P}^{\pi}_{\theta'}) = \sum_{t=1}^{n} \mathbb{E}^{\pi}_{\theta} \Big[D_{\mathrm{KL}}\Big(\mathbb{P}_{\theta}(\boldsymbol{y}_t \mid \mathcal{H}_{t-1}) \,\Big\|\, \mathbb{P}_{\theta'}(\boldsymbol{y}_t \mid \mathcal{H}_{t-1}) \Big) \Big].$$

Conditioned on $\mathcal{H}_{t-1}$ (hence on S_t), the observations are a product of Poissons, so

$$D_{\mathrm{KL}}\Big(\mathbb{P}_{\theta}(\boldsymbol{y}_t \mid \mathcal{H}_{t-1}) \,\Big\|\, \mathbb{P}_{\theta'}(\boldsymbol{y}_t \mid \mathcal{H}_{t-1}) \Big) = \sum_{j \in S_t} d_{\mathrm{Pois}}(\lambda p_j, \lambda p'_j),$$

where $p_j = p_j(S_t; \theta)$, $p'_j = p_j(S_t; \theta')$, and $d_{\mathrm{Pois}}(\mu, \nu) = \mu \log(\mu/\nu) + \nu - \mu$. Thus

$$D_{\mathrm{KL}}\Big(\mathbb{P}_{\theta}(\boldsymbol{y}_t \mid \mathcal{H}_{t-1}) \,\Big\|\, \mathbb{P}_{\theta'}(\boldsymbol{y}_t \mid \mathcal{H}_{t-1}) \Big) = \lambda \sum_{j \in S_t} \left[p_j \log \frac{p_j}{p'_j} + p'_j - p_j \right].$$

$x_{j,i} \in [-1/\sqrt{d}, 1/\sqrt{d}]$ and $\theta_i - \theta'_i = 2\Delta/\sqrt{d}$, so $|(\theta-\theta')^\top x_j| \le 2\Delta/d$ for all arms j. Writing

$$\frac{p_j(S;\theta)}{p_j(S;\theta')} = \exp\big((\theta-\theta')^\top x_j\big) \cdot \frac{1+\sum_{\ell\in S}\exp(\theta'^\top x_\ell)}{1+\sum_{\ell\in S}\exp(\theta^\top x_\ell)},$$

and since each term in the denominator changes by the same factor at most, one obtains the ratio bound $e^{-4\Delta/d} \le \frac{p_j(S;\theta)}{p_j(S;\theta')} \le e^{4\Delta/d}$, $\forall j \in S$. Let $x = p_j/p'_j$. Then $p_j \log\frac{p_j}{p'_j} + p'_j - p_j = p'_j\big(x\log x - x + 1\big)$. By Taylor's theorem, for any $x > 0$, $x\log x - x + 1 \le \frac{(x-1)^2}{2\min\{x,1\}}$. The ratio bound gives $\min\{x,1\} \ge e^{-4\Delta/d}$ and hence $x\log x - x + 1 \le \frac{e^{4\Delta/d}}{2}(x-1)^2 \le \frac{e^{4\Delta/d}}{2}\big(e^{4\Delta/d}-1\big)^2$. Therefore,

$$D_{\mathrm{KL}}\Big(\mathbb{P}_\theta(\boldsymbol{y}_t \mid \mathcal{H}_{t-1}) \,\Big\|\, \mathbb{P}_{\theta'}(\boldsymbol{y}_t \mid \mathcal{H}_{t-1})\Big) \le \lambda \cdot \frac{e^{4\Delta/d}}{2}\big(e^{4\Delta/d}-1\big)^2 \sum_{j\in S_t} p'_j.$$

Finally, $\sum_{j\in S_t} p'_j \le 1$, so $D_{\mathrm{KL}}\Big(\mathbb{P}_\theta(\boldsymbol{y}_t \mid \mathcal{H}_{t-1}) \,\Big\|\, \mathbb{P}_{\theta'}(\boldsymbol{y}_t \mid \mathcal{H}_{t-1})\Big) \le \frac{\lambda}{2}e^{4\Delta/d}\big(e^{4\Delta/d}-1\big)^2$. Summing over $t = 1,\dots,n$ gives the first inequality in (18). For the second inequality, use $e^{4\Delta/d} - 1 \le 4\frac{\Delta}{d}e^{4\Delta/d}$, so $\frac{1}{2}e^{4\Delta/d}(e^{4\Delta/d}-1)^2 \le 8\frac{\Delta^2}{d^2}e^{12\Delta/d}$. Substituting (18) into (17) yields the result. □

C.1. Proof for Theorem 3

For the lower bound, we construct the hard instance as follows. Fix $n, d \ge 1$ and define

$$\mathcal{N}_{n,d} := \{(u,l) : u \in \{-1,+1\}^d,\ l \in [n]\}, \qquad x_{(u,l)} := \frac{u}{\sqrt{d}}.$$

The labels l make the replicas distinct arms. The candidate pool contains exactly n replicas of each sign type and has cardinality $n\,2^d$. Take $\Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le 1\}$.

Theorem 5 (Formal version of minimax lower bound in Theorem 3) *For the replicated-hypercube instance above, fix any arrival rate $\lambda > 0$. Under the Poisson–MNL feedback model and the intrinsic-quality objective in (4), fix $\Delta \in (0,1]$ and put $\kappa_\Delta := 8\lambda n\frac{\Delta^2}{d^2}\exp\Big(\frac{12\Delta}{d}\Big)$. Every policy $\pi \in \Pi_n$ satisfies*

$$\sup_{\theta\in\Theta} \mathrm{Regret}_n^{F^{IQ}}(\pi;\lambda,\theta) \ge \frac{n\Delta}{4e}e^{-\kappa_\Delta}.$$

In particular, the admissible choice $\Delta = \frac{1}{8}\min\{1, d/\sqrt{\lambda n}\}$ gives

$$\sup_{\theta\in\Theta} \mathrm{Regret}_n^{F^{IQ}}(\pi;\lambda,\theta) \ \ge\ \frac{1}{32e^2}\min\left\{n,\ d\sqrt{\frac{n}{\lambda}}\right\}.$$

Proof. For any $\theta \in \Theta$, choose $u^\star \in \arg\max_{u\in\{-1,+1\}^d} u^\top\theta$. The n copies of type $u^\star$ have the largest attraction weight in the pool. For the general bound, fix $\Delta \in (0,1]$ and set

$$\theta^v := \frac{\Delta}{\sqrt{d}}v, \quad v \in \{-1,+1\}^d.$$

If $v^{(i)}$ denotes v with coordinate i flipped, Lemma 3 gives the uniform neighboring-history bound

$$D_{\mathrm{KL}}(\mathbb{P}^{\pi}_{\theta^v} \| \mathbb{P}^{\pi}_{\theta^{v^{(i)}}}) \le \kappa_\Delta.$$

Write the arm selected at time t as $a_t = (u_t, l_t)$, and define the coordinate mismatch count $M_i(v) := \sum_{t=1}^{n} \mathbb{I}\{u_{t,i} \neq v_i\}$. Fix a coordinate i and a vertex v with $v_i = +1$. For the event $E_i(v) := \left\{\sum_{t=1}^{n} \mathbb{I}\{u_{t,i} = -1\} > \frac{n}{2}\right\}$, we have $M_i(v) > n/2$ on $E_i(v)$, while $M_i(v^{(i)}) \ge n/2$ on $E_i(v)^c$. Lemma 3 therefore implies

$$\mathbb{E}^{\pi}_{\theta^v} M_i(v) + \mathbb{E}^{\pi}_{\theta^{v^{(i)}}} M_i(v^{(i)}) \ge \frac{n}{2}\left[\mathbb{P}^{\pi}_{\theta^v}(E_i(v)) + \mathbb{P}^{\pi}_{\theta^{v^{(i)}}}(E_i(v)^c)\right] \ge \frac{n}{4} e^{-\kappa_\Delta}. \tag{19}$$

Partitioning the 2^d vertices into 2^{d-1} pairs along coordinate i, summing over all coordinates, and dividing by 2^d gives $\frac{1}{2^d}\sum_{v\in\{-1,+1\}^d}\sum_{i=1}^{d} \mathbb{E}^{\pi}_{\theta^v} M_i(v) \ge \frac{nd}{8} e^{-\kappa_\Delta}$.

Next, we translate mismatches into regret. Let $H(u,v) := \sum_{i=1}^{d} \mathbb{I}\{u_i \neq v_i\}$. Under θ^v, an arm of type u has utility $x_{(u,k)}^{\top}\theta^v = \Delta\,(1 - 2H(u,v)/d)$. Under Assumption 1, the mean-value theorem yields $e^{\Delta} - \exp(x_{(u,k)}^{\top}\theta^v) \ge \frac{2\Delta}{ed} H(u,v)$, and therefore $\mathrm{Regret}_n^{FIQ}(\pi;\lambda,\theta^v) \ge \frac{2\Delta}{ed}\sum_{i=1}^{d} \mathbb{E}^{\pi}_{\theta^v} M_i(v)$. Averaging over v, applying Equation (19), and using that a supremum is at least an average, we have

$$\sup_{\theta\in\Theta} \mathrm{Regret}_n^{FIQ}(\pi;\lambda,\theta) \ge \frac{n\Delta}{4e} e^{-\kappa_\Delta}.$$

Finally, choose $\Delta = \frac{1}{8}\min\{1, d/\sqrt{\lambda n}\}$. Then $\lambda n\Delta^2/d^2 \le 1/64$ and $12\Delta/d \le 3/2$, so $\kappa_\Delta \le e^{3/2}/8 < 1$. Hence $\sup_{\theta\in\Theta} \mathrm{Regret}_n^{FIQ}(\pi;\lambda,\theta) \ge \frac{n\Delta}{4e^2} = \frac{1}{32e^2}\min\left\{n,\ d\sqrt{\frac{n}{\lambda}}\right\}$. □

Appendix D: Supporting Material for ETE-FSE

D.1. Fixed-set estimator and policy

The growing active set creates a statistical complication: as S_t expands, the dimension of $\boldsymbol{y}_t$ changes, and observations for earlier selections are generated under progressively larger choice sets. A maximum-likelihood estimator that pools all observations must therefore account for a sequence of changing choice environments.

Definition 3 (Fixed-Set Estimator, FSE) *Fix $m \le n$ and let S_m be the active set after the m-th expansion. For any $t \ge m$, define the fixed-set estimator $\widehat{\theta}_t^m$ to be the conditional maximum-likelihood estimator built only from observations on S_m:*

$$\widehat{\theta}_t^m \in \arg\min_{\vartheta\in\Theta} \mathcal{L}_t^m(\vartheta), \qquad \mathcal{L}_t^m(\vartheta) := \sum_{\tau=m}^{t}\left[\left(\sum_{i\in S_m} y_{i,\tau}\right)\log\left(\sum_{j\in S_m} e^{x_j^{\top}\vartheta}\right) - \sum_{i\in S_m} y_{i,\tau} x_i^{\top}\vartheta\right]. \tag{20}$$

Algorithm 2 Explore-then-Exploit with Fixed-Set Estimator (ETE-FSE)

Input: exploration length m.
Initialize: $S_0 = \varnothing$.
for $t = 1, 2, \ldots, n$ **do**
  if $t \le m$ **then** ▷ Exploration
    Sample an arm uniformly: $a_t \sim \mathrm{Unif}\big(\mathcal{N} \setminus S_{t-1}\big)$, $S_t \leftarrow S_{t-1} \cup \{a_t\}$.
  else ▷ Exploitation
    Select an arm greedily: $a_t \in \arg\max_{i \in \mathcal{N} \setminus S_{t-1}} x_i^\top \widehat{\theta}^m_{t-1}$, $S_t \leftarrow S_{t-1} \cup \{a_t\}$.
  end if
  Observe counts $y_{i,t}$ for all $i \in S_t$.
  if $t \ge m$ **then**
    Update the fixed-set estimator: $\widehat{\theta}^m_t \leftarrow \arg\min_{\vartheta \in \Theta} \mathcal{L}^m_t(\vartheta)$. ▷ FSE, Definition 3
  end if
end for
Output: S_n.

D.2. Random-exploration design

Assumption 6 (κ_1-Isotropic Candidate Features) *The features $\{x_i\}_{i \in \mathcal{N}}$ are i.i.d. draws from a distribution with $\mathbb{E}[x] = 0$, and $\Sigma := \mathbb{E}[xx^\top]$ satisfies $\lambda_{\min}(\Sigma) \ge (\kappa_1 d)^{-1}$ for some $\kappa_1 \ge 1$.*

Remark 7 *Under Assumption 1, we have $\mathrm{tr}(\Sigma) = \mathbb{E}\|x\|_2^2 \le 1$, while $\mathrm{tr}(\Sigma) = \sum_{j=1}^d \lambda_j(\Sigma) \ge d\,\lambda_{\min}(\Sigma)$. Hence $\lambda_{\min}(\Sigma) \le 1/d$ and Assumption 6 is tight up to a constant κ_1.*

Proposition 7 (Random exploration growth rate) *Under Assumption 6 and Assumption 1, fix $m \ge 2$ and let $x_1, \ldots, x_m$ be i.i.d. samples. Let $D_m = \sum_{i=1}^m x_i x_i^\top$, $B_m = \frac{1}{m}\sum_{i=1}^m (x_i - \bar{x}_m)(x_i - \bar{x}_m)^\top$, $\widetilde{B}_m = m B_m$. For any $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\lambda_{\min}(D_m) \ge (1 - \varepsilon)\frac{m}{\kappa_1 d}, \qquad \varepsilon := \sqrt{\frac{2\kappa_1 d}{m} \log \frac{2d}{\delta}}. \tag{21}$$

On the same event,

$$\lambda_{\min}(\widetilde{B}_m) \ge (1 - \varepsilon)\frac{m}{\kappa_1 d} - 8\Big(d \log 5 + \log \frac{2}{\delta}\Big). \tag{22}$$

In particular, if $m \geq 8\kappa_1 d \cdot \max\left\{\log\frac{2d}{\delta},\ 2\left(d\log 5 + \log\frac{2}{\delta}\right)\right\}$, *there exists a constant* $C_2 > 0$ *such that, with probability at least* $1-\delta$,

$$\lambda_{\min}(\widetilde{B}_m) \geq C_2\,\frac{m}{\kappa_1 d}.$$

Proof. *Step 1 (matrix Chernoff for* D_m*).* $\mathbb{E}[D_m] = m\Sigma$. Moreover, $0 \preceq x_i x_i^\top \preceq I_d$ because $\|x_i\| \leq 1$. The matrix Chernoff bound implies that for any $\varepsilon \in (0,1)$,

$$\mathbb{P}(\lambda_{\min}(D_m) \leq (1-\varepsilon)\lambda_{\min}(\mathbb{E}[D_m])) \leq d\exp\left(-\frac{\varepsilon^2\,\lambda_{\min}(\mathbb{E}[D_m])}{2}\right).$$

Since $\lambda_{\min}(\mathbb{E}[D_m]) = m\lambda_{\min}(\Sigma) \geq m/(\kappa_1 d)$, we get $\mathbb{P}\left(\lambda_{\min}(D_m) \leq (1-\varepsilon)\frac{m}{\kappa_1 d}\right) \leq d\exp\left(-\frac{\varepsilon^2}{2}\cdot\frac{m}{\kappa_1 d}\right)$. Setting the right-hand side to $\delta/2$ yields $\varepsilon = \sqrt{\frac{2\kappa_1 d}{m}\log\frac{2d}{\delta}}$ and proves (21) with probability $1-\delta/2$.

Step 2 (centering penalty). Using $\widetilde{B}_m = D_m - m\bar{x}_m\bar{x}_m^\top$ and Weyl's inequality,

$$\lambda_{\min}(\widetilde{B}_m) \ \geq\ \lambda_{\min}(D_m) - \lambda_{\max}(m\bar{x}_m\bar{x}_m^\top) = \lambda_{\min}(D_m) - m\|\bar{x}_m\|^2.$$

Step 3 (bounding $\|\bar{x}_m\|$*).* Let $S_m = \sum_{i=1}^m x_i$. Fix $u \in \mathbb{S}^{d-1}$ and define $Z_i = u^\top x_i$. Then $\mathbb{E}[Z_i] = 0$ and $|Z_i| \leq \|x_i\| \leq 1$ a.s., so Hoeffding gives $\mathbb{P}(u^\top S_m \geq \xi) \leq \exp(-\xi^2/(2m))$. Let $\mathcal{N}$ be a $1/2$-net of $\mathbb{S}^{d-1}$ with $|\mathcal{N}| \leq 5^d$, and use $\|S_m\| = \sup_{\|u\|=1} u^\top S_m \leq 2\max_{u\in\mathcal{N}} u^\top S_m$. A union bound yields

$$\mathbb{P}\big(\|S_m\| \geq 2\xi\big) \ \leq\ 5^d\exp\left(-\frac{\xi^2}{2m}\right).$$

Choosing $\xi = \sqrt{2m\big(d\log 5 + \log(2/\delta)\big)}$ gives, with probability at least $1-\delta/2$,

$$\|S_m\| \ \leq\ 2\sqrt{2m\big(d\log 5 + \log(2/\delta)\big)} \quad\Rightarrow\quad m\|\bar{x}_m\|^2 = \frac{\|S_m\|^2}{m} \ \leq\ 8\left(d\log 5 + \log\frac{2}{\delta}\right).$$

Step 4. Combining Steps 1 and 3 and applying a union bound gives (22).

Remark 8 *The upper bound above gives* $\lambda_{\min}(\widetilde{B}_m) \leq 4m/d$ *for all bounded features. Proposition 7 shows that, under* κ_1*-isotropy and i.i.d. exploration,* $\lambda_{\min}(\widetilde{B}_m) \gtrsim m/(\kappa_1 d)$ *once* m *is at least of order* $\kappa_1 d\big(d + \log(d/\delta)\big)$. *Thus the* m/d *scaling is tight up to constants.*

D.3. Fixed-set estimation error

Proposition 8 (Estimation Error) *Let* $\mathcal{D}_m := \sigma\big((x_i)_{i\in\mathcal{N}}, a_1, \ldots, a_m\big)$ *record the candidate features and the uniform exploration indices of ETE-FSE. Under Assumption 1, there exists a constant* $C_1 > 0$ *such that, for every* $t \geq m$, *on* $\{\lambda_{\min}(\widetilde{B}_m) > 0\}$,

$$\mathbb{E}\left[\|\widehat{\theta}_t^m - \theta\|_2^2 \,\middle|\, \mathcal{D}_m\right] \leq C_1\,\frac{d}{\lambda}\,\frac{1}{\lambda_{\min}(\widetilde{B}_m)}\,\frac{t+1}{t-m+1}. \tag{23}$$

Proof. Fix (λ, θ) and condition throughout on $\mathcal{D}_m$, with $b := \lambda_{\min}(\widetilde{B}_m) > 0$. *Step 1 (Conditional multinomial).* Fix $\tau \geq m$ and condition on $\mathcal{F}_{\tau-} := \sigma(\mathcal{D}_m, \mathcal{H}_{\tau-1}, a_\tau)$. By Poisson thinning, $\{y_{i,\tau}\}_{i \in S_m}$ are independent Poisson random variables with means $\lambda p_i(S_\tau; \theta)$. Let $M_\tau := \sum_{i \in S_m} y_{i,\tau}$,

$$(y_{i,\tau})_{i \in S_m} \,\Big|\, (M_\tau, \mathcal{F}_{\tau-}) \sim \text{Multinomial}\Big(M_\tau,\ \big(q_i(S_m; \theta)\big)_{i \in S_m}\Big).$$

Thus (20) sums the periodwise conditional softmax negative log-likelihoods on S_m, with total sample size $M := \sum_{\tau=m}^{t} M_\tau = \sum_{\tau=m}^{t} \sum_{i \in S_m} y_{i,\tau}$. For one categorical draw, the per-sample Hessian is $H(\theta) := X_m^\top \big(\text{diag}(q(\theta)) - q(\theta) q(\theta)^\top\big) X_m$, where X_m is the $m \times d$ matrix with rows $\{x_i^\top\}_{i \in S_m}$, and $q(\vartheta) := q(S_m; \vartheta)$.

Step 2 (Hessian sandwich). Define $q_{\min} := \inf_{\|\vartheta\|_2 \leq 1} \min_{i \in S_m} q_i(\vartheta)$, $q^\star_{\max} := \max_{i \in S_m} q_i(\theta)$. Analyzing the variance form implies that for all $\|\vartheta\|_2 \leq 1$, $H(\vartheta) \succeq q_{\min} \widetilde{B}_m$, $H(\theta) \preceq q^\star_{\max} \widetilde{B}_m$.

Step 3 (basic inequality and score second moment). Write $u := \widehat{\theta}_t^m - \theta$ and $G_t := \nabla \mathcal{L}_t^m(\theta)$. Since $\mathcal{L}_t^m(\widehat{\theta}_t^m) \leq \mathcal{L}_t^m(\theta)$ and the segment joining the two parameters lies in the unit ball, Step 2 gives

$$\frac{M q_{\min}}{2} \|u\|_{\widetilde{B}_m}^2 \leq -G_t^\top u \leq \|G_t\|_{\widetilde{B}_m^{-1}} \|u\|_{\widetilde{B}_m}.$$

For $M > 0$, it follows that $\|u\|_2^2 \leq 4 \|G_t\|_{\widetilde{B}_m^{-1}}^2 / (M^2 q_{\min}^2 b)$. Let $\rho_\tau := \lambda \sum_{i \in S_m} p_i(S_\tau; \theta)$. Given $\mathcal{F}_{\tau-}$, the period-τ score increment has mean zero and covariance $\rho_\tau H(\theta)$. Orthogonality of these martingale differences therefore yields

$$\mathbb{E}[\|G_t\|_{\widetilde{B}_m^{-1}}^2 \mid \mathcal{D}_m] = \text{tr}(\widetilde{B}_m^{-1} H(\theta))\, \mathbb{E}[M \mid \mathcal{D}_m] \leq q^\star_{\max} d\, \mathbb{E}[M \mid \mathcal{D}_m].$$

Step 4 (controlling the adaptive sample size). Since $|S_\tau| = \tau$ and $|x_i^\top \theta| \leq 1$,

$$\frac{e^{-2} \lambda m}{\tau + 1} \leq \rho_\tau \leq \frac{e^2 \lambda m}{\tau + 1}. \qquad \nu := e^{-2} \lambda m \sum_{\tau=m}^{t} \frac{1}{\tau + 1}.$$

Hence $\mathbb{E}[M \mid \mathcal{D}_m] \leq e^4 \nu$. Moreover, $M_\tau \mid \mathcal{F}_{\tau-} \sim \text{Poisson}(\rho_\tau)$. Iterating its conditional Laplace transform gives, for $s > 0$,

$$\mathbb{E}[e^{-sM} \mid \mathcal{D}_m] \leq \exp\big((e^{-s} - 1)\nu\big), \qquad \mathbb{P}(M < \nu/2 \mid \mathcal{D}_m) \leq e^{-\nu/8},$$

where the last bound follows by taking $s = \log 2$ and applying Markov's inequality.

Step 5 (bound $q^\star_{\max}/q^2_{\min}$). Under Assumption 1, for any $\|\vartheta\|_2 \leq 1$, $q_i(S_m; \vartheta) \in [\frac{e^{-2}}{m}, \frac{e^2}{m}]$. Thus $q_{\min} \geq e^{-2}/m$, $q^\star_{\max} \leq e^2/m$, and $\frac{q^\star_{\max}}{q^2_{\min}} \leq e^6 m$.

Step 6 (combining results). On $\{M \geq \nu/2\}$ use Step 3, and on its complement use $\|u\|_2^2 \leq 4$, which also covers $M = 0$. Steps 3–5 give

$$\mathbb{E}[\|u\|_2^2 \mid \mathcal{D}_m] \leq \frac{16e^4 q^\star_{\max} d}{q^2_{\min} b\nu} + 4e^{-\nu/8} \leq C\frac{dm}{b\nu}.$$

The last inequality uses $b \leq \mathrm{tr}(\widetilde{B}_m)/d \leq m/d$ and $e^{-\nu/8} \leq 8/(e\nu)$, with a numerical constant $C > 0$. Finally, $\nu \geq e^{-2}\lambda m(t-m+1)/(t+1)$, proving (23) after absorbing constants into C_1. □

D.4. Regret guarantees

The key proof idea is to relate the regret of greedy exploitation to the estimation error. We first establish the expected mean-square error of the estimator in Proposition 8, which depends on the smallest eigenvalue of the centered design matrix $\widetilde{B}_m$. Then Proposition 7 establishes that $\lambda_{\min}(\widetilde{B}_m)$ grows at rate $\Theta(m/d)$ with high probability under an isotropic feature condition. Combining these results with a proper choice of exploration length m leads to the $O(n^{2/3})$ regret upper bound in Theorem 6.

Theorem 6 (ETE-FSE Frequentist Regret Bound) *Under Assumption 1 and Assumption 6, assume* $m \geq 8\kappa_1 d \cdot \max\left\{\log\frac{2dn}{\delta},\ 2\left(d\log 5 + \log\frac{2n}{\delta}\right)\right\}$. *Then, with probability at least* $1-\delta$ *over the candidate features, there exists a constant* $C > 0$ *depending only on* κ_1 *such that, for any* $\lambda \in [\underline{\lambda}, \overline{\lambda}]$, $\theta \in \Theta$

$$\mathrm{Regret}_n^{FIQ}\left(\pi^{\text{ETE-FSE}}; \lambda, \theta\right) \leq Cm + C\frac{d}{\sqrt{\underline{\lambda}}}\sqrt{\frac{(n-m)^2}{m} + (n-m)\log(n-m+1)}. \tag{24}$$

In particular, taking $m = \lceil n^{2/3} \rceil$, *there exists a constant* $C' > 0$ *such that*

$$\mathrm{Regret}_n^{FIQ}\left(\pi^{\text{ETE-FSE}}; \lambda, \theta\right) \leq C'\left(1 + \frac{d}{\sqrt{\underline{\lambda}}}\right) n^{2/3}.$$

Proof. Let $X := (x_i)_{i\in\mathcal{N}}$ and $\mathcal{E}_m := \{\lambda_{\min}(\widetilde{B}_m) \geq C_2 m/(\kappa_1 d)\}$. Uniform exploration of i.i.d. candidate features gives i.i.d. selected features, so Proposition 7, applied with failure probability δ/n, gives $\mathbb{P}(\mathcal{E}_m^c) \leq \delta/n$. Writing $r_m(X) := \mathbb{P}(\mathcal{E}_m^c \mid X)$, Markov's inequality yields $\mathbb{P}_X(r_m(X) > 1/n) \leq \delta$. Fix a candidate set with $r_m(X) \leq 1/n$; this event does not depend on (λ, θ). The pathwise completion argument in the proof of Lemma 7 applies to any policy and bounds terminal loss by the sum of greedy attraction gaps. Under Assumption 1, the per-round regret is bounded by $2(e - e^{-1})$: for any two arms i, j, $|\langle x_i - x_j, \theta\rangle| \leq \|x_i - x_j\| \cdot \|\theta\| \leq 2$. Hence the exploration phase contributes at most $2(e-e^{-1})m$.

For each $t = m, \dots, n-1$, let $i_t^\star \in \arg\max_{i \in \mathcal{N} \setminus S_t} \langle x_i, \theta \rangle$ and let $\hat{i}_t := a_{t+1}$ be ETE-FSE's next choice. By optimality of $\hat{i}_t$ under $\hat{\theta}_t^m$, $\langle x_{i_t^\star}, \hat{\theta}_t^m \rangle \le \langle x_{\hat{i}_t}, \hat{\theta}_t^m \rangle$. Add and subtract θ to obtain

$$\begin{aligned}\langle x_{i_t^\star} - x_{\hat{i}_t}, \theta \rangle &= \langle x_{i_t^\star} - x_{\hat{i}_t}, \theta - \hat{\theta}_t^m \rangle + \langle x_{i_t^\star} - x_{\hat{i}_t}, \hat{\theta}_t^m \rangle \\ &\le \langle x_{i_t^\star} - x_{\hat{i}_t}, \theta - \hat{\theta}_t^m \rangle \;\le\; \|x_{i_t^\star} - x_{\hat{i}_t}\| \, \|\hat{\theta}_t^m - \theta\| \;\le\; 2\|\hat{\theta}_t^m - \theta\|.\end{aligned}$$

Therefore,

$$\sum_{t=m}^{n-1} \langle x_{i_t^\star} - x_{\hat{i}_t}, \theta \rangle \;\le\; 2 \sum_{t=m}^{n-1} \|\hat{\theta}_t^m - \theta\|. \tag{25}$$

Since both utilities lie in $[-1, 1]$, the corresponding attraction gap is at most e times the utility gap. By Cauchy–Schwarz, $\sum_{t=m}^{n-1} \|\hat{\theta}_t^m - \theta\| \;\le\; \sqrt{n-m} \, \sqrt{\sum_{t=m}^{n-1} \|\hat{\theta}_t^m - \theta\|^2}$. On $\mathcal{E}_m$, condition on $\mathcal{D}_m$. Taking expectations and applying Jensen's inequality gives

$$\mathbb{E}\Big[\sum_{t=m}^{n-1} \|\hat{\theta}_t^m - \theta\| \;\Big|\; \mathcal{D}_m \Big] \;\le\; \sqrt{n-m} \, \sqrt{\sum_{t=m}^{n-1} \mathbb{E}\big[\|\hat{\theta}_t^m - \theta\|^2 \;\big|\; \mathcal{D}_m \big]},$$

so Proposition 8 yields

$$\mathbb{E}\big[\|\hat{\theta}_t^m - \theta\|^2 \;\big|\; \mathcal{D}_m \big] \;\le\; C_1 \cdot \frac{d}{\lambda} \cdot \frac{\kappa_1 d}{C_2 m} \cdot \frac{t+1}{t-m+1} \;=:\; \frac{C_0 \kappa_1 d^2}{\lambda m} \cdot \frac{t+1}{t-m+1},$$

where $C_0 := C_1 / C_2$. Let $s = t - m + 1 \in \{1, \dots, n-m\}$. Then $\frac{t+1}{t-m+1} = \frac{m+s}{s} = 1 + \frac{m}{s}$. Hence

$$\sum_{t=m}^{n-1} \frac{t+1}{t-m+1} = (n-m) + m \sum_{s=1}^{n-m} \frac{1}{s} \;\le\; (n-m) + 2m \log(n-m+1).$$

Combining these bounds with (25) and the attraction-gap bound gives the right-hand side of (24) for the expected terminal loss conditional on $\mathcal{D}_m$ on $\mathcal{E}_m$. For the fixed candidate set, average over the exploration indices. The complement $\mathcal{E}_m^c$ contributes at most $n(e - e^{-1}) r_m(X) \le e - e^{-1}$, which is absorbed into Cm. This proves (24), with C depending only on κ_1. With $m = \lceil n^{2/3} \rceil$, the exploration term is $O(n^{2/3})$ and the second term in (24) is also $O\big(\frac{d}{\sqrt{\underline{\lambda}}} n^{2/3}\big)$. □

Corollary 7 (ETE-FSE Bayesian Regret Bound) *Under the same assumptions as Theorem 6, with probability at least $1 - \delta$ over the candidate features, there exists a constant $C > 0$ such that*

$$\mathrm{BayesRegret}_n^{F^{IQ}}\big(\pi^{\text{ETE-FSE}}\big) \le Cm + C \frac{d}{\sqrt{\underline{\lambda}}} \sqrt{\frac{(n-m)^2}{m} + (n-m)\log(n-m+1)}.$$

In particular, taking $m = \lceil n^{2/3} \rceil$, there exists a constant $C' > 0$ such that

$$\mathrm{BayesRegret}_n^{F^{IQ}}\big(\pi^{\text{ETE-FSE}}\big) \le C' \left(1 + \frac{d}{\sqrt{\underline{\lambda}}}\right) n^{2/3}.$$

Appendix E: Supporting Lemmas, Propositions, and Proof of Theorem 4

E.1. Technical Lemmas

Lemma 4 (One-period mutual information inequality) *Let G be any probability law supported on $[\underline{\lambda}, \overline{\lambda}] \times \Theta$, and $(\lambda, \theta) \sim G$. For a nonempty set S of size s, let $\boldsymbol{Y}^S$ be the feedback generated by the Poisson–MNL model with parameters (λ, θ). Then*

$$I((\lambda, \theta); \boldsymbol{Y}^S) \geq \kappa_0 b_s \operatorname{tr}\left(B_S \operatorname{Cov}(\theta)\right), \tag{26}$$

where the mutual information $I(\cdot;\cdot)$ and covariance $\operatorname{Cov}(\cdot)$ are induced by G and the model.

Proof. We separate the proof into the Poisson, categorical, and mutual information steps. Denote the normalized squared Hellinger distance as $h^2(P, Q) := 1 - \int \sqrt{dP\, dQ}$. First, the exact Hellinger formula for product-Poisson laws separates the arrival from the conditional inside-choice distribution. Second, the categorical step converts this categorical Hellinger distance into a constant multiple of $(\theta - \theta')^\top B_S (\theta - \theta')$. Third, we bound the mutual information below by the expected Hellinger distance between two independent draws.

1. Poisson step. For nonnegative mean vectors μ, μ', Poisson distributions have $h^2(P_\mu, P_{\mu'}) = 1 - \exp\left[-\frac{1}{2}\|\sqrt{\mu} - \sqrt{\mu'}\|_2^2\right]$. Consider parameter pairs (λ, θ) and (λ', θ'). Their inside-count mean vectors are $\mu = \rho q(S;\theta)$ and $\mu' = \rho' q(S;\theta')$, where $\rho = \lambda p_{\text{in}}(S;\theta)$ and similarly for ρ'. Expanding coordinate by coordinate,

$$\begin{aligned}\|\sqrt{\mu} - \sqrt{\mu'}\|_2^2 &= \rho + \rho' - 2\sqrt{\rho\rho'} \sum_{i \in S} \sqrt{q_i(S;\theta) q_i(S;\theta')} \\ &= (\sqrt{\rho} - \sqrt{\rho'})^2 + 2\sqrt{\rho\rho'}\, h^2(q(S;\theta), q(S;\theta')).\end{aligned}$$

Under Assumption 1, $\rho, \rho' \geq \rho_s := \frac{\lambda s}{e+s}$. Let $h_q^2 = h^2(q(S;\theta), q(S;\theta')) \in [0,1]$. Then $\frac{1}{2}\|\sqrt{\mu} - \sqrt{\mu'}\|_2^2 \geq \sqrt{\rho\rho'} h_q^2 \geq \rho_s h_q^2$. Combining this with the Poisson Hellinger distance yields $h^2(P_\mu, P_{\mu'}) \geq 1 - e^{-\rho_s h_q^2}$. The function $u \mapsto 1 - e^{-\rho_s u}$ is concave on $[0,1]$, so

$$h^2(P_\mu, P_{\mu'}) \geq (1 - e^{-\rho_s}) h_q^2 = b_s h^2(q(S;\theta), q(S;\theta')).$$

2. Categorical step. Let $\psi(u) = \log \sum_{i \in S} e^{x_i^\top u}$ and $v = \theta - \theta'$. For the conditional inside-choice distributions, $q_i(S;\theta) = \exp\left\{x_i^\top \theta - \psi(\theta)\right\}$. Their Hellinger affinity is

$$\begin{aligned}\sum_{i \in S} \sqrt{q_i(S;\theta) q_i(S;\theta')} &= \exp\left\{-\frac{\psi(\theta) + \psi(\theta')}{2}\right\} \sum_{i \in S} \exp\left\{x_i^\top \frac{\theta + \theta'}{2}\right\} \\ &= \exp\left\{\psi\left(\frac{\theta + \theta'}{2}\right) - \frac{\psi(\theta) + \psi(\theta')}{2}\right\} = e^{-J},\end{aligned}$$

where $J := \frac{\psi(\theta)+\psi(\theta')}{2} - \psi\left(\frac{\theta+\theta'}{2}\right)$. Consequently, $h^2(q(S;\theta), q(S;\theta')) = 1 - e^{-J}$.

We first prove $\nabla^2\psi(u) \succeq e^{-2}B_S$. For every point u on the segment joining θ and θ', we have $\|u\|_2 \le 1$, and $q_i(S;u) \ge \frac{e^{-1}}{se} = \frac{e^{-2}}{s}$. To compare the corresponding covariance with the design covariance, let $z \in \mathbb{R}^d$. Using the variational characterization of variance,

$$z^\top \nabla^2\psi(u) z = z^\top \operatorname{Cov}_{q(S;u)}(x_i) z = \min_{c\in\mathbb{R}} \sum_{i\in S} q_i(S;u)(z^\top x_i - c)^2$$
$$\ge e^{-2} \min_{c\in\mathbb{R}} \frac{1}{s} \sum_{i\in S} (z^\top x_i - c)^2 = e^{-2} z^\top B_S z.$$

Because this inequality holds for every $z \in \mathbb{R}^d$, it follows that $\nabla^2\psi(u) = \operatorname{Cov}_{q(S;u)}(x_i) \succeq e^{-2}B_S$.

To track the constant J, let $f(\alpha) = \psi(\theta' + \alpha v)$, $0 \le \alpha \le 1$. Then $f''(\alpha) \ge e^{-2} v^\top B_S v$. Applying the midpoint inequality to the convex function $f(\alpha) - \alpha^2 e^{-2} v^\top B_S v/2$ gives

$$J = \frac{f(0)+f(1)}{2} - f\left(\frac{1}{2}\right) \ge \frac{e^{-2}}{8} v^\top B_S v.$$

Next we linearize the exponential bound. Assumption 1 gives $B_S \preceq I_d$ and $0 \le v^\top B_S v \le 4$. The function $x \mapsto 1 - \exp[-x/(8e^2)]$ is concave, so on $[0,4]$ we have $1 - \exp\left(-\frac{x}{8e^2}\right) \ge \frac{1-\exp[-1/(2e^2)]}{4} x$,

$$h^2(q(S;\theta), q(S;\theta')) \ge \frac{1-\exp[-1/(2e^2)]}{4} (\theta-\theta')^\top B_S (\theta-\theta').$$

3. Mutual information step. Let P_G^S denote the marginal law of $\boldsymbol{Y}^S$ obtained by averaging the channel law over G, $P_G^S(\boldsymbol{y}) = \mathbb{E}_G\left[P_{\lambda,\theta}^S(\boldsymbol{y})\right]$. The conditional representation of mutual information is $I_G((\lambda,\theta);\boldsymbol{Y}^S) = \mathbb{E}_G\left[D_{\mathrm{KL}}\left(P_{\lambda,\theta}^S \middle\| P_G^S\right)\right]$. Since $D_{\mathrm{KL}}(P\|Q) \ge 2h^2(P,Q)$,

$$I_G((\lambda,\theta);\boldsymbol{Y}^S) \ge 2\mathbb{E}_G\left[h^2\left(P_{\lambda,\theta}^S, P_G^S\right)\right].$$

Let (λ',θ') be an independent draw from G. Then

$$\mathbb{E}_G\left[h^2\left(P_{\lambda,\theta}^S, P_G^S\right)\right] = 1 - \sum_{\boldsymbol{y}\in\mathbb{N}_0^S} \sqrt{P_G^S(\boldsymbol{y})}\, \mathbb{E}_G\left[\sqrt{P_{\lambda,\theta}^S(\boldsymbol{y})}\right],$$
$$\mathbb{E}\left[h^2\left(P_{\lambda,\theta}^S, P_{\lambda',\theta'}^S\right)\right] = 1 - \sum_{\boldsymbol{y}\in\mathbb{N}_0^S} \left(\mathbb{E}_G\left[\sqrt{P_{\lambda,\theta}^S(\boldsymbol{y})}\right]\right)^2.$$

Because $\sum_{\boldsymbol{y}\in\mathbb{N}_0^S} P_G^S(\boldsymbol{y}) = 1$, subtracting the second expression from twice the first gives

$$2\mathbb{E}_G\left[h^2\left(P_{\lambda,\theta}^S, P_G^S\right)\right] - \mathbb{E}\left[h^2\left(P_{\lambda,\theta}^S, P_{\lambda',\theta'}^S\right)\right] = \sum_{\boldsymbol{y}\in\mathbb{N}_0^S} \left(\sqrt{P_G^S(\boldsymbol{y})} - \mathbb{E}_G\left[\sqrt{P_{\lambda,\theta}^S(\boldsymbol{y})}\right]\right)^2 \ge 0.$$

Therefore, $I_G((\lambda,\theta);\boldsymbol{Y}^S) \geq \mathbb{E}\left[h^2\left(P^S_{\lambda,\theta}, P^S_{\lambda',\theta'}\right)\right]$. Combining this with the preceding Poisson and categorical step bounds gives

$$I_G((\lambda,\theta);\boldsymbol{Y}^S) \geq b_s \frac{1-\exp[-1/(2e^2)]}{4} \mathbb{E}_G\left[(\theta-\theta')^\top B_S(\theta-\theta')\right].$$

Because θ and θ' are i.i.d., $\mathbb{E}_G[(\theta-\theta')^\top B_S(\theta-\theta')] = \operatorname{tr}(B_S \mathbb{E}_G[(\theta-\theta')(\theta-\theta')^\top]) = 2\operatorname{tr}(B_S \operatorname{Cov}_G(\theta))$. Therefore, $I_G((\lambda,\theta);\boldsymbol{Y}^S) \geq \kappa_0 b_s \operatorname{tr}(B_S \operatorname{Cov}_G(\theta))$.

Lemma 5 (Rank-one update of the centered design) *For every available candidate $a \in \mathcal{A}_t$,*

$$tB_{S_{t-1}\cup\{a\}} = \widetilde{B}_{t-1} + \frac{t-1}{t}(x_a - \bar{x}_{t-1})(x_a - \bar{x}_{t-1})^\top, \tag{27}$$

Consequently, $\mathbb{E}_t[B_t] = \frac{1}{t}\left[\widetilde{B}_{t-1} + \frac{t-1}{t}\mathbb{E}_t\left\{(x_{a_t} - \bar{x}_{t-1})(x_{a_t} - \bar{x}_{t-1})^\top\right\}\right]$. Moreover,

$$\mathbb{E}_t\left\{(x_{a_t} - \bar{x}_{t-1})(x_{a_t} - \bar{x}_{t-1})^\top\right\} \succeq \operatorname{Cov}_t(x_{a_t^\star}).$$

Proof. The new feature mean is $\bar{x}_{S_{t-1}\cup\{a\}} = \bar{x}_{S_{t-1}} + \frac{x_a - \bar{x}_{S_{t-1}}}{t}$. Expanding every centered square about $\bar{x}_{S_{t-1}}$, the cross terms vanish because $\sum_{i\in S_{t-1}}(x_i - \bar{x}_{S_{t-1}}) = 0$. The remaining terms give (27). Divide by t and average over the conditional law of a_t to obtain the first identity. Finally, the standard covariance decomposition around the fixed vector $\bar{x}_{t-1}$ gives $\mathbb{E}_t[(x_{a_t} - \bar{x}_{t-1})(x_{a_t} - \bar{x}_{t-1})^\top] = \operatorname{Cov}_t(x_{a_t}) + (\mathbb{E}_t x_{a_t} - \bar{x}_{t-1})(\mathbb{E}_t x_{a_t} - \bar{x}_{t-1})^\top$. The second term is positive semi-definite, and probability matching gives $\operatorname{Cov}_t(x_{a_t}) = \operatorname{Cov}_t(x_{a_t^\star})$.

Lemma 6 (Parameter uncertainty bound via information gain) *For every $t \geq 2$,*

$$\mathcal{I}_t \geq \frac{\kappa_0 b_t}{t} \operatorname{tr}(\widetilde{B}_{t-1}\Sigma_t).$$

Proof. For $a \in \mathcal{A}_t$ with $\pi_{a,t} > 0$, condition on $a_t = a$. The active set is the fixed size-t set $S_{t-1} \cup \{a\}$, so Lemma 4 gives

$$I_t((\lambda,\theta);\boldsymbol{Y}_t \mid a_t = a) \geq \kappa_0 b_t \operatorname{tr}(B_{S_{t-1}\cup\{a\}}\Sigma_t).$$

By the conditional chain rule, $\mathcal{I}_t$ is the $\pi_{a,t}$-weighted average of these fixed-action information gains, so $\mathcal{I}_t \geq \kappa_0 b_t \operatorname{tr}(\mathbb{E}_t[B_t]\Sigma_t)$. Lemma 5 then gives $\mathbb{E}_t[B_t] \succeq \frac{1}{t}\widetilde{B}_{t-1}$, and $\mathcal{I}_t \geq \frac{\kappa_0 b_t}{t}\operatorname{tr}(\widetilde{B}_{t-1}\Sigma_t)$.

E.2. Proofs and Supporting Lemmas for Thompson Sampling

Lemma 7 (Regret decomposition)

$$\mathrm{BayesRegret}_n^{FIQ}(\pi^{\mathrm{TS}}) \le \sum_{t=1}^{n} \mathbb{E}[\Delta_t].$$

Proof. Write $w_i = e^{x_i^\top \theta}$. For a selected prefix S_t, define the value of its best size-n completion

$$\Phi_t := \sum_{i \in S_t} w_i + \max_{\substack{C \subseteq \mathcal{N} \setminus S_t \\ |C| = n-t}} \sum_{i \in C} w_i.$$

Then Φ_0 is the true top-n total and Φ_n is the selected total. Let C_{t-1} attain the maximum defining Φ_{t-1}. If $a_t \in C_{t-1}$, fixing a_t does not change the best completed set, so $\Phi_t = \Phi_{t-1}$. Otherwise, remove a least-weight member of C_{t-1} and replace it by a_t. Because $a_t^\star$ is a highest-weight remaining candidate, the resulting loss is no larger than $w_{a_t^\star} - w_{a_t}$. Hence $\Phi_{t-1} - \Phi_t \le w_{a_t^\star} - w_{a_t}$. Telescoping and taking the expectation proves the lemma.

Proof of Lemma 2. Because $a_t^\star$ is greedy under θ, $x_{a_t^\star}^\top \theta \ge x_{a_t}^\top \theta$. Both sides lie in $[-1, 1]$, so the mean-value theorem gives $e^{x_{a_t^\star}^\top \theta} - e^{x_{a_t}^\top \theta} \le e(x_{a_t^\star} - x_{a_t})^\top \theta$. The Thompson action is conditionally independent of θ, and its conditional law is $(\pi_{a,t})_{a \in \mathcal{A}_t}$. Therefore

$$\begin{aligned}\frac{\Delta_t}{e} &\le \sum_a \pi_{a,t} x_a^\top m_t(a) - \left(\sum_a \pi_{a,t} x_a\right)^\top \mathbb{E}_t[\theta] \\ &= \sum_a \pi_{a,t} \big(x_a - \mathbb{E}_t x_{a_t^\star}\big)^\top \big(m_t(a) - \mathbb{E}_t[\theta]\big),\end{aligned}$$

where the last equality uses $\sum_a \pi_{a,t} m_t(a) = \mathbb{E}_t[\theta]$. Next, insert $(4I_d + \widetilde{B}_{t-1})^{-1/2}(4I_d + \widetilde{B}_{t-1})^{1/2}$ into the expression and apply Cauchy–Schwarz. Pointwise on the conditional history,

$$\left(\frac{\Delta_t}{e}\right)^2 \le \left(\sum_a \pi_{a,t} \|x_a - \mathbb{E}_t x_{a_t^\star}\|^2_{(4I_d + \widetilde{B}_{t-1})^{-1}}\right) \cdot \left(\sum_a \pi_{a,t} \|m_t(a) - \mathbb{E}_t \theta\|^2_{4I_d + \widetilde{B}_{t-1}}\right).$$

Probability matching shows that the first factor equals ℓ_t and simplifies the second factor to a matrix trace. This gives (11). The law of total covariance states

$$\Sigma_t = \sum_{a : \pi_{a,t} > 0} \pi_{a,t} \operatorname{Cov}_t(\theta \mid a_t^\star = a) + \operatorname{Cov}_t(m_t(a_t^\star)),$$

so $\operatorname{Cov}_t(m_t(a_t^\star)) \preceq \Sigma_t$. Therefore the trace term in (11) is at most $\operatorname{tr}((4I_d + \widetilde{B}_{t-1})\Sigma_t)$. By Lemma 6, $\operatorname{tr}(\widetilde{B}_{t-1}\Sigma_t) \le \frac{t\mathcal{I}_t}{\kappa_0 b_t}$. Moreover, $\operatorname{tr}(\Sigma_t) = \mathbb{E}_t\|\theta - \mathbb{E}_t\theta\|_2^2 \le \mathbb{E}_t\|\theta\|_2^2 \le 1$. It follows that

$$\operatorname{tr}((4I_d + \widetilde{B}_{t-1})\Sigma_t) \le 4 + \frac{t\mathcal{I}_t}{\kappa_0 b_t}.$$

Substituting this inequality into (11) proves the result.

Lemma 8 (Elliptical potential for the centered design) *For every $n \geq 2$,*

$$\sum_{t=2}^{n} \mathbb{E}[\ell_t] \leq \frac{d}{\log(3/2)} \log\left(1 + \frac{n}{4d}\right).$$

Proof. Fix $t \geq 2$, put $r_t = (t-1)/t$, and define the realized rank-one increment

$$u_t := r_t (x_{a_t} - \bar{x}_{t-1})^\top (4I_d + \widetilde{B}_{t-1})^{-1} (x_{a_t} - \bar{x}_{t-1}).$$

We first relate ℓ_t to u_t. By Lemma 5,

$$r_t \ell_t \leq r_t \operatorname{tr}\left(\mathbb{E}_t[(x_{a_t} - \bar{x}_{t-1})(x_{a_t} - \bar{x}_{t-1})^\top](4I_d + \widetilde{B}_{t-1})^{-1}\right) = \mathbb{E}_t[u_t].$$

Because $4I_d + \widetilde{B}_{t-1} \succeq 4I_d$ and $\|x_{a_t} - \bar{x}_{t-1}\|_2 \leq 2$, we have $0 \leq u_t \leq r_t$. Concavity of $u \mapsto \log(1+u)$ on $[0, r_t]$ gives

$$\frac{\log(1+u)}{\log(1+r_t)} \geq \frac{u}{r_t}, \qquad 0 \leq u \leq r_t.$$

Since $r_t \geq 1/2$, combining this inequality with the preceding trace bound yields

$$\ell_t \leq \frac{\mathbb{E}_t[u_t]}{r_t} \leq \frac{\mathbb{E}_t[\log(1+u_t)]}{\log(1+r_t)} \leq \frac{1}{\log(3/2)} \mathbb{E}_t[\log(1+u_t)].$$

By the rank-one update (27), $4I_d + \widetilde{B}_t = 4I_d + \widetilde{B}_{t-1} + r_t(x_{a_t} - \bar{x}_{t-1})(x_{a_t} - \bar{x}_{t-1})^\top$. The matrix determinant lemma therefore gives

$$\log(1+u_t) = \log\det(4I_d + \widetilde{B}_t) - \log\det(4I_d + \widetilde{B}_{t-1}).$$

Take expectations and sum over $t = 2, \ldots, n$ to telescope the determinant terms. Because $\widetilde{B}_1 = 0$,

$$\sum_{t=2}^{n} \mathbb{E}[\ell_t] \leq \frac{1}{\log(3/2)} \mathbb{E}[\log\det(I_d + \widetilde{B}_n/4)].$$

For every realized selected set, $\operatorname{tr}(\widetilde{B}_n) = \sum_{i \in S_n} \|x_i - \bar{x}_n\|_2^2 = \sum_{i \in S_n} \|x_i\|_2^2 - n\|\bar{x}_n\|_2^2 \leq n$. If $\xi_1, \ldots, \xi_d$ are the eigenvalues of $\widetilde{B}_n$, concavity of the logarithm gives

$$\log\det(I_d + \widetilde{B}_n/4) = \sum_{j=1}^{d} \log(1 + \xi_j/4) \leq d \log\left(1 + \frac{\operatorname{tr}(\widetilde{B}_n)}{4d}\right) \leq d \log\left(1 + \frac{n}{4d}\right).$$

Substituting into the preceding determinant bound finishes the proof.

Lemma 9 (Mutual Information Upper Bound) *Under Assumption 1,*

$$I((\lambda, \theta); \mathbf{H}_n) \leq d \log(1 + 8Q_n) + \log\left(1 + 2Q_n \log\frac{\overline{\lambda}}{\underline{\lambda}}\right) + \frac{1}{2}. \tag{28}$$

Proof. We quantize the joint parameter and decompose its information into the part carried by the quantized pair and the residual within each quantization cell. Set $\varepsilon = Q_n^{-1} \le 1$ and $\eta = \log \lambda$. Choose a maximal $\varepsilon/4$-separated subset of the Euclidean unit ball and a maximal $\varepsilon/2$-separated subset of $[\log \underline{\lambda}, \log \overline{\lambda}]$. Let θ_ε and η_ε be nearest centers to θ and η, respectively, with deterministic tie-breaking, and set $\lambda_\varepsilon = e^{\eta_\varepsilon}$. The mapping satisfies $\|\theta - \theta_\varepsilon\|_2 \le \varepsilon/4$ and $|\log \lambda - \log \lambda_\varepsilon| \le \varepsilon/2$. Because $(\lambda_\varepsilon, \theta_\varepsilon)$ is a function of (λ, θ), the chain rule gives

$$I((\lambda,\theta);\mathbf{H}_n) = I((\lambda_\varepsilon,\theta_\varepsilon);\mathbf{H}_n) + I((\lambda,\theta);\mathbf{H}_n \mid \lambda_\varepsilon,\theta_\varepsilon).$$

We bound these two terms in order.

For the first term, the volume comparison bounds the number of preference centers by $(1+8/\varepsilon)^d$: the $\varepsilon/8$-balls around the separated points are disjoint and lie in the ball of radius $1+\varepsilon/8$. The rate net has at most $1 + (2/\varepsilon)\log(\overline{\lambda}/\underline{\lambda})$ points. Consequently,

$$I((\lambda_\varepsilon,\theta_\varepsilon);\mathbf{H}_n) \le H(\lambda_\varepsilon,\theta_\varepsilon) \le d\log(1+8Q_n) + \log\left(1 + 2Q_n \log\frac{\overline{\lambda}}{\underline{\lambda}}\right).$$

For the residual term, fix a nonempty active set S and a supported realization (λ, θ), with assigned center $(\lambda_\varepsilon, \theta_\varepsilon)$. For $i \in S$, let $\mu_i = \lambda p_i(S;\theta)$ and $\mu_{i,\varepsilon} = \lambda_\varepsilon p_i(S;\theta_\varepsilon)$. The log-rate and linear-feature terms in $\log \mu_i$ change by at most $\varepsilon/2$ and $\varepsilon/4$, respectively. Moreover, the log-partition function is 1-Lipschitz and changes by at most $\varepsilon/4$. Hence $|\log(\mu_i/\mu_{i,\varepsilon})| \le \varepsilon$.

Set $u_{i,\varepsilon} = \log(\mu_i/\mu_{i,\varepsilon})$ and $f(u) = u - 1 + e^{-u}$. Since $f(0) = f'(0) = 0$ and $f''(u) = e^{-u} \le e$ on $[-1,1]$, Taylor's theorem and the product-Poisson KL formula give

$$D_{\mathrm{KL}}(P^S_{\lambda,\theta} \| P^S_{\lambda_\varepsilon,\theta_\varepsilon}) = \sum_{i\in S} \mu_i f(u_{i,\varepsilon}) \le \frac{e}{2}\sum_{i\in S}\mu_i u_{i,\varepsilon}^2 \le \frac{e\overline{\lambda}\varepsilon^2}{2},$$

where the last inequality uses $\sum_{i\in S}\mu_i = \lambda p_{\mathrm{in}}(S;\theta) \le \overline{\lambda}$.

It remains to lift the one-period bound to the complete observed history. Let $P^{(n)}_{\lambda,\theta}$ be the law of $\mathbf{H}_n$ under the fixed environment (λ,θ) and π^{TS}, and let $\mathbb{E}_{\lambda,\theta}$ denote expectation under this law. Given the same past, both path laws use the same action kernel $\pi_t^{\mathrm{TS}}(\cdot \mid \mathbf{H}_{t-1})$, so it cancels from the likelihood ratio. The adaptive KL chain rule gives

$$D_{\mathrm{KL}}(P^{(n)}_{\lambda,\theta} \| P^{(n)}_{\lambda_\varepsilon,\theta_\varepsilon}) = \sum_{t=1}^{n} \mathbb{E}_{\lambda,\theta}\left[D_{\mathrm{KL}}\left(P^{S_t}_{\lambda,\theta}\middle\| P^{S_t}_{\lambda_\varepsilon,\theta_\varepsilon}\right)\right] \le \frac{ne\overline{\lambda}\varepsilon^2}{2} \le \frac{1}{2}, \tag{29}$$

where the last inequality uses $Q_n^2 \ge en\overline{\lambda}$.

For each value $c = (\lambda_c, \theta_c)$ of the quantized pair having positive probability, define its conditional mixture path law, for every measurable history event A, by $\overline{P}_c^{(n)}(A) := \mathbb{E}\left[P_{\lambda,\theta}^{(n)}(A) \mid (\lambda_\varepsilon, \theta_\varepsilon) = c\right]$. Adding and subtracting the log density of this mixture gives

$$\mathbb{E}[D_{\mathrm{KL}}(P_{\lambda,\theta}^{(n)} \| P_{\lambda_\varepsilon,\theta_\varepsilon}^{(n)})] = I((\lambda,\theta); \mathbf{H}_n \mid \lambda_\varepsilon, \theta_\varepsilon) + \mathbb{E}[D_{\mathrm{KL}}(\overline{P}_{\lambda_\varepsilon,\theta_\varepsilon}^{(n)} \| P_{\lambda_\varepsilon,\theta_\varepsilon}^{(n)})].$$

The last term is nonnegative, and the left-hand side is at most $1/2$ by (29). Thus the residual information is at most $1/2$. Combining the two bounds with the decomposition proves (28).

E.3. Proof of Corollary 1

Proof. For $x \geq 0$, concavity on $[0,1]$ and monotonicity afterward give $1 - e^{-x} \geq (1 - e^{-1}) \min\{x, 1\}$. Since $n/(e+n) \geq 2/(e+2)$ for $n \geq 2$,

$$b_n \geq 1 - \exp\left(-\frac{2\underline{\lambda}}{e+2}\right) \geq (1 - e^{-1}) \min\left\{\frac{2\underline{\lambda}}{e+2}, 1\right\} \geq \frac{2(1-e^{-1})}{e+2} \min\{\underline{\lambda}, 1\}.$$

Substitution into (15) gives (16).

Appendix F: Proof of Proposition 6

Proof. Let $W_{t-1} := \sum_{j=1}^{t-1} e^{c_j/2}$ be the total attraction of the baseline candidates. Since $c_j \geq 3/4$, $W_{t-1} \geq (t-1)e^{3/8}$.

Condition first on $a_t = i_+$; the other action is symmetric. Enlarge the feedback by revealing the outside count. For one arrival, the probabilities of choosing i_+ under θ_+ and θ_- are

$$p_+ = \frac{e^r}{1 + W_{t-1} + e^r}, \qquad p_- = \frac{e^{-r}}{1 + W_{t-1} + e^{-r}},$$

respectively. Conditional on not choosing i_+, the distribution over the outside option and the baseline candidates is the same under both parameter values: its probabilities are proportional to $1, e^{c_1/2}, \ldots, e^{c_s/2}$. Thus all information about the sign of the second coordinate is contained in the binary event that an arrival chooses i_+. The two categorical KL divergences therefore reduce to the Bernoulli divergences between p_+ and p_-. For two product-Poisson laws with common total rate λ, KL is λ times the KL between their categorical probability vectors. Therefore,

$$D_{\mathrm{KL}}(P_+ \| P_-) + D_{\mathrm{KL}}(P_- \| P_+) = 2\lambda r (p_+ - p_-),$$

because the two Bernoulli log odds differ by $2r$. Direct subtraction gives

$$p_+ - p_- = \frac{(1 + W_{t-1})(e^r - e^{-r})}{(1 + W_{t-1} + e^r)(1 + W_{t-1} + e^{-r})} \leq \frac{e^r - e^{-r}}{W_{t-1}}.$$

Let P_+ and P_- be the corresponding product-Poisson laws of the augmented count vector, and let $P_0 = (P_+ + P_-)/2$. Convexity of KL in its second argument gives

$$\frac{1}{2}D_{\mathrm{KL}}(P_+\|P_0) + \frac{1}{2}D_{\mathrm{KL}}(P_-\|P_0) \le \frac{1}{4}\{D_{\mathrm{KL}}(P_+\|P_-) + D_{\mathrm{KL}}(P_-\|P_+)\}.$$

The left-hand side of the bound is the mutual information between the equiprobable sign and the augmented feedback. The preceding bounds show that this information is at most $\lambda r(e^r - e^{-r})/(2W_{t-1})$. The same conclusion holds conditional on $a_t = i_-$. Moreover, the Thompson action is independent of the true parameter given the history, and the data-processing inequality implies that deleting the outside count cannot increase mutual information. Hence

$$\mathcal{I}_t \le \frac{\lambda r(e^r - e^{-r})}{2W_{t-1}}.$$

Combining this inequality with $\Delta_t = (e^r - e^{-r})/2$ and $W_s \ge (t-1)e^{3/8}$ proves (9). □

Appendix G: Implementation Details for Gibbs Thompson Sampling

We provide the conditional distributions and elliptical-slice update omitted from Section 7.1. Define

$$M_{<t} := \sum_{\tau=1}^{t-1} M_\tau, \qquad \log \mathcal{L}^{\theta}_{<t}(\theta;\lambda) := \sum_{\tau=1}^{t-1}\left[\sum_{i\in S_\tau} y_{i\tau}\log p_i(S_\tau;\theta) - \lambda p_{\mathrm{in}}(S_\tau;\theta)\right].$$

λ-update. Conditional on the current θ,

$$G_t(\lambda \mid \theta) \propto \lambda^{M_{<t}} \exp\left\{-\lambda \sum_{\tau<t} p_{\mathrm{in}}(S_\tau;\theta)\right\} \mathbb{I}\{\underline{\lambda} \le \lambda \le \bar{\lambda}\}.$$

Thus, for $t \ge 2$, this is a Gamma distribution with shape $M_{<t} + 1$ and rate $\sum_{\tau<t} p_{\mathrm{in}}(S_\tau;\theta)$, truncated to $[\underline{\lambda}, \bar{\lambda}]$.

θ-update. Conditional on the current λ,

$$G_t(\theta \mid \lambda) \propto \mathcal{L}^{\theta}_{<t}(\theta;\lambda) \exp\left(-\frac{\|\theta\|_2^2}{2}\right) \mathbb{I}\{\theta \in \Theta\}.$$

We implement the elliptical-slice transition of Murray et al. (2010) as follows. Let $\theta_c \in \Theta$ be the current state. Draw $U \sim \mathrm{Unif}(0,1)$ and set

$$h = \log \mathcal{L}^{\theta}_{<t}(\theta_c;\lambda) + \log U.$$

Draw $\nu \sim \mathcal{N}(0, I_d)$ and $\phi \sim \mathrm{Unif}[0, 2\pi)$, and initialize

$$(\phi^-, \phi^+) = (\phi - 2\pi, \phi).$$

For a trial angle ϕ, form

$$\theta(\phi) = \theta_c \cos\phi + \nu \sin\phi.$$

Accept the proposal if

$$\theta(\phi) \in \Theta \qquad \text{and} \qquad \log \mathcal{L}^{\theta}_{<t}(\theta(\phi); \lambda) \geq h.$$

Otherwise, set $\phi^- = \phi$ if $\phi < 0$, and $\phi^+ = \phi$ otherwise, redraw $\phi \sim \mathrm{Unif}[\phi^-, \phi^+]$, and repeat until acceptance. The support condition $\theta(\phi) \in \Theta$ enforces the truncated Gaussian prior used in the experiments.