POSTERIOR TOP-K SELECTION UNDER NOISY REVIEWS: EQUIVALENCE AND GAIN RELATIVE TO POSTERIOR MEAN

Anonymous authors
Paper under double-blind review

ABSTRACT

Selection problems, including conference acceptance, grant review, and hiring, often require choosing exactly K candidates from noisy and heterogeneous evaluations. Ranking aggregate scores does not directly answer the decision-relevant question: how likely is each candidate to belong to the latent top K? We formalize this target through the posterior membership probability $p_i = \Pr(i \in \mathrm{TopK}(q) \mid D)$. Selecting the K largest p_i is Bayes-optimal for expected overlap with the latent top-K set. Taking this established decision rule as a starting point, we study when it differs from posterior-mean ranking, why it differs, and how much the resulting roster change matters. We show that exchangeable centered residuals yield a structural equivalence regime. For independent posteriors with continuous densities, we derive a boundary approximation under local regularity; its Gaussian case predicts uncertainty-driven reordering. For arbitrary joint posteriors, we further show that the consequence of disagreement depends jointly on how many candidates change and how far those candidates lie from the membership-probability boundary, and establish when this characterization stabilizes in growing populations. Simulations illustrate these theoretical predictions. Our analysis of ICLR 2026 reviews finds identical rosters at the realized cohort capacity, consistent with nearly homogeneous posterior uncertainty across papers. The Jester results show occasional roster differences, but these yield only small estimated gains under the fitted posterior, while evaluation against an independent high-information reference does not establish a clear advantage. Overall, roster changes alone do not determine the value of posterior Top-K relative to posterior-mean ranking.

1 INTRODUCTION

Fixed-capacity selection in peer review, recommendation, screening, and prioritization requires choosing exactly K candidates from a pool of n, where candidates are indexed by $i \in \{1, \ldots, n\}$. Let D denote the observed data, let q_i denote candidate i's latent quality, and write $q = (q_1, \ldots, q_n)$. Posterior-mean ranking orders candidates by $m_i = \mathbb{E}[q_i \mid D]$. In contrast, posterior Top-K ranking orders them by $p_i = \mathbb{P}(i \in T_K(q) \mid D)$, where $T_K(q)$ is the set of indices corresponding to the K largest components of q. Thus, m_i measures expected latent quality, whereas p_i measures the posterior probability that candidate i belongs to the latent top-K set. Selecting the K candidates with the largest p_i is well known to be Bayes-optimal for maximizing expected overlap with the latent top-K set (Lin et al., 2006; Ferguson & Chang, 2020).

Bayes optimality, however, does not explain the relationship between the two rules. They may produce exactly the same roster, disagree only among nearly interchangeable candidates, or make genuinely consequential changes. We therefore ask: **Q1, equivalence:** when do the criteria coincide? **Q2, disagreement:** when and why do they differ? **Q3, magnitude:** how large is the resulting expected-overlap difference?

Our central quantitative insight is that roster disagreement has two distinct dimensions: how many candidates are affected and how far their membership probabilities lie from the selection boundary. We refer to these dimensions as *disagreement mass* and *disagreement depth*, respectively. Shallow disagreements involve candidates with nearly interchangeable membership probabilities, whereas

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

deep disagreements involve candidates whose membership probabilities lie farther from the selection cutoff. For any joint posterior, we show that

$$\frac{\Delta_D}{K} = \frac{\delta_n \bar{d}_n}{\alpha_n},$$

where Δ_D/K is the expected-overlap improvement per selected position, δ_n is the fraction of candidates involved in disagreement, $\bar{d}_n$ is their average distance from the membership-probability cutoff, and $\alpha_n = K/n$ is the fraction of candidates selected. Thus, many shallow changes can matter less than fewer deep changes. Formal definitions are given in Sections 3 and 4.3. We further establish conditions under which this finite-sample decomposition stabilizes as the candidate pool grows.

Our theoretical contributions proceed in three steps:

(1) Structural equivalence. We first identify conditions under which the two ranking rules agree. When posterior uncertainty can be represented by exchangeable centered residuals, candidates ordered by posterior mean are ordered in the same way by their Top-K membership probabilities, even when their qualities are dependent. Consequently, posterior-mean selection maximizes expected overlap with the latent top-K set (Section 4.1).

(2) Boundary geometry. We next explain how disagreement can arise. For independent posteriors with continuous densities, we approximate Top-K membership probabilities by tail probabilities at a common quality boundary under local regularity. In the Gaussian case, these approximate scores rank candidates by standardized boundary positions and can reverse posterior-mean ordering under heterogeneous uncertainty (Section 4.2).

(3) Disagreement magnitude. Finally, we quantify how much such disagreement matters. For any joint posterior, our finite-sample mass–depth identity separates the number of affected candidates from the decision-relevant size of the changes. We also provide a general population result showing when the disagreement mass and normalized expected-overlap difference stabilize as the candidate pool grows, and when the disagreement depth stabilizes as well (Section 4.3).

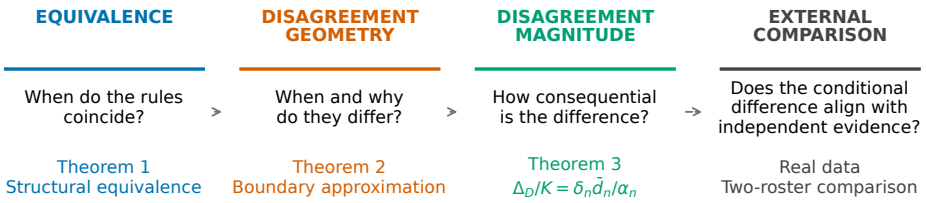


Figure 1: Overview of the paper. We move from structural equivalence, to the mechanism of disagreement, to its decision-relevant magnitude, and finally to external interpretation using independent evidence.

Simulations support the three theoretical stages. In real data, ICLR illustrates near-equivalence under weak fitted uncertainty heterogeneity, while Jester shows small model-conditional gains that are not resolved under independent-reference evaluation. Together, these results emphasize that disagreement matters through both its mass and depth.

2 Related Work

2.1 Bayesian ranking and loss-based selection

Bayesian ranking has long acknowledged that the appropriate ordering depends on the inferential target rather than on a single universal point estimate. Berger & Deely (1988) studied probabilities that related means are largest, while Laird & Louis (1989) developed empirical-Bayes rank estimates that incorporate posterior rank uncertainty. Shen & Louis (1998) made this target dependence explicit: posterior means are optimal for parameter estimation under squared-error loss, but can be inadequate when the goal is to recover ranks or the distribution of latent effects. Lin et al. (2006)

further developed loss functions tailored to percentile-based selection and ranking. Related empirical Bayes ideas have since been used for gene ranking and screening in high-throughput applications (Noma et al., 2010; Ferguson & Chang, 2020).

The work closest to our selection objective shifts attention from estimating ranks to recovering a top set. Henderson & Newton (2016) construct a ranking procedure that maximizes expected overlap between reported and latent top lists, using posterior probabilities of exceeding a population quantile. Gu & Koenker (2023) directly compare posterior-mean and posterior-tail selection under capacity constraints, showing equivalence in their homogeneous-variance model and divergence under heterogeneous precision. Their target is defined through posterior exceedance of a population threshold, whereas we study exact finite-pool Top-K membership under a common fitted joint posterior and characterize equivalence, disagreement geometry, and its expected-overlap magnitude.

2.2 Ranking uncertainty and top-K stability

Uncertainty in latent quality can make an observed ranking unstable even when its point estimates are well defined. Goldstein & Spiegelhalter (1996) emphasized the need to account for model-based uncertainty when interpreting institutional league tables, while Hall & Miller (2009) showed that conventional bootstrap procedures can be inconsistent for empirical ranks. Zuk et al. (2007) studied noisy-ranking reliability through top-K overlap and Kendall correlation, showing that top-K lists can be particularly sensitive to noise. Bowen (2022) developed Bayesian procedures for rank intervals and selection with error-rate control. These works ask how much confidence should be placed in a ranking or selected set. We address a distinct question: holding the posterior fixed, we compare two posterior ranking criteria and quantify the expected-overlap consequence of the candidates on which they disagree.

2.3 Sequential ranking-and-selection and active top-m identification

A separate literature studies how additional information should be acquired to identify the best or top-m alternatives efficiently. Confidence-based methods compare selected and omitted candidates through confidence bounds, as in PAC subset selection (Kalyanakrishnan et al., 2012). Bayesian ranking-and-selection work studies the value of information and posterior-based stopping (Frazier & Powell, 2010; Eckman & Henderson, 2022), while more recent methods develop sequential allocation rules for top-m selection (Zhang et al., 2023; Liu et al., 2024). Related settings include pairwise-comparison ranking (Xiao et al., 2023), top-k ranking observations in Bayesian optimization (Nguyen et al., 2021), and identifiability and experimental design for LLM judging (Xu et al., 2026). Unlike sequential ranking-and-selection, our main comparison holds the fitted posterior and capacity fixed, isolating the effect of the ranking criterion.

3 Problem Setup

We consider a fixed-capacity selection problem in which exactly K of n candidates must be selected after observing data D. Conditional on D, let $q = (q_1, \dots, q_n)$ denote their latent qualities with a specified finite-mean joint posterior. For $1 \le K < n$, define

$$m_i = \mathbb{E}[q_i \mid D], \quad p_i = \mathbb{P}(i \in T_K(q) \mid D),$$

where $T_K(q) = \mathrm{TopK}(q)$ is the latent top-K set, i.e., the set of K candidates with the largest underlying qualities q_i.

Posterior-mean ranking selects

$$A_{\text{mean}} = \mathrm{TopK}(m),$$

whereas posterior Top-K selects

$$A_{\text{p}} = \mathrm{TopK}(p).$$

The two scores answer different posterior questions: m_i summarizes expected latent quality, while p_i directly targets membership in the latent top-K set.

We evaluate any size-K roster A by its posterior expected overlap with the latent top-K set,

$$R_D(A) = \mathbb{E}[|A \cap T_K(q)| \mid D].$$

By linearity of expectation,

$$R_D(A) = \sum_{i \in A} p_i.$$

Hence selecting the K largest membership probabilities maximizes expected overlap,

$$A_p \in \arg\max_{|A|=K} R_D(A).$$

Our central comparison is therefore

$$\Delta_D = R_D(A_p) - R_D(A_{\text{mean}}) \geq 0,$$

which we call the *model-conditional expected-overlap difference*. Thus, Δ_D measures how much expected overlap is lost by ranking posterior means rather than posterior top-K membership probabilities under the same fitted posterior.

Unless stated otherwise, all finite-sample results condition on D. The two ranking rules use the same score-tie convention; when latent qualities have atoms, iid random priorities are used to define $T_K(q)$ without breaking permutation symmetry.

4 Theory: Equivalence, Disagreement and Magnitude

We first establish structural equivalence, then characterize boundary disagreement, and finally quantify its expected-overlap consequence through an exact mass–depth identity and its population limit.

4.1 Equivalence under Residual Symmetry

We first ask when posterior uncertainty preserves the posterior-mean ranking. Exchangeable centered residuals provide a sufficient condition, while still allowing dependence.

Theorem 1 (A structural equivalence regime)**.** *Let $q = m + \varepsilon$, where $\mathbb{E}[\varepsilon_i \mid D] = 0$ and the conditional joint law of ε is exchangeable. Then*

$$m_i \geq m_j \implies p_i \geq p_j, \qquad m_i = m_j \implies p_i = p_j.$$

Every posterior mean top-K roster is expected-overlap optimal, so $\Delta_D = 0$. With $P = I_n - n^{-1}\mathbf{1}\mathbf{1}^\top$, suppose further that $P\varepsilon$ has full support on $\mathbf{1}^\perp$: every nonempty relatively open set has positive probability. Then $m_i > m_j$ implies $p_i > p_j$, and a common score-tie rule gives $A_p = A_{\text{mean}}$.

Theorem 1 indicates that exchangeability is sufficient for posterior-mean selection to attain the optimal expected overlap, while full contrast support further guarantees roster equality; see Appendix A.1 for the coupling argument.

Corollary 1 (Exchangeable Gaussian covariance)**.** *Let $q \mid D \sim N(m, \sigma^2\{(1-\rho)I_n + \rho\mathbf{1}\mathbf{1}^\top\})$, with $\sigma^2 > 0$. Strict order holds for $-1/(n-1) \leq \rho < 1$, including the singular negative endpoint. At $\rho = 1$, all residuals are a common shift and only weak order is guaranteed.*

Since the contrast covariance is $\sigma^2(1-\rho)P$, changing ρ can alter membership confidence without changing the selected roster. We next examine how this equivalence can break.

4.2 Boundary Mechanism

To see how heterogeneous uncertainty can break this equivalence, let $n \geq 2$, $1 \leq K < n$, and suppose the q_i are conditionally independent given D, with continuous densities f_i and cdfs F_i. No identical distributions or common location–scale form are required. Define

$$\pi_i(t) = 1 - F_i(t), \qquad G_n(t) = \frac{1}{n}\sum_i \pi_i(t), \qquad f_n(t) = \frac{1}{n}\sum_i f_i(t) = -G_n'(t).$$

Choose t_n with $G_n(t_n) = K/n$. Continuity ensures existence; the local positive-density condition below and monotonicity ensure uniqueness (Appendix A.2).

Let T_{-i} be the Kth largest quality among the other $n-1$ candidates. Independence and continuity give

$$p_i = \mathbb{E}[\pi_i(T_{-i}) \mid D].$$

The boundary approximation uses $v_i = \pi_i(t_n)$ and $A_v = \mathrm{TopK}(v)$.

Theorem 2 (Coordinate-specific boundary approximation). *Suppose $f_n(t) \geq c > 0$ for $|t - t_n| \leq r$, where $r > 0$, and choose finite $L_i \geq \sup_{|t-t_n|\leq r} f_i(t)$. For every i and $0 < h \leq r$,*

$$\mathbb{P}(|T_{-i} - t_n| > h \mid D) \leq 2\exp\{-2n^2c^2h^2/(n-1)\}. \tag{1}$$

The conditional membership probabilities satisfy the deterministic bounds

$$|p_i - \pi_i(t_n)| \leq e_i := \min\left\{1, \frac{\sqrt{\pi(n-1)}}{\sqrt{2}\,nc} L_i + 2\exp\left(-\frac{2n^2c^2r^2}{n-1}\right)\right\}. \tag{2}$$

Corollary 2 (Gaussian boundary approximation). *Under the local regularity conditions of Theorem 2, suppose $q_i \mid D \sim N(m_i, s_i^2)$ independently, with $s_i > 0$. The Gaussian tail and a valid density bound are*

$$\pi_i(t_n) = \Phi((m_i - t_n)/s_i), \qquad L_i = \frac{1}{s_i\sqrt{2\pi}}$$

so

$$|p_i - \Phi((m_i - t_n)/s_i)| \leq \min\left\{1, \frac{\sqrt{n-1}}{2ncs_i} + 2\exp\left(-\frac{2n^2c^2r^2}{n-1}\right)\right\}. \tag{3}$$

The roster A_v equivalently ranks $(m_i - t_n)/s_i$, using the same score-tie convention.

Here Φ and φ denote the standard normal cdf and density. Proofs of Theorem 2 and Corollary 2 are given in Appendix A.2.

With equal posterior scales, the approximation preserves posterior-mean ordering. Heterogeneous scales can alter this ordering because the effect of uncertainty depends on the mean's position relative to the quality boundary. For fixed m and t,

$$\frac{\partial}{\partial s}\Phi\left(\frac{m-t}{s}\right) = -\frac{m-t}{s^2}\varphi\left(\frac{m-t}{s}\right).$$

Thus, greater uncertainty lowers the probability of exceeding t when $m > t$ and raises it when $m < t$. If $m_i > m_j$ but $(m_i - t_n)/s_i < (m_j - t_n)/s_j$, the approximation ranks j above i despite its lower posterior mean.

Such a reversal concerns the approximate scores, not necessarily the exact membership probabilities. Decision-transfer bounds and sufficient conditions for exact pairwise ordering and roster recovery are given in Appendix A.3. The analytic bounds can be conservative (Appendix C.5).

4.3 Exact Finite and General Population Magnitude

Having characterized where the rankings can disagree, we now quantify the expected-overlap consequence of that disagreement. For a probability cutoff $\tau \in [p_{(K+1)}, p_{(K)}]$, define

$$\alpha_n = K/n, \quad \delta_n = |A_p \triangle A_{\mathrm{mean}}|/n, \quad \bar{d}_n = \frac{\sum_{i \in A_p \triangle A_{\mathrm{mean}}} |p_i - \tau|}{|A_p \triangle A_{\mathrm{mean}}|},$$

with $\bar{d}_n = 0$ when the difference is empty. Here δ_n measures how much the two rosters disagree, while $\bar{d}_n$ measures how far the disputed candidates lie from the probability cutoff; the symmetric difference $A_p \triangle A_{\mathrm{mean}} = (A_p \setminus A_{\mathrm{mean}}) \cup (A_{\mathrm{mean}} \setminus A_p)$ counts both entering and leaving candidates.

Theorem 3 (Exact finite magnitude). *For every joint posterior and every size-K comparator B,*

$$R_D(A_p) - R_D(B) = \sum_{i \in A_p \triangle B} |p_i - \tau|. \tag{4}$$

In particular, including score ties and empty disagreement,

$$\frac{\Delta_D}{K} = \frac{\delta_n \bar{d}_n}{\alpha_n}.$$

Theorem 3 requires no Gaussianity, independence, population model, or limit. Its direct implication is that roster disagreement alone does not determine decision-theoretic consequence: many shallow exchanges may matter less than fewer deep ones.

We next ask whether this finite decomposition stabilizes as the candidate pool grows. Let $(M_i, U_i)_{i\geq 1}$ be one iid sequence with law H on $\mathbb{R} \times [0,1]$, and let the deterministic capacities satisfy $K_n/n \to \alpha \in (0,1)$. The baseline roster ranks M_i, the posterior mean in our application; U_i is the limiting membership score. Assume U_i uniformly approximates the finite membership probabilities $p_{i,n}$:

$$a_n := \max_{i\leq n} |p_{i,n} - U_i| \longrightarrow 0 \quad \text{almost surely,} \tag{5}$$

and there are thresholds b and λ satisfying

$$H(M > b) = H(U > \lambda) = \alpha, \qquad H(M = b) = H(U = \lambda) = 0. \tag{6}$$

Write $B_m = \{M > b\}$, $B_p = \{U > \lambda\}$ and $\delta = H(B_p \triangle B_m)$.

Theorem 4 (Population magnitude and stabilization)**.** (a) Magnitude and mass. *Under* (5)–(6)*, almost surely,*

$$\frac{\Delta_D}{K_n} \longrightarrow g(\alpha, H) := \frac{\mathbb{E}_H[U\mathbf{1}_{B_p}] - \mathbb{E}_H[U\mathbf{1}_{B_m}]}{\alpha} = \frac{\mathbb{E}_H[|U-\lambda|\mathbf{1}_{B_p \triangle B_m}]}{\alpha}, \tag{7}$$

$$\delta_n \longrightarrow \delta. \tag{8}$$

Here $0 \leq g \leq 1$*, and* $g = 0$ *exactly when* $\delta = 0$*.*

(b) Depth. *If* $\delta > 0$*, for every admissible finite-sample cutoff* τ_n *in its admissible interval,*

$$\bar{d}_n \longrightarrow \bar{d} := \mathbb{E}_H[|U-\lambda| \mid B_p \triangle B_m], \qquad g = \delta\bar{d}/\alpha.$$

If $\delta = 0$*, set* $\bar{d} = 0$*; then* $\delta_n \bar{d}_n \to 0$*, without a claim that* $\bar{d}_n \to 0$*.*

Theorem 4 gives the mass–depth decomposition a stable population counterpart. As the pool grows at a fixed selection fraction, persistent disagreement yields a positive limiting gain, determined by both the number and depth of exchanges. Thus, many borderline exchanges may matter less than fewer deep ones. If disagreement vanishes, so does the gain per selected position.

In the independent Gaussian setting of Section 4.2, Corollary 3 in Appendix B.2 verifies the uniform approximation condition (5) under its stated assumptions. Corollary 4 gives limiting equivalence at the symmetric median, whereas Appendix B.4 provides a complementary Gaussian example with heterogeneous scales and strictly positive limiting normalized gain for $0 < \alpha < 1/2$.

5 Simulation Study

We next use simulations to examine the theoretical predictions of Section 4 in the same order. Posterior means are available directly from the simulated models, while posterior Top-K membership probabilities are approximated from posterior draws. Reported finite-sample overlap differences use full-draw plug-in membership estimates; a separate split-draw calculation provides an additional numerical check. Full designs and numerical details are given in Appendix C.

A: structural equivalence. Across 24 exchangeable Gaussian configurations spanning independent, positively correlated, and singular negatively correlated residuals, the two rankings select identical rosters, with no resolved expected-overlap difference. Apparent probability-order inversions vanish under refined Monte Carlo estimation, as predicted by Theorem 1.

B: boundary approximation and roster disagreement. Across 18 independent Gaussian arrays, the maximum error between $v_i = \Phi((m_i - t_n)/s_i)$ and $\hat{p}_i$ ranges from 2.2×10^{-4} to 1.7×10^{-2} (Figure 2(a)). The approximate roster A_v matches $\widehat{A}_p$ in all 18 arrays, whereas A_{mean} differs in 7, with at most 16 replacements (Table 1). Entering candidates have lower posterior means but higher standardized offsets $(m_i - t_n)/s_i$, as in Corollary 2. Figure 3(a) illustrates this mechanism using a saved Experiment C array. Empirical agreement is not analytic certification; Appendix C.5 reports the conservative bounds.

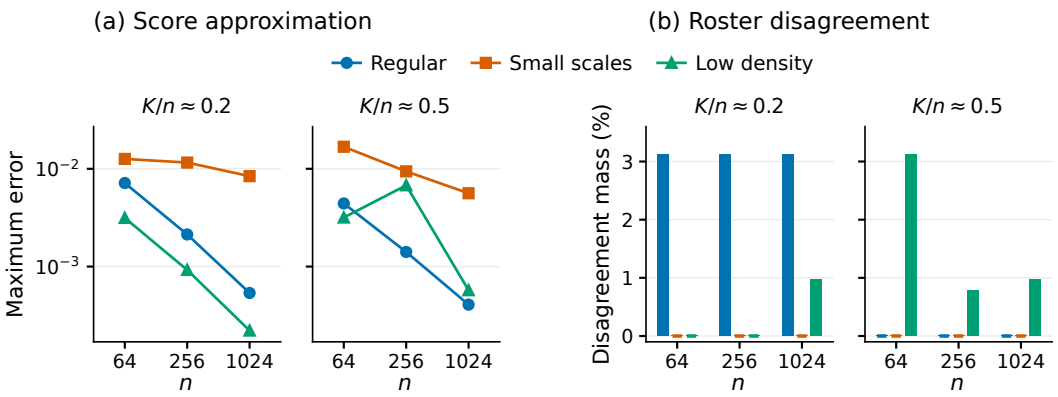


Figure 2: (a) Maximum score approximation error $\max_i |v_i - \widehat{p}_i|$. (b) Symmetric disagreement mass $|A_{\text{mean}} \triangle \widehat{A}_p|/n$, shown as percentages; baseline marks indicate zero. Columns distinguish capacity and colors distinguish setting across all 18 saved arrays.

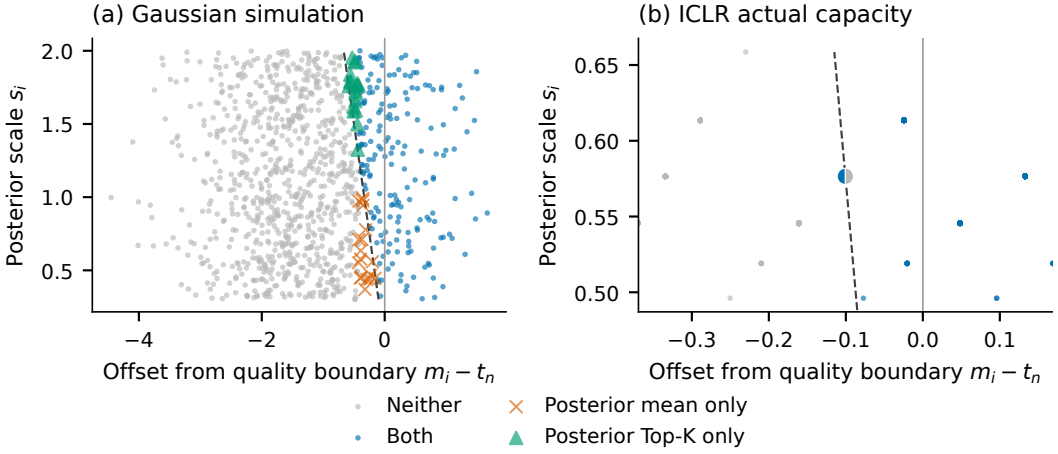


Figure 3: Selection status and standardized boundary geometry. Dashed lines mark approximate membership cutoffs. (a) Experiment C, separated-exchange sequence 0: independent $m_i \sim N(0,1)$, $s_i \sim \text{Unif}[0.3, 2]$, $n = 1024$, $K = 205$; colors use 65,536-draw Rao–Blackwell estimates. (b) ICLR scored cohort, $n = 19474$, $K = 5353$: rosters agree; the split marker denotes a partially selected identical-parameter group under the shared tie rule.

C: disagreement magnitude and stabilization. Population calculations show that many shallow exchanges can yield a smaller g than fewer, deeper exchanges, as predicted by $g = \delta\bar{d}/\alpha$ (Figure 4(a)). Across four regimes and nested sizes $n = 64, 256, 1024$, Monte Carlo estimates of finite normalized gains illustrate stabilization toward population targets (Figure 4(b)); Figure 6 separates mass and depth. At the symmetric median, finite rosters may differ even though limiting mass and gain vanish.

6 Real-Data Analysis

We next examine how the mechanisms studied theoretically and in simulation manifest in two complementary real-data posteriors. For ICLR 2026, we fit an independent Gaussian empirical-Bayes model, yielding paper-level posteriors $q_i \mid D \sim N(m_i, s_i^2)$, and study whether posterior-scale heterogeneity is sufficient to induce boundary reordering at the realized capacity within the scored cohort. For Jester, a Gaussian item-plus-user model, $y_{ij} = q_i + b_j + e_{ij}$, retains the induced posterior dependence across items and allows us to relate model-conditional disagreement magnitude to an

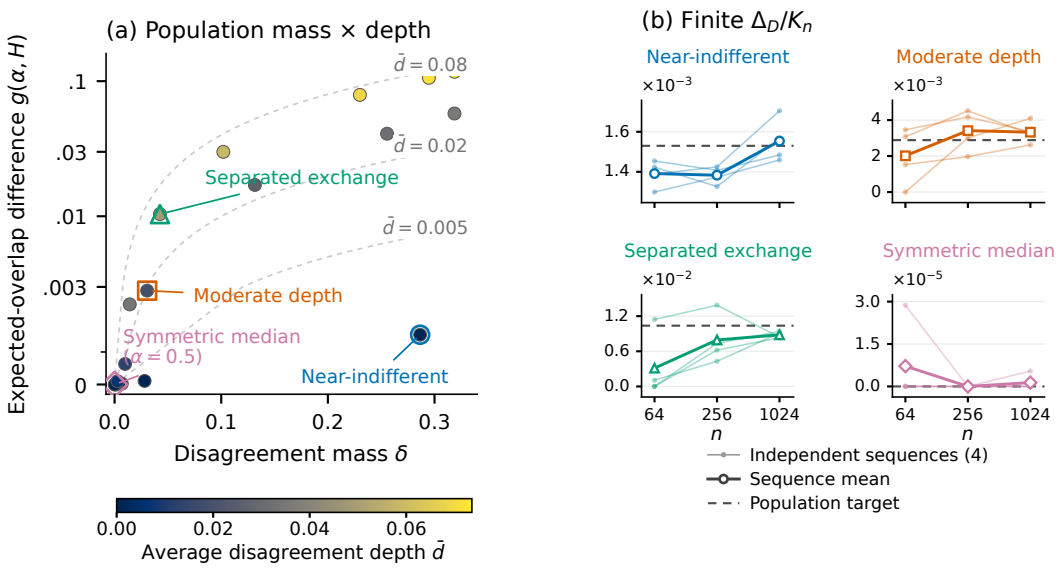


Figure 4: Disagreement magnitude and stabilization. (a) Population relation among disagreement mass δ, average depth $\bar{d}$, and normalized overlap difference g; color indicates $\bar{d}$, and gray lines show fixed depths at reference capacity $\alpha_{\text{ref}} = 0.2$. (b) Full-draw Monte Carlo estimates of finite Δ_D/K_n across four representative regimes; light paths show independent parameter sequences, open markers their means, and dashed lines the population targets. The finite sequences illustrate stabilization, not a convergence rate.

independent-reference comparison. Both analyses condition on fitted hyperparameters; Appendix C gives implementation details and Appendix C.6 discusses practical implications.

6.1 ICLR 2026: Near-Equivalence under Weak Uncertainty Heterogeneity

The public scored snapshot (JerMa88, n.d.) contains $n = 19{,}474$ papers and $75,857$ deduplicated ratings. Matching this analysis cohort to the final ICLR 2026 program gives $K = 5{,}353$ and $\alpha_{\text{actual}} \approx 0.2749$ (ICLR, 2026). Acceptance information is used only to determine the realized capacity.

Posterior scales are nearly homogeneous across papers, with a coefficient of variation of 0.0297. This places the fitted posterior close to the homogeneous-uncertainty regime associated with structural equivalence. The boundary approximation of Corollary 2 provides the corresponding mechanism: when s_i varies little, ranking by $(m_i - t_n)/s_i$ remains close to ranking by m_i, leaving little opportunity for uncertainty-driven boundary reversals.

At the realized capacity, the posterior-mean ranking, the standardized-boundary approximation, and the Monte Carlo estimate of the posterior Top-K ranking select exactly the same roster (Figure 3(b)). Consequently, the estimated disagreement mass and normalized expected-overlap difference are both zero. Together with the weak scale heterogeneity, this provides a real-data example of near-equivalence under the fitted posterior. Appendix C.3 reports capacity-sensitivity and numerical diagnostics.

6.2 Jester: Shallow Disagreement and Independent-Reference Evaluation

The Jester data allow us to separate what the fitted posterior predicts from what independent, higher-information ratings support. We use 100 Jester items (SeppeV, n.d.) and a fixed random split of complete-case users into disjoint training and reference pools. Sparse training panels provide either 2/6 or 4/12 ratings per item, and we compare two ranking rules at four capacities. Reference users are excluded from model fitting and used only for independent-reference comparison. Appendix C.4 gives full implementation details and numerical diagnostics.

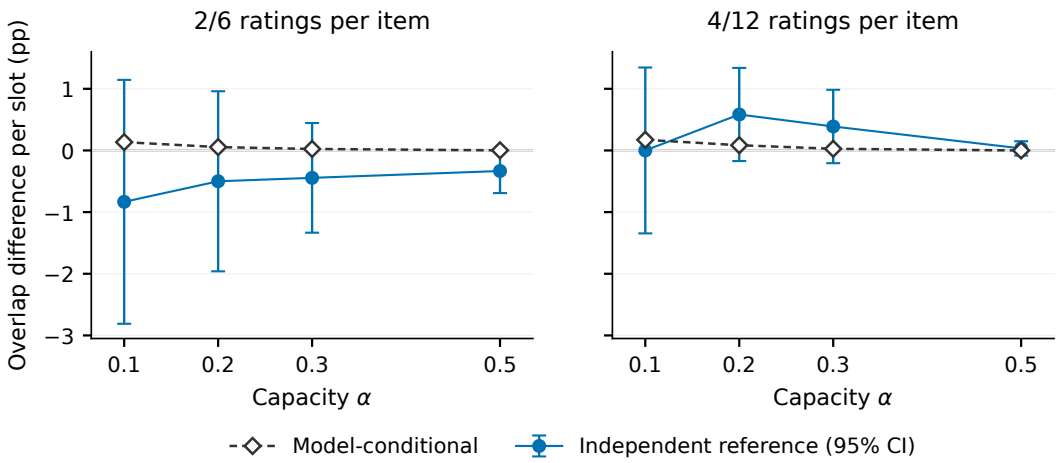


Figure 5: Jester model-conditional and independent-reference overlap differences. Open diamonds and dashed lines show mean full-draw plug-in estimates of the model-conditional differences; filled circles and error bars show mean independent-reference differences with paired-panel 95% intervals.

Full-draw plug-in estimates of posterior Top-K gains are positive but small across all eight design–capacity settings (Figure 5; Table 3). Theorem 3 explains why: $\Delta_D/K = \delta_n \bar{d}_n/\alpha_n$. Although the two rules disagree on a nonzero fraction of items, the disputed items lie close to the membership-probability cutoff, so the estimated model-conditional gain remains small. The split-draw check does not resolve a positive gain for the 2/6 design at $K = 50$ (Appendix C.4). Under the independent reference, all eight paired-panel intervals include zero. Thus conditional optimality does not by itself imply an externally detectable advantage, which additionally depends on model adequacy and reference uncertainty.

7 Limitations

Our results have three main limitations. First, our comparison is model-conditional: the membership probabilities p_i and the resulting rosters depend on the fitted posterior. Posterior misspecification may therefore affect both quantities, so model-conditional gains do not by themselves imply improvements under the true data-generating process. Second, Theorems 1 and 2 provide sufficient regimes rather than a complete characterization of agreement and disagreement. Residual symmetry explains one equivalence regime, while the boundary approximation requires independent posteriors with continuous densities and local regularity. Posterior dependence and violations of local density regularity remain outside the scope of the boundary approximation. Third, the empirical evaluation is limited to the datasets and fitted models studied. The Jester comparison uses a high-information reference rather than exact ground truth, and its confidence intervals condition on fixed user pools without accounting for reference-user sampling uncertainty. Broader independent evaluations are needed to assess the practical value of roster differences across settings.

8 Conclusion

Posterior Top-K and posterior-mean ranking can coincide, disagree only near the selection boundary, or differ in a decision-relevant way under the same fitted posterior. We show that residual symmetry yields an equivalence regime, while heterogeneous uncertainty can induce boundary reorderings, and that the consequence of disagreement is determined not by roster changes alone but by their mass and depth. This finite decomposition also admits a stable population counterpart. Simulations support these mechanisms, while the real-data analyses illustrate both near-equivalence and small, shallow disagreement whose external advantage is not resolved. Overall, roster disagreement alone does not determine the value of Posterior Top-K relative to posterior-mean ranking. Its decision consequence depends jointly on disagreement mass and depth. Conclusions beyond the fitted posterior require independent evidence.

AI USE STATEMENT

Generative AI tools were used as auxiliary support for proof exposition and checking, implementation and debugging of experimental and plotting code, manuscript drafting, limited literature search and bibliographic verification. The authors independently determined the research questions, modeling assumptions, theorem statements, experimental design, and scientific conclusions, and manually verified all AI-assisted mathematical arguments, code, citations, and reported results against the underlying derivations, data, and saved outputs. No AI-generated result was accepted without author verification. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Appendices A and B provide detailed assumptions and proofs of the theoretical results, while Appendix C documents the experimental designs and numerical diagnostics. The supplementary material includes anonymized code for all reported experiments, together with configuration files, random seeds, environment specifications, data-preparation instructions, and scripts for regenerating the empirical figures and tables. It includes a minimal numerical ICLR extract and saved experimental outputs; Jester ratings must be acquired separately from the specified pinned source. The accompanying README describes data access, preprocessing, posterior fitting, and the commands and verification procedures for reproducing the results.

REFERENCES

James O. Berger and John Deely. A Bayesian approach to ranking and selection of related means with alternatives to analysis-of-variance methodology. *Journal of the American Statistical Association*, 83(402):364–373, 1988. doi: 10.1080/01621459.1988.10478606.

Dillon Bowen. Bayesian ranking and selection with applications to field studies, economic mobility, and forecasting, 2022. URL https://arxiv.org/abs/2208.02038.

David J. Eckman and Shane G. Henderson. Posterior-based stopping rules for Bayesian ranking-and-selection procedures. *INFORMS Journal on Computing*, 34(3):1711–1728, 2022. doi: 10.1287/ijoc.2021.1132.

John Ferguson and Joseph Chang. An empirical Bayesian ranking method, with applications to high throughput biology. *Bioinformatics*, 36(1):177–185, 2020. doi: 10.1093/bioinformatics/btz471.

Peter I. Frazier and Warren B. Powell. Paradoxes in learning and the marginal value of information. *Decision Analysis*, 7(4):378–403, 2010. doi: 10.1287/deca.1100.0190.

Harvey Goldstein and David J. Spiegelhalter. League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 159(3):385–409, 1996. doi: 10.2307/2983325.

Jiaying Gu and Roger Koenker. Invidious comparisons: Ranking and selection as compound decisions. *Econometrica*, 91(1):1–41, 2023. doi: 10.3982/ECTA19304.

Peter Hall and Hugh Miller. Using the bootstrap to quantify the authority of an empirical ranking. *The Annals of Statistics*, 37(6B):3929–3959, 2009. doi: 10.1214/09-AOS699.

Nicholas C. Henderson and Michael A. Newton. Making the cut: Improved ranking and selection for large-scale inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):781–804, 2016. doi: 10.1111/rssb.12131. URL https://pages.stat.wisc.edu/~newton/papers/publications/rvalues.pdf.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

ICLR. ICLR 2026 conference program, 2026. URL https://iclr.cc/virtual/2026/papers.html.

JerMa88. ICLR peer reviews 2026. Hugging Face dataset, n.d. URL https://huggingface.co/datasets/JerMa88/ICLR_Peer_Reviews_2026.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 655–662. Omnipress, 2012. URL https://www.cs.utexas.edu/~pstone/Papers/bib2html-links/ICML12-shivaram.pdf.

Nan M. Laird and Thomas A. Louis. Empirical Bayes ranking methods. *Journal of Educational Statistics*, 14(1):29–46, 1989. doi: 10.3102/10769986014001029.

Rongheng Lin, Thomas A. Louis, Susan M. Paddock, and Greg Ridgeway. Loss function based ranking in two-stage, hierarchical models. *Bayesian Analysis*, 1(4):915–946, 2006. doi: 10.1214/06-BA130. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2896056/.

Haitao Liu, Zheng Xing, Hui Xiao, and Ek Peng Chew. Marginal improvement procedures for top-m selection. *Automatica*, 170:111875, 2024. doi: 10.1016/j.automatica.2024.111875.

Quoc Phong Nguyen, Sebastian Tay, Bryan Kian Hsiang Low, and Patrick Jaillet. Top-k ranking Bayesian optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9135–9143, 2021. doi: 10.1609/aaai.v35i10.17103.

Hisashi Noma, Shigeyuki Matsui, Takashi Omori, and Tosiya Sato. Bayesian ranking and selection methods using hierarchical mixture models in microarray studies. *Biostatistics*, 11(2):281–289, 2010. doi: 10.1093/biostatistics/kxp047.

SeppeV. Rated jokes dataset from Jester. Hugging Face dataset, n.d. URL https://huggingface.co/datasets/SeppeV/rated_jokes_dataset_from_jester.

Wei Shen and Thomas A. Louis. Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):455–471, 1998. doi: 10.1111/1467-9868.00135.

Hui Xiao, Yao Zhang, Gang Kou, Si Zhang, and Juergen Branke. Ranking and selection for pairwise comparison. *Naval Research Logistics*, 70(3):284–302, 2023. doi: 10.1002/nav.22093.

Jian Xu, Delu Zeng, John Paisley, and Qibin Zhao. When can you debias an LLM judge? identifiability limits, a test, and designs for top-k ranking, 2026. URL https://arxiv.org/abs/2607.02104v2. Version 2, July 18, 2026.

Gongbo Zhang, Yijie Peng, Jianghua Zhang, and Enlu Zhou. Asymptotically optimal sampling policy for selecting top-m alternatives. *INFORMS Journal on Computing*, 35(6):1261–1285, 2023. doi: 10.1287/ijoc.2021.0333.

Or Zuk, Liat Ein-Dor, and Eytan Domany. Ranking under uncertainty. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 466–473. AUAI Press, 2007. URL https://arxiv.org/abs/1206.5280.

A PROOFS OF THE FOUR MAIN RESULTS

A.1 EXCHANGEABLE RESIDUALS AND STRICTNESS

Proof of Theorem 1. All expectations are conditional on D. For vectors $x, z \in \mathbb{R}^n$, let $I_k(x, z)$ indicate whether candidate k belongs to the top-K set when candidates are ranked by x, with z used to break ties. Take iid uniform priorities $Z_1, \dots, Z_n$, independent of ε.

Weak ordering. Fix $i \neq j$, put $d = m_i - m_j \geq 0$, and define

$$m^{(0)} = m - de_i, \qquad X = m^{(0)} + \varepsilon.$$

Here e_i is the ith coordinate basis vector. Since $m_i^{(0)} = m_j^{(0)}$, exchangeability of the residuals and priorities gives

$$(X, Z) \overset{d}{=} (\mathsf{S}_{ij}X, \mathsf{S}_{ij}Z), \qquad \mathbb{E}I_i(X, Z) = \mathbb{E}I_j(X, Z),$$

where S_{ij} swaps coordinates i, j. Raising only x_i cannot increase the number of competitors above i and cannot decrease the number above j. Hence, almost surely in Z,

$$I_i(x + de_i, Z) - I_i(x, Z) \geq 0, \qquad I_j(x, Z) - I_j(x + de_i, Z) \geq 0.$$

Using $q = X + de_i$ and the above inequalities, we have

$$\begin{aligned} p_i - p_j &= \mathbb{E}I_i(X + de_i, Z) - \mathbb{E}I_j(X + de_i, Z) \\ &= \mathbb{E}\big[I_i(X + de_i, Z) - I_i(X, Z)\big] + \mathbb{E}\big[I_j(X, Z) - I_j(X + de_i, Z)\big] \geq 0. \end{aligned} \tag{9}$$

When $d = 0$, both increments vanish, so $p_i = p_j$. For $d > 0$, each increment is either zero or one. Strict inequality therefore holds exactly when the coupling makes i enter or j leave with positive probability.

Strict ordering under full contrast support. Suppose $d > 0$. Choose a set $H \subseteq \{1, \dots, n\} \setminus \{i, j\}$ of size $K - 1$ and define

$$x_i = 0, \quad x_j = d/2, \quad x_k = 2d \ \text{ for } \ k \in H, \quad x_k = -2d \ \text{ for } \ k \notin H \cup \{i, j\}.$$

At x, candidate j occupies the last selected position; at $x + de_i$, candidate i occupies it. Every comparison needed for these statements has a gap of at least $d/2$. For $\|w\|_\infty < d/8$,

$$|w_k - w_\ell| \leq 2\|w\|_\infty < d/4 < d/2,$$

so the same memberships hold at $x + w$ and $x + de_i + w$, independently of priorities. Put

$$v = P(x - m^{(0)}), \qquad \mathcal{E} = \{\|P\varepsilon - v\|_\infty < d/8\}.$$

The defining ball is a nonempty relatively open subset of $\mathbf{1}^\perp$, so full contrast support gives $\mathbb{P}(\mathcal{E} \mid D) > 0$. On $\mathcal{E}$, write $w = P\varepsilon - v$. Since $P(X - x - w) = 0$ and $\ker(P) = \operatorname{span}\{\mathbf{1}\}$, there exists a scalar c such that

$$X = x + w + c\mathbf{1}.$$

Common shifts preserve all rankings. Consequently,

$$\mathcal{E} \subseteq \{I_i(X + de_i, Z) - I_i(X, Z) = 1,\ I_j(X, Z) - I_j(X + de_i, Z) = 1\},$$

and (9) yields $p_i - p_j \geq 2\mathbb{P}(\mathcal{E} \mid D) > 0$.

Optimality and roster equality. If A is any posterior-mean top-K roster, then

$$i \in A,\ j \notin A \quad \Longrightarrow \quad m_i \geq m_j \quad \Longrightarrow \quad p_i \geq p_j.$$

Hence every selected index has a membership probability at least as large as that of any omitted index, so A is itself a top-K set of the membership probabilities. Thus $\sum_{i \in A} p_i = \max_{|B|=K} \sum_{i \in B} p_i$, proving expected-overlap optimality and $\Delta_D = 0$. Under full contrast support, unequal means give the same strict ordering of probabilities, and equal means give equal probabilities. A common score-tie rule therefore gives $A_p = A_{\text{mean}}$. □

Proof of Corollary 1. For $\Sigma = \sigma^2\{(1-\rho)I_n + \rho\mathbf{1}\mathbf{1}^\top\}$,

$$\Sigma\mathbf{1} = \sigma^2\{1+(n-1)\rho\}\mathbf{1}, \qquad \Sigma v = \sigma^2(1-\rho)v \quad (v \in \mathbf{1}^\perp).$$

Thus $\Sigma \succeq 0$ exactly when $-1/(n-1) \le \rho \le 1$. Every permutation matrix S satisfies $S\Sigma S^\top = \Sigma$, so the centered Gaussian law is exchangeable, including at singular endpoints. On the contrast space,

$$\mathrm{Cov}(P\varepsilon) = P\Sigma P = \sigma^2(1-\rho)P.$$

For $\rho < 1$, this covariance is positive definite in orthonormal coordinates on $\mathbf{1}^\perp$. The resulting Gaussian density is positive everywhere on that space, verifying full contrast support even at $\rho = -1/(n-1)$. Theorem 1 gives strict ordering. At $\rho = 1$,

$$\mathrm{Var}(\varepsilon_i - \varepsilon_j) = 2\sigma^2(1-\rho) = 0,$$

so all residuals equal a common shift, which preserves each realized ranking. □

A.2 Conditional boundary approximation

Proof of Theorem 2. Condition throughout on D. The continuous function G_n has limits 1 and 0 at $-\infty$ and $+\infty$, respectively. Since $0 < K/n < 1$, a solution of $G_n(t_n) = K/n$ exists. The local lower bound $f_n \ge c > 0$ makes G_n strictly decreasing on $[t_n - r, t_n + r]$; monotonicity outside this interval then rules out any other solution. Each continuous density is bounded on this compact interval, so finite local bounds L_i exist.

Continuity and conditional independence imply that quality ties have probability zero. The variable q_i is conditionally independent of T_{-i}, so

$$p_i = \mathbb{P}(q_i > T_{-i} \mid D) = \mathbb{E}[\mathbb{P}(q_i > T_{-i} \mid T_{-i}, D) \mid D] = \mathbb{E}[\pi_i(T_{-i}) \mid D].$$

No moment condition is needed for this identity or the bounds below. Define

$$N_{-i}(t) = \sum_{j \ne i} \mathbf{1}\{q_j > t\}, \qquad \mu_{-i}(t) = \mathbb{E}[N_{-i}(t) \mid D] = \sum_{j \ne i} \pi_j(t) = nG_n(t) - \pi_i(t).$$

Continuity of the coordinates gives, up to null events,

$$\{T_{-i} > t\} = \{N_{-i}(t) \ge K\}, \qquad \{T_{-i} < t\} = \{N_{-i}(t) \le K-1\}.$$

Cutoff concentration. For $0 < h \le r$, integrate $f_n = -G_n' \ge c$:

$$G_n(t_n + h) = G_n(t_n) - \int_{t_n}^{t_n+h} f_n(u)\,du \le K/n - ch,$$
$$G_n(t_n - h) = G_n(t_n) + \int_{t_n-h}^{t_n} f_n(u)\,du \ge K/n + ch.$$

Since $0 \le \pi_i \le 1$,

$$\mu_{-i}(t_n + h) \le K - nch - \pi_i(t_n + h) \le K - nch,$$
$$\mu_{-i}(t_n - h) \ge K + nch - \pi_i(t_n - h) \ge K + nch - 1.$$

The adverse count deviations therefore satisfy

$$K - \mu_{-i}(t_n + h) \ge K - (K - nch) = nch,$$
$$\mu_{-i}(t_n - h) - (K-1) \ge (K + nch - 1) - (K-1) = nch.$$

Each count is a sum of $n-1$ conditionally independent Bernoulli variables. Hoeffding's inequality (Hoeffding, 1963) gives

$$\mathbb{P}(T_{-i} > t_n + h \mid D) \le \mathbb{P}\{N_{-i}(t_n + h) - \mu_{-i}(t_n + h) \ge nch \mid D\} \le e^{-2n^2c^2h^2/(n-1)},$$
$$\mathbb{P}(T_{-i} < t_n - h \mid D) \le \mathbb{P}\{\mu_{-i}(t_n - h) - N_{-i}(t_n - h) \ge nch \mid D\} \le e^{-2n^2c^2h^2/(n-1)}.$$

Adding the tails proves (1). The constant follows directly from the centered moment-generating-function bound: for a sum N of N_0 independent variables with range length one,

$$\mathbb{P}(N - \mathbb{E}N \ge a) \le \inf_{\theta > 0} \exp\{-\theta a + N_0\theta^2/8\} = \exp(-2a^2/N_0).$$

Membership approximation. Set $X = |T_{-i} - t_n|$ and $A = 2n^2c^2/(n-1)$. On $X \le r$, the local density bound gives $|\pi_i(T_{-i}) - \pi_i(t_n)| \le L_i X$; on $X > r$, the difference is at most one. Thus the locally valid inequality is

$$|\pi_i(T_{-i}) - \pi_i(t_n)| \le L_i \min(X, r) + \mathbf{1}\{X > r\}.$$

Consequently,

$$\begin{aligned}
|p_i - \pi_i(t_n)| &\le \mathbb{E}[|\pi_i(T_{-i}) - \pi_i(t_n)| \mid D] \\
&\le L_i \mathbb{E}[\min(X, r) \mid D] + \mathbb{P}(X > r \mid D) \\
&= L_i \int_0^r \mathbb{P}(X > h \mid D)\, dh + \mathbb{P}(X > r \mid D) \\
&\le 2L_i \int_0^r e^{-Ah^2}\, dh + 2e^{-Ar^2} \\
&\le 2L_i \int_0^\infty e^{-Ah^2}\, dh + 2e^{-Ar^2} = \frac{L_i\sqrt{\pi}}{\sqrt{A}} + 2e^{-Ar^2} \\
&= \frac{\sqrt{\pi(n-1)}}{\sqrt{2}\, nc} L_i + 2e^{-2n^2c^2r^2/(n-1)}.
\end{aligned}$$

The integral identity follows from $\min(X, r) = \int_0^r \mathbf{1}\{X > h\}\, dh$ and Tonelli's theorem. Extending the exponential integral does not assert a cutoff-tail bound outside $[0, r]$. Combining this estimate with $|p_i - \pi_i(t_n)| \le 1$ proves (2). □

Proof of Corollary 2. For $q_i \mid D \sim N(m_i, s_i^2)$,

$$\pi_i(t_n) = \Phi((m_i - t_n)/s_i), \qquad f_i(t) = \frac{1}{s_i}\varphi((t - m_i)/s_i) \le \frac{1}{s_i\sqrt{2\pi}}.$$

Hence $L_i = (s_i\sqrt{2\pi})^{-1}$ is a valid local bound. Substituting into (2) and using

$$\frac{\sqrt{\pi(n-1)}}{\sqrt{2}\, nc} \frac{1}{s_i\sqrt{2\pi}} = \frac{\sqrt{n-1}}{2ncs_i}$$

proves (3). Strict monotonicity of Φ shows that ranking $v_i = \pi_i(t_n)$ is equivalent to ranking $(m_i - t_n)/s_i$, with the same score-tie convention. With this Gaussian choice, define $s_* = \min_i s_i$ and $e_n = \max_i e_i$. Since the Gaussian radius is nonincreasing in s_i, e_n equals the uniform radius obtained by replacing s_i with s_* in the bound of Corollary 2. □

A.3 Nonuniform and local decision transfer

More generally, coordinate-wise approximation errors also translate into decision guarantees.

Proposition 1 (Nonuniform decision transfer)**.** *Suppose* $|v_i - p_i| \le e_i$, $A_v = \mathrm{TopK}(v)$, $E = A_p \setminus A_v$, *and* $F = A_v \setminus A_p$. *Then*

$$0 \le R_D(A_p) - R_D(A_v) \le \sum_{i \in E} e_i + \sum_{j \in F} e_j. \tag{10}$$

Moreover, $v_i - v_j > e_i + e_j$ *implies* $p_i > p_j$. *In particular, the observable condition*

$$\min_{i \in A_v}(v_i - e_i) > \max_{j \notin A_v}(v_j + e_j)$$

certifies that $A_v = A_p$ *is the unique optimal roster. For* $e_i = e$, (10) *recovers* $2r_v e \le 2Ke$, *where* $r_v = |E|$.

Proof of Proposition 1. Let $E = A_p \setminus A_v$ and $F = A_v \setminus A_p$. Both rosters have size K, so $|E| = |F|$. Optimality of A_v for v gives

$$\sum_{i \in E} v_i \le \sum_{j \in F} v_j.$$

Consequently,

$$\begin{aligned} 0 \le R_D(A_p) - R_D(A_v) &= \sum_{i\in E} p_i - \sum_{j\in F} p_j \\ &= \sum_{i\in E}(p_i - v_i) + \left[\sum_{i\in E} v_i - \sum_{j\in F} v_j\right] + \sum_{j\in F}(v_j - p_j) \\ &\le \sum_{i\in E} e_i + \sum_{j\in F} e_j. \end{aligned}$$

For each pair,

$$p_i - p_j \ge (v_i - e_i) - (v_j + e_j) = v_i - v_j - e_i - e_j.$$

Thus strict interval separation gives strict probability order. If it holds for every selected–omitted pair, A_v is the unique optimal roster. For $e_i = e$, the bound is $2|E|e \le 2Ke$. No independence is used. □

Certifying a Gaussian ranking reversal. For the Gaussian scores in Corollary 2, the separation $\pi_j(t_n) - \pi_i(t_n) > e_i + e_j$ implies

$$p_j - p_i \ge \pi_j(t_n) - \pi_i(t_n) - e_i - e_j > 0.$$

Thus, if $m_i > m_j$ but $(m_i - t_n)/s_i < (m_j - t_n)/s_j$, the approximation predicts a reversal, while this strict separation certifies the exact order $p_j > p_i$.

The computable interval bound. The bound in (10) depends on the unknown roster A_p. For a computable alternative, define $\ell_i = \max(0, v_i - e_i)$ and $u_i = \min(1, v_i + e_i)$. Sort omitted upper endpoints decreasingly as $u^{\text{out}}_{[j]}$ and selected lower endpoints increasingly as $\ell^{\text{in}}_{[j]}$. Then

$$R_D(A_p) - R_D(A_v) \le Q(v, e) := \max_{0\le r\le \min(K, n-K)} \sum_{j=1}^{r} (u^{\text{out}}_{[j]} - \ell^{\text{in}}_{[j]}). \tag{11}$$

The empty sum is zero. The quantity Q is the largest regret allowed by the coordinate intervals. It is conservative for membership vectors, which also satisfy $\sum_i p_i = K$.

Optimization interpretation of $Q(v, e)$. Let $\mathcal{B} = \prod_i [\ell_i, u_i]$. For $x \in \mathcal{B}$, define regret relative to A_v by

$$\mathcal{R}(x) = \max_{|B|=K} \left(\sum_{i\in B} x_i - \sum_{i\in A_v} x_i \right).$$

Writing $O = B \setminus A_v$, $I = A_v \setminus B$, and $|O| = |I| = r$ gives

$$\begin{aligned} \max_{x\in\mathcal{B}} \mathcal{R}(x) &= \max_{0\le r\le \min(K,n-K)} \; \max_{\substack{O\subseteq A_v^c,\; I\subseteq A_v \\ |O|=|I|=r}} \left(\sum_{i\in O} u_i - \sum_{j\in I} \ell_j \right) \\ &= \max_{0\le r\le \min(K,n-K)} \sum_{j=1}^{r} (u^{\text{out}}_{[j]} - \ell^{\text{in}}_{[j]}) = Q(v, e). \end{aligned}$$

The endpoint vector $x_i = u_i$ outside A_v, $x_i = \ell_i$ inside A_v attains the maximum. Since $p \in \mathcal{B}$ and A_p maximizes $\sum_{i\in A} p_i$ over $|A| = K$,

$$\mathcal{R}(p) = \sum_{i\in A_p} p_i - \sum_{i\in A_v} p_i = R_D(A_p) - R_D(A_v) \le \max_{x\in\mathcal{B}} \mathcal{R}(x) = Q(v, e),$$

which proves (11). This endpoint vector need not satisfy $\sum_i x_i = K$, so sharpness over the box does not imply sharpness over membership vectors. Sorting costs $O(n \log n)$, and $0 \le Q \le \min(K, n - K)$.

A.4 PROOFS FOR THE PROPORTIONAL-GROWTH COMPARISON

Proof of Theorem 3. For a size-K comparator B, set $C = A_p \setminus B$ and $F = B \setminus A_p$. Then $|C| = |F|$ and, for every $\tau \in [p_{(K+1)}, p_{(K)}]$,

$$p_i \geq \tau \quad (i \in C), \qquad p_j \leq \tau \quad (j \in F).$$

Canceling common members and using $\tau(|C| - |F|) = 0$ gives

$$\begin{aligned} R_D(A_p) - R_D(B) &= \sum_{i \in C} p_i - \sum_{j \in F} p_j \\ &= \sum_{i \in C} (p_i - \tau) + \sum_{j \in F} (\tau - p_j) \\ &= \sum_{i \in A_p \triangle B} |p_i - \tau|. \end{aligned}$$

This proves (4), including score ties. With $B = A_{\text{mean}}$,

$$|A_p \triangle A_{\text{mean}}| = n\delta_n, \qquad \Delta_D = n\delta_n \bar{d}_n, \qquad K = n\alpha_n,$$

and therefore

$$\frac{\Delta_D}{K} = \frac{\delta_n \bar{d}_n}{\alpha_n}.$$

When the symmetric difference is empty, both sides are zero by the convention $\bar{d}_n = 0$.

□

Proof of Theorem 4(a)–(b). Work on the probability-one event where $a_n \to 0$ and the strong laws below hold for the single iid sequence (M_i, U_i). We do not assume independence of the triangular array $p_{i,n}$.

Threshold sets and selected sums. Define

$$C_{p,n} = \{i \leq n : U_i > \lambda\}, \qquad C_{m,n} = \{i \leq n : M_i > b\}, \qquad A_U = \text{TopK}(U).$$

Each threshold set and its corresponding top-K_n roster are nested. Thus

$$|A_U \triangle C_{p,n}| = \big||C_{p,n}| - K_n\big|, \qquad |A_{\text{mean}} \triangle C_{m,n}| = \big||C_{m,n}| - K_n\big|. \tag{12}$$

The strong law gives

$$\begin{aligned} &|C_{p,n}|/n \longrightarrow \alpha, && |C_{m,n}|/n \longrightarrow \alpha, \\ &\frac{1}{n} \sum_{i \in C_{p,n}} U_i \longrightarrow \mu_p := \mathbb{E}_H[U \mathbf{1}_{B_p}], && \frac{1}{n} \sum_{i \in C_{m,n}} U_i \longrightarrow \mu_m := \mathbb{E}_H[U \mathbf{1}_{B_m}]. \end{aligned}$$

Since $0 \leq U_i \leq 1$ and $K_n/n \to \alpha$,

$$\left| \frac{1}{n} \sum_{i \in A_U} U_i - \frac{1}{n} \sum_{i \in C_{p,n}} U_i \right| \leq \left| \frac{|C_{p,n}|}{n} - \alpha_n \right| \longrightarrow 0,$$

$$\left| \frac{1}{n} \sum_{i \in A_{\text{mean}}} U_i - \frac{1}{n} \sum_{i \in C_{m,n}} U_i \right| \leq \left| \frac{|C_{m,n}|}{n} - \alpha_n \right| \longrightarrow 0.$$

Hence the two normalized selected sums converge to μ_p and μ_m.

Uniform approximation and the value limit. For every size-K_n set A,

$$\left| \sum_{i \in A} p_{i,n} - \sum_{i \in A} U_i \right| \leq K_n a_n.$$

Maximizing the two inequalities obtained from this bound gives

$$\max_{|A|=K_n} \sum_{i\in A} U_i - K_n a_n \le \max_{|A|=K_n} \sum_{i\in A} p_{i,n} \le \max_{|A|=K_n} \sum_{i\in A} U_i + K_n a_n.$$

Applying the same bound to A_{mean} yields

$$\left| R_D(A_p) - \sum_{i\in A_U} U_i \right| \le K_n a_n, \qquad \left| R_D(A_{\text{mean}}) - \sum_{i\in A_{\text{mean}}} U_i \right| \le K_n a_n. \tag{13}$$

Consequently,

$$\left| \frac{\Delta_D}{K_n} - \frac{n^{-1}\sum_{i\in A_U} U_i - n^{-1}\sum_{i\in A_{\text{mean}}} U_i}{\alpha_n} \right| \le 2a_n \longrightarrow 0, \qquad \frac{\Delta_D}{K_n} \longrightarrow \frac{\mu_p - \mu_m}{\alpha}.$$

Equal population capacities imply $H(B_p \setminus B_m) = H(B_m \setminus B_p)$. Therefore

$$\begin{aligned} \mu_p - \mu_m &= \mathbb{E}_H[U\mathbf{1}_{B_p\setminus B_m}] - \mathbb{E}_H[U\mathbf{1}_{B_m\setminus B_p}] \\ &= \mathbb{E}_H[(U-\lambda)\mathbf{1}_{B_p\setminus B_m}] + \mathbb{E}_H[(\lambda-U)\mathbf{1}_{B_m\setminus B_p}] \\ &= \mathbb{E}_H[|U-\lambda|\mathbf{1}_{B_p\triangle B_m}]. \end{aligned}$$

This proves both expressions in (7). The integrand is positive on the symmetric difference except on the null set $\{U = \lambda\}$. Hence $g = 0$ exactly when $\delta = 0$. Also $0 \le \mu_p - \mu_m \le \mu_p \le \alpha$, so $0 \le g \le 1$.

Convergence of disagreement mass. Let $X_n = C_{p,n} \setminus A_p$ and $Y_n = A_p \setminus C_{p,n}$. If both sets are nonempty, then every $x \in X_n$, $y \in Y_n$ satisfies

$$p_{y,n} \ge p_{x,n}, \qquad U_x > \lambda \ge U_y, \qquad 0 < U_x - U_y \le (p_{x,n} + a_n) - (p_{y,n} - a_n) \le 2a_n.$$

Hence

$$0 < U_x - \lambda \le U_x - U_y \le 2a_n, \qquad 0 \le \lambda - U_y < U_x - U_y \le 2a_n,$$

so every index in $X_n \cup Y_n$ lies in $\{|U_i - \lambda| \le 2a_n\}$. If one set is empty, the discrepancy is exactly the difference of cardinalities. In either case,

$$\frac{|A_p \triangle C_{p,n}|}{n} \le \left| \frac{|C_{p,n}|}{n} - \alpha_n \right| + \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{|U_i - \lambda| \le 2a_n\}.$$

For every fixed $\eta > 0$, eventually $a_n \le \eta$. The strong law and continuity of probability from above give

$$\limsup_n \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{|U_i - \lambda| \le 2a_n\} \le H(|U-\lambda| \le 2\eta) \longrightarrow H(U = \lambda) = 0 \quad (\eta \downarrow 0).$$

Use a countable sequence of η to stay on one probability-one event. Together with (12), this gives

$$\left| \delta_n - \frac{|C_{p,n} \triangle C_{m,n}|}{n} \right| \le \frac{|A_p \triangle C_{p,n}| + |A_{\text{mean}} \triangle C_{m,n}|}{n} \longrightarrow 0,$$

$$\frac{|C_{p,n} \triangle C_{m,n}|}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{B_p\triangle B_m}(M_i, U_i) \longrightarrow \delta.$$

This proves (8).

Depth and zero limiting mass. If $\delta > 0$, then $\delta_n > 0$ eventually. Theorem 3 implies, for every admissible τ_n,

$$\bar{d}_n = \frac{\alpha_n}{\delta_n}\frac{\Delta_D}{K_n} \longrightarrow \frac{\alpha g}{\delta} = \frac{\mathbb{E}_H[|U-\lambda|\mathbf{1}_{B_p\triangle B_m}]}{H(B_p \triangle B_m)} = \bar{d}, \qquad g = \frac{\delta \bar{d}}{\alpha}.$$

No convergence of τ_n is needed. If $\delta = 0$, set $\bar{d} = 0$. The conclusion is only

$$\delta_n \bar{d}_n = \alpha_n \frac{\Delta_D}{K_n} \longrightarrow \alpha g = 0.$$

Division by $\delta_n \to 0$ is not justified, so no convergence of $\bar{d}_n$ to zero is asserted. □

B ADDITIONAL ASSUMPTIONS, EXAMPLES, AND FINITE BOUNDS

B.1 WEAK ORDERING NEED NOT BE STRICT

Example 1 (Weak ordering need not be strict)**.** Take $n = 3$, $K = 1$, $m = (4, 1, 0)$, and iid $\mathrm{Unif}[-0.1, 0.1]$ residuals. Then

$$p = (1, 0, 0)$$

although $m_2 > m_3$.

The support intervals are strictly separated, so candidate 1 is always the maximum. Thus exchangeability alone need not turn the weak ordering in Theorem 1 into a strict one; the additional contrast-support condition is doing real work.

B.2 GAUSSIAN SPECIALIZATION AND SYMMETRIC-MEDIAN EQUIVALENCE

Corollary 3 (Independent Gaussian specialization)**.** *Let $(M_i, S_i)_{i\geq 1}$ be iid on $\mathbb{R} \times [s_*, s^*]$, with $0 < s_* \leq s^* < \infty$. For each prefix of size n, let $q_i \mid (M_i, S_i) \sim N(M_i, S_i^2)$ independently, and let the deterministic capacities satisfy $K_n/n \to \alpha \in (0, 1)$. Define*

$$G(t) = \mathbb{E}\Phi((M - t)/S), \quad G(t_\alpha) = \alpha, \quad U = \Phi((M - t_\alpha)/S). \tag{14}$$

Then the uniform approximation condition (5) holds. If the threshold conditions in (6) also hold, Theorem 4 applies with the induced law of (M, U).

Proof of Corollary 3. Corollary 5 in Appendix B.3 establishes, under these assumptions,

$$t_n \to t_\alpha, \qquad \max_{i\leq n} |p_{i,n} - U_i| \to 0 \quad \text{almost surely.}$$

Hence condition (5) holds, and Theorem 4 applies whenever (6) is satisfied. □

Gaussianity is used here to verify the uniform approximation required by Theorem 4; the finite mass–depth identity itself does not rely on Gaussianity. Scale heterogeneity can nevertheless coexist with limiting equivalence:

Corollary 4 (Heterogeneous uncertainty at the symmetric median)**.** *Under the setup of Corollary 3, suppose additionally that $(M - \mu, S) \overset{d}{=} (-(M - \mu), S)$ and $H(M = \mu) = 0$. Then at $\alpha = 1/2$, $t_{1/2} = \mu$, $B_p = B_m = \{M > \mu\}$, and $g = 0$. This gives $|A_p \triangle A_{\mathrm{mean}}| = o(n)$ and $\Delta_D = o(K_n)$.*

Proof of Corollary 4. Reflection symmetry and $\Phi(x) + \Phi(-x) = 1$ imply

$$G(\mu) = \mathbb{E}\Phi((M - \mu)/S) = \mathbb{E}\Phi(-(M - \mu)/S) = 1 - G(\mu) = 1/2.$$

Strict monotonicity gives $t_{1/2} = \mu$. Since $S > 0$,

$$U > 1/2 \iff M > \mu, \qquad H(U = 1/2) = H(M = \mu) = 0,$$

and symmetry gives $H(M > \mu) = 1/2$. Thus $B_p = B_m$, $\delta = 0$, and Theorem 4 gives $g = 0$, $|A_p \triangle A_{\mathrm{mean}}| = o(n)$, and $\Delta_D = o(K_n)$. □

B.3 TRANSITION TO A PARAMETER POPULATION

Assumption 1 (Population for proportional growth)**.** The arrays are prefixes of an infinite iid sequence $(M_i, S_i) \sim H$ on $\mathbb{R} \times [s_*, s^*]$, with $0 < s_* \leq s^* < \infty$. Conditional on each prefix the qualities are independent Gaussian. The deterministic capacities satisfy $\alpha_n = K_n/n \to \alpha \in (0, 1)$. Use G, t_α, U as defined in (14).

Let

$$f(t) = -G'(t) = \mathbb{E}_H\left[\frac{\varphi((M - t)/S)}{S}\right].$$

The scale lower bound makes f continuous and strictly positive at every finite t, providing the local density and Lipschitz control used below.

Corollary 5 (Population threshold and uniform approximation)**.** *Under Assumption 1,*

$$t_n \to t_\alpha, \qquad \max_{i\le n} |p_{i,n} - U_i| \to 0$$

almost surely over the parameter sequence, where $U_i = \Phi((M_i - t_\alpha)/S_i)$.

Proof of Corollary 5. Write

$$f_i(t) = \frac{\varphi((M_i - t)/S_i)}{S_i}, \qquad f_n(t) = \frac{1}{n}\sum_{i=1}^{n} f_i(t), \qquad f(t) = \mathbb{E} f_i(t).$$

Since $S_i \ge s_* > 0$, uniformly in i and t,

$$0 < f_i(t) \le b := \frac{1}{s_*\sqrt{2\pi}}, \qquad |f_i'(t)| \le L := \frac{1}{s_*^2\sqrt{2\pi e}}.$$

Differentiation under the expectation and dominated convergence give $-G' = f > 0$, continuity of f, and $G(-\infty) = 1$, $G(\infty) = 0$. Thus t_α is unique.

Threshold convergence. For every fixed $\varepsilon > 0$, strict monotonicity of G gives

$$G(t_\alpha - \varepsilon) > \alpha > G(t_\alpha + \varepsilon).$$

The strong law and $\alpha_n \to \alpha$ imply

$$G_n(t_\alpha \pm \varepsilon) \to G(t_\alpha \pm \varepsilon) \quad \text{almost surely,}$$

and hence, eventually almost surely,

$$G_n(t_\alpha - \varepsilon) > \alpha_n > G_n(t_\alpha + \varepsilon).$$

Since $G_n(t_n) = \alpha_n$, monotonicity gives $|t_n - t_\alpha| < \varepsilon$. Intersecting the probability-one events for $\varepsilon = 1/k$, $k \ge 1$, yields $t_n \to t_\alpha$ almost surely.

Uniform local density bound. By continuity and positivity, choose fixed $r, \gamma > 0$ such that

$$f(t) \ge 2\gamma \qquad (|t - t_\alpha| \le 2r).$$

Cover $[t_\alpha - 2r, t_\alpha + 2r]$ with a fixed finite grid of mesh at most $\gamma/(2L)$, including both endpoints. At each grid point t_j, the strong law gives $f_n(t_j) \to f(t_j)$ almost surely. Because the grid is finite, eventually

$$f_n(t_j) \ge \frac{3\gamma}{2} \qquad \text{simultaneously for all grid points.}$$

The common Lipschitz bound for f_n then gives, for each t in the interval and a grid point with $|t - t_j| \le \gamma/(2L)$,

$$f_n(t) \ge f_n(t_j) - L|t - t_j| \ge \frac{3\gamma}{2} - \frac{\gamma}{2} = \gamma.$$

This holds uniformly on the interval eventually almost surely.

Membership approximation and transfer. On the common probability-one event above, $t_n \to t_\alpha$ implies that eventually

$$[t_n - r, t_n + r] \subseteq [t_\alpha - 2r, t_\alpha + 2r].$$

Corollary 2 therefore applies with $c = \gamma$ and radius r. Since $S_i \ge s_*$, it gives

$$\max_{i\le n} |p_{i,n} - \pi_i(t_n)| \le \frac{\sqrt{n-1}}{2n\gamma s_*} + 2\exp\left(-\frac{2n^2\gamma^2 r^2}{n-1}\right) \longrightarrow 0.$$

Finally, $|\pi_i'(t)| = f_i(t) \le b$ uniformly in i, t, and $U_i = \pi_i(t_\alpha)$. Hence

$$\max_{i\le n} |\pi_i(t_n) - U_i| \le b|t_n - t_\alpha| \longrightarrow 0.$$

Combining the two bounds yields

$$\max_{i\le n} |p_{i,n} - U_i| \le \max_{i\le n} |p_{i,n} - \pi_i(t_n)| + b|t_n - t_\alpha| \longrightarrow 0 \qquad \text{almost surely.}$$

□

The uniform approximation in Corollary 5 is stronger than coordinatewise convergence and is precisely the condition used by Theorem 4 to transfer the population score ordering to the growing finite rosters.

B.4 A GAUSSIAN EXAMPLE WITH POSITIVE LIMITING GAIN

Complementing the symmetric-median zero-gain result of Corollary 4, this Gaussian example has positive limiting normalized gain. In the iid, independent Gaussian posterior setup of Corollary 3, take independent $M \sim N(0, \sigma^2)$, $S \sim \mathrm{Unif}[s_l, s_h]$, with $\sigma > 0$, $0 < s_l < s_h < \infty$, and $K_n/n \to \alpha \in (0, 1/2)$. Set

$$U = \Phi((M - t_\alpha)/S), \quad G(t_\alpha) = \alpha, \quad b = \sigma\Phi^{-1}(1-\alpha), \quad B_m = \{M > b\}.$$

Probabilities $\mathbb{P}$ and expectations $\mathbb{E}$ use the (M, S) law; H is the induced law of (M, U) in Theorem 4. Conditional on S, U is a strictly increasing continuous transform of continuous M, so M, U are atomless and the capacity thresholds in (6) exist without boundary mass.

For an independent standard normal Z, $M + SZ \mid S = s \sim N(0, \sigma^2 + s^2)$. Since $b > 0$,

$$\mathbb{P}(M + sZ > b) = 1 - \Phi\left(\frac{b}{\sqrt{\sigma^2 + s^2}}\right) > 1 - \Phi(b/\sigma) = \alpha.$$

Integrating over S gives $G(b) > \alpha$; strict decrease of G yields $t_\alpha > b$. Put $D = t_\alpha - b > 0$, choose $s_l < s_L < s_H < s_h$ and $0 < d < D(s_H - s_L)/(s_H + s_L)$, and define

$$E_+ = \{b - d \le M \le b - d/2,\ S \ge s_H\}, \qquad E_- = \{b + d/2 \le M \le b + d,\ S \le s_L\}.$$

Then $E_+ \subset B_m^c$ and $E_- \subset B_m$. Positive area and strictly positive Gaussian–uniform product density on both interiors give $\mathbb{P}(E_\pm) > 0$. Since $d < D$, both lie below t_α, where $\Phi((m - t_\alpha)/s)$ increases with both m and s. Hence

$$U \ge u_+ := \Phi\big(-(D + d)/s_H\big) \text{ on } E_+, \qquad U \le u_- := \Phi\big(-(D - d)/s_L\big) \text{ on } E_-.$$

The choice of d gives $(D + d)/s_H < (D - d)/s_L$, so $u_+ > u_-$.

Set $r = \min\{\mathbb{P}(E_-), \mathbb{P}(E_+)\} > 0$. By nonatomicity, choose $F_\pm \subseteq E_\pm$ with $\mathbb{P}(F_-) = \mathbb{P}(F_+) = r$. Then $B' = (B_m \setminus F_-) \cup F_+$ has $\mathbb{P}(B') = \alpha$ and

$$\mathbb{E}[U\mathbf{1}_{B'}] - \mathbb{E}[U\mathbf{1}_{B_m}] = \mathbb{E}[U\mathbf{1}_{F_+}] - \mathbb{E}[U\mathbf{1}_{F_-}] \ge r(u_+ - u_-).$$

The threshold region $B_p = \{U > \lambda\}$, with $\mathbb{P}(B_p) = \alpha$, maximizes $\mathbb{E}[U\mathbf{1}_B]$ over mass-α regions: equal masses and the sign of $U - \lambda$ give

$$\mathbb{E}[U(\mathbf{1}_{B_p} - \mathbf{1}_{B'})] = \mathbb{E}[(U - \lambda)(\mathbf{1}_{B_p} - \mathbf{1}_{B'})] \ge 0.$$

Consequently,

$$g(\alpha, H) \ge r(u_+ - u_-)/\alpha > 0. \tag{15}$$

By Corollary 3 and Theorem 4, the finite-pool normalized expected-overlap gain of Posterior Top-K over posterior-mean ranking satisfies $\Delta_D/K_n \to g(\alpha, H) > 0$ almost surely.

C EXPERIMENTAL DESIGN, NUMERICAL DIAGNOSTICS, AND PRACTICAL INTERPRETATION

C.1 MONTE CARLO ESTIMATION AND NUMERICAL CHECKS

Posterior Top-K and posterior-mean ranking use the same fitted posterior. Experiment A, ICLR, and Jester use ordinary membership counts over independent posterior draws; each draw selects K candidates and hence $\sum_i \hat{p}_i = K$. Experiments B and C and the four refinements of A use Gaussian Rao–Blackwell estimates

$$\hat{p}_i = L^{-1} \sum_{\ell=1}^{L} \Phi\big((m_i - T_{-i}^{(\ell)})/s_i\big),$$

where $T_{-i}^{(\ell)}$ is the Kth largest of the other candidates' qualities. These estimates are unbiased, but their finite-draw sum need not equal K; we do not renormalize them. Correlated Jester draws retain the full fitted posterior covariance.

The full-draw roster $\widehat{A}_p = \mathrm{TopK}(\hat{p})$ gives the plug-in gain

$$\widehat{G} = K^{-1}\left(\sum_{i\in\widehat{A}_p}\hat{p}_i - \sum_{i\in A_{\mathrm{mean}}}\hat{p}_i\right).$$

This quantity is nonnegative by construction; its sign alone does not resolve a positive gain under the exact posterior. For a numerical check, we select $\widehat{A}_p^{(1)} = \mathrm{TopK}(\hat{p}^{(1)})$ on the first half of the draws and evaluate it using the independent second half:

$$\widehat{G}_{\mathrm{split}} = K^{-1}\left(\sum_{i\in\widehat{A}_p^{(1)}}\hat{p}_i^{(2)} - \sum_{i\in A_{\mathrm{mean}}}\hat{p}_i^{(2)}\right).$$

This evaluates the first-half roster, which may differ from $\widehat{A}_p$; negative split estimates are retained. It is a Monte Carlo diagnostic, distinct from the independent Jester reference comparison. First-half/full-draw roster agreement is $|\widehat{A}_p^{(1)} \cap \widehat{A}_p|/K$.

Coordinate standard errors use independent batch means (1,024 draws per batch, or 128 for ICLR) of the actual estimator, including Rao–Blackwellization and pooling. Each second-half batch supplies a paired roster contrast. The sample standard deviation of these contrasts divided by the square root of the batch count gives the split-gain standard error, accounting for within-draw coordinate dependence. For mean split gains over J arrays or panels, independent second-half streams give standard error $\sqrt{\sum_{j=1}^{J}\mathrm{SE}_j^2}/J$, conditional on the fitted posteriors and first-half rosters. These standard errors exclude panel, hyperparameter, and reference-user uncertainty.

All finite normalizations use $\alpha_n = K/n$. Scores are sorted decreasingly, with generation/item order breaking ties, except for the fixed ICLR priority described in Appendix C.3. With $R = |\widehat{A}_p \setminus A_{\mathrm{mean}}|$, symmetric disagreement is $2R$ and $\widehat{\delta}_n = 2R/n$. A cutoff between the Kth and $(K+1)$st ordered $\hat{p}_i$ gives the average disagreement depth and verifies $\widehat{G} = \widehat{\delta}_n\widehat{\bar{d}}_n/\alpha_n$; depth is zero for empty disagreement, and cutoff ties contribute zero depth. For each coordinate, the estimator averages L iid contributions in $[0,1]$ with mean p_i. Hoeffding's inequality and a union bound give

$$\|\hat{p} - p\|_\infty \le \sqrt{\frac{\log(2n/\eta)}{2L}} \quad \text{with probability at least } 1-\eta.$$

The bound requires no independence across coordinates and can be used with Proposition 1. Here η denotes Monte Carlo failure probability, distinct from disagreement mass δ. Numerical diagnostics follow; commands, environment details, and complete array/panel records are in the supplementary README and `docs/implementation_details.md`.

C.2 Simulation designs and population targets

The designs below follow Experiments A–C in Section 5.

Experiment A: structural equivalence. We use equicorrelated Gaussian posteriors with $n \in \{24, 96\}$, $K/n \in \{1/4, 1/2\}$, unit marginal scales, and equally spaced means

$$m_i = \eta\left(i - \frac{n+1}{2}\right), \qquad i = 1,\dots,n, \qquad \eta \in \{0.04, 0.3\}.$$

For each combination, the common residual correlation is $\rho \in \{0, 0.6, -1/(n-1)\}$, giving 24 configurations. The initial membership-count estimates use 65,536 independent posterior draws per configuration. All 24 estimated rosters agree with posterior-mean ranking.

The four configurations exhibiting apparent probability-order inversions in the initial estimates are refined using 262,144 Rao–Blackwell draws from the rank-equivalent independent array $m + \sqrt{1-\rho}\,Z$, where Z has iid standard-normal coordinates. The equicorrelated representation differs from this array only by a common random shift, which does not affect realized rankings. All four apparent inversions are absent in the refined estimates; no probability postprocessing is applied. This experiment tests the structural-equivalence prediction of Theorem 1 across independent, positively correlated, and singular negatively correlated Gaussian residuals.

n	K/n	Setting	Max err.	MC SE	Repl.	Mass (%)	Agree.
64	0.203	regular	0.0072	0.0004	1	3.125	1.000
64	0.203	small scale	0.0126	0.0006	0	0.000	1.000
64	0.203	low density	0.0031	0.0004	0	0.000	1.000
64	0.500	regular	0.0044	0.0004	0	0.000	1.000
64	0.500	small scale	0.0168	0.0006	0	0.000	1.000
64	0.500	low density	0.0032	0.0003	1	3.125	1.000
256	0.199	regular	0.0021	0.0002	4	3.125	1.000
256	0.199	small scale	0.0116	0.0006	0	0.000	1.000
256	0.199	low density	0.0009	0.0002	0	0.000	1.000
256	0.500	regular	0.0014	0.0002	0	0.000	1.000
256	0.500	small scale	0.0094	0.0005	0	0.000	1.000
256	0.500	low density	0.0068	0.0002	1	0.781	1.000
1024	0.200	regular	0.0005	0.0001	16	3.125	1.000
1024	0.200	small scale	0.0084	0.0003	0	0.000	1.000
1024	0.200	low density	0.0002	0.0001	5	0.977	1.000
1024	0.500	regular	0.0004	0.0001	0	0.000	1.000
1024	0.500	small scale	0.0056	0.0003	0	0.000	1.000
1024	0.500	low density	0.0006	0.0001	5	0.977	1.000

Table 1: The 18 Gaussian arrays in Figure 2. Max err. is $\max_i |v_i - \hat{p}_i|$; MC SE is the largest coordinate standard error. Repl. is $|\widehat{A}_p \setminus A_{\text{mean}}|$, Mass is $100|\widehat{A}_p \triangle A_{\text{mean}}|/n$, and Agree. is $|A_v \cap \widehat{A}_p|/K$.

Experiment B: boundary approximation and roster disagreement. The arrays in Figure 2 have conditionally independent Gaussian qualities $q_i \mid m_i, s_i \sim N(m_i, s_i^2)$. The regular and small-scale settings use iid $m_i \sim N(0,1)$ with independent $s_i \sim \text{Unif}[0.5, 1.5]$ and $s_i \sim \text{Unif}[0.04, 0.2]$, respectively. The low-density setting uses independent

$$m_i \sim \tfrac{1}{2}N(-3, 0.15^2) + \tfrac{1}{2}N(3, 0.15^2), \qquad s_i \sim \text{Unif}[0.15, 0.3].$$

Each setting is evaluated at $n \in \{64, 256, 1024\}$ and capacities near $\alpha = 0.2$ and $\alpha = 0.5$. The integer capacity is $K = \text{round}(n\alpha)$. Each of the 18 arrays uses 131,072 independent posterior draws with the Gaussian Rao–Blackwell estimator described in Appendix C.1.

For each array and neighborhood radius r, the approximation-error bound in Corollary 2 uses the density lower bound

$$c(r) = \frac{1}{n}\sum_{i=1}^{n} \min\left\{\frac{\phi((m_i - t_n - r)/s_i)}{s_i}, \frac{\phi((m_i - t_n + r)/s_i)}{s_i}\right\}.$$

Each Gaussian component density is unimodal, so its minimum on $[t_n - r, t_n + r]$ occurs at an endpoint. Averaging these componentwise minima gives a valid lower bound for f_n throughout the interval.

The uniform error bound uses $s_* = \min_i s_i$; each coordinate-specific bound uses s_i. We minimize each capped bound separately over the same 70-point geometric grid of neighborhood radii,

$$\mathcal{G}_r = \left\{0.02\ \text{median}(s_1, \ldots, s_n)\ 100^{j/69} : j = 0, \ldots, 69\right\}.$$

Thus the candidate radii range from $0.02\ \text{median}(s_1, \ldots, s_n)$ to $2\ \text{median}(s_1, \ldots, s_n)$.

Table 1 distinguishes approximation accuracy from disagreement between the two principal selection rules.

Experiment C: disagreement magnitude and stabilization. The population experiments in Figures 4 and 6 use independent $M \sim N(0, \sigma^2)$ and $S \sim \text{Unif}[l, h]$, with four representative regimes:

Regime	σ	l	h	α
Near-indifferent	0.001	0.99	1.01	0.2
Moderate depth	0.5	0.5	1	0.2
Separated exchange	1	0.3	2	0.2
Symmetric median	1	1	2	0.5

For each regime, four independent parameter sequences are evaluated at nested prefixes $n \in \{64, 256, 1024\}$. The full 54-point population grid crosses $\sigma \in \{0.001, 0.04, 0.15, 0.5, 1, 2\}$, scale intervals $[0.99, 1.01]$, $[0.5, 1]$, and $[0.3, 2]$, and capacities $\alpha \in \{0.1, 0.2, 0.5\}$. Figure 4(a) displays its 18 points at $\alpha = 0.2$, together with the four representative regimes; the symmetric-median marker has $\alpha = 0.5$. Gray reference lines are $g = \delta\bar{d}/\alpha_{\mathrm{ref}}$ for fixed depths 0.005, 0.02, and 0.08 at $\alpha_{\mathrm{ref}} = 0.2$.

For each parameter sequence and pool size, we set $K = \mathrm{round}(n\alpha)$ and estimate membership probabilities using $65{,}536$ independent posterior draws with the Gaussian Rao–Blackwell estimator. The four regimes, four parameter sequences per regime, and three nested pool sizes give 48 finite arrays.

Population targets follow the Gaussian specialization of Theorem 4. For an independent $Z \sim N(0,1)$, $M + SZ \mid S = s \sim N(0, \sigma^2 + s^2)$, so the quality boundary solves

$$G(t) = \mathbb{E}_S\left[\bar{\Phi}\left(\frac{t}{\sqrt{\sigma^2 + S^2}}\right)\right], \qquad G(t_\alpha) = \alpha,$$

where $\bar{\Phi} = 1 - \Phi$. The posterior-mean boundary is $b = \sigma\Phi^{-1}(1-\alpha)$. With $U = \Phi((M - t_\alpha)/S)$ and $z = \Phi^{-1}(\lambda)$, the membership threshold satisfies

$$\mathbb{E}_S\left[\bar{\Phi}\left(\frac{t_\alpha + Sz}{\sigma}\right)\right] = \alpha.$$

Conditional on $S = s$, the rules disagree between b and $t_\alpha + sz$. Integrating the population mass–depth expression of Theorem 4 over this region gives

$$g(\alpha, H) = \frac{1}{\alpha}\mathbb{E}_S\left[\int_{\min\{b, t_\alpha + Sz\}}^{\max\{b, t_\alpha + Sz\}} \left|\Phi\left(\frac{m - t_\alpha}{S}\right) - \lambda\right| \phi_\sigma(m)\, dm\right],$$

where $\phi_\sigma(m) = \phi(m/\sigma)/\sigma$. Disagreement mass and average depth are computed over the same region. Population integrals use Gauss–Legendre quadrature with 128 nodes per dimension, checked with 256 nodes; the largest change in g is 2.46×10^{-6} across the 54 grid points. The corresponding maximum changes in mass and depth are below 4.94×10^{-4} and 1.58×10^{-4}, respectively.

Figure 3(a) reuses the separated-exchange sequence 0 at $n = 1024$, $K = 205$: $m_i \sim N(0,1)$ and independent $s_i \sim \mathrm{Unif}[0.3, 2]$. Parameter and posterior-draw seeds are 20273121 and 20282931. Its colors use the full-draw roster from 65,536 Rao–Blackwell draws; its dashed boundary uses the smallest selected $\hat{p}_i$. All coordinates are displayed without filtering. The supplementary `configs/simulations.json` and `docs/experiment_map.csv` give every parameter/draw seed, saved array, and figure correspondence.

Replaying the 18 B and 48 C arrays reproduces their saved results. The C diagnostics below summarize four sequences per row: Max SE is the largest coordinatewise Monte Carlo standard error across the four arrays, Half agree. is the minimum first-half/full-draw roster agreement, and the remaining columns are mean full-draw gain, mean split gain, and the Monte Carlo SE of the latter, in percentage points. The 35 C arrays with nonzero full-draw gain also have positive split estimates exceeding 1.96 paired Monte Carlo SE; the other 13 have identical rosters and zero contrast. These are numerical checks of the finite estimates, not convergence-rate claims.

Regime	n	Max SE	Half agree.	Plug-in	Split	Split SE
Near-indifferent	64	0.00023	0.923	0.1391	0.1380	0.000863
Near-indifferent	256	0.00012	0.980	0.1383	0.1383	0.000106
Near-indifferent	1024	0.00006	0.995	0.1554	0.1554	0.000020
Moderate depth	64	0.00045	1.000	0.2014	0.2015	0.000544
Moderate depth	256	0.00022	1.000	0.3412	0.3413	0.000281
Moderate depth	1024	0.00012	1.000	0.3326	0.3326	0.000155
Separated exchange	64	0.00094	1.000	0.3122	0.3134	0.000793
Separated exchange	256	0.00051	1.000	0.7932	0.7927	0.000913
Separated exchange	1024	0.00032	1.000	0.8808	0.8807	0.000555
Symmetric median	64	0.00037	1.000	0.0007	0.0007	0.000145
Symmetric median	256	0.00020	1.000	0.0000	0.0000	0.000000
Symmetric median	1024	0.00010	1.000	0.0001	0.0001	0.000002

C.3 ICLR 2026 DATA, POSTERIOR FITTING, AND CAPACITY SENSITIVITY

Data and analysis cohort. We use official numeric review scores from a fixed public snapshot of `JerMa88/ICLR_Peer_Reviews_2026` (JerMa88, n.d.), archived on September 23, 2026. A paper enters the cohort if it has at least one finite numeric official-review score; withdrawn, desk-rejected, and decision-unknown papers are not excluded by status. We verify each score against the official-review note preserved in the snapshot, retain the latest saved record for each reviewer–paper pair (modification time, creation time, then review ID), and remove two duplicate records. The resulting $n = 19{,}474$ papers have 75,857 ratings, with 2–9 per paper and 14,834 papers having four. This snapshot is not a reconstruction of information available at the final-decision date.

Realized capacity. Snapshot decision fields are not used as ground truth or model inputs. To determine the capacity within this scored cohort, we match canonical OpenReview paper IDs to visible main-conference oral/poster entries in the separately archived official program (ICLR, 2026), obtaining

$$K_{\text{actual}} = 5{,}353, \qquad \alpha_{\text{actual}} = \frac{5{,}353}{19{,}474} \approx 0.2749.$$

The archived program has 5,576 eligible event records and 5,353 unique paper IDs, all present in the cohort; oral/poster appearances of the same paper count once. Thus K_{actual} is an observed cohort intersection, not an independent claim about the conference's total number of acceptances. Only this aggregate capacity enters the analysis; the program roster is not treated as latent top-K ground truth.

Empirical-Bayes posterior. We fit the independent Gaussian model

$$y_{ir} = q_i + e_{ir}, \qquad q_i \sim N(\mu, v_q), \qquad e_{ir} \sim N(0, v_e).$$

Hyperparameters (μ, v_q, v_e) are estimated by marginal likelihood and held fixed for posterior analysis. For paper i with r_i ratings and mean $\overline{y}_i$, the independent posterior $q_i \mid D \sim N(m_i, s_i^2)$ has

$$s_i^2 = (v_q^{-1} + r_i/v_e)^{-1}, \qquad m_i = \mu + s_i^2 r_i(\overline{y}_i - \mu)/v_e.$$

The fitted values are $(\mu, v_q, v_e) \approx (4.204, 0.624, 2.845)$. Posterior scales have mean 0.5808, SD 0.01723, CV 0.02967, and range $[0.458, 0.6586]$, indicating weak uncertainty heterogeneity.

Actual-capacity analysis. Membership probabilities use 32,768 independent joint Gaussian draws, shared across capacities. The 79 groups with exactly equal stored (m_i, s_i) coincide with groups having equal rating count and sum. By exchangeability within each group, we average its membership frequencies, separately for the full sample, first half, and each 128-draw batch. Individual qualities remain independently sampled; pooling does not make them share a quality draw. Standard errors use the pooled batch scores, retaining dependence within each draw, as in Appendix C.1. Score ties use the saved order of SHA-256 paper-ID keys throughout. At the realized capacity, the common quality boundary and midpoint membership-probability cutoff are

$$t_n \approx 4.677, \qquad \widehat{\tau} \approx 0.4307.$$

Posterior-mean selection, the standardized-boundary approximation A_v, and the full-draw Posterior Top-K estimate $\widehat{A}_p$ select the same roster. Hence the estimated disagreement mass, depth (zero for empty disagreement), and normalized expected-overlap difference are all zero.

The maximum coordinate difference between the boundary approximation and $\widehat{p}$ is 0.004896; the median, 95th-percentile, and maximum coordinate Monte Carlo SEs are 0.0000464, 0.000163, and 0.002712. First-half and full-draw rosters also agree. The computable bound $Q(v, e)/K = 0.002895$ remains conservative, so empirical agreement is not an analytic exact-recovery certificate. Figure 3(b) shows the 1,999-member identical-parameter boundary group, of which 1,499 members are selected by the shared tie rule. This local tie structure does not establish full residual exchangeability.

Capacity sensitivity. The realized capacity is the main analysis. Table 2 assesses dependence on capacity at $K = \text{round}(n\alpha)$ for $\alpha \in \{0.1, 0.2, 0.3, 0.5\}$, using the same fitted posterior and draws.

K	K/n	Repl.	$\widehat{\delta}_n$	$\widehat{\bar{d}}_n$	$\widehat{\Delta}_D/K$	Agree
1,947	0.1000	6	0.000616	0.001572	9.69×10^{-6}	1.000
3,895	0.2000	1	0.000103	0.006501	3.34×10^{-6}	1.000
5,842	0.3000	0	0	0	0	1.000
9,737	0.5000	0	0	0	0	1.000

Table 2: ICLR 2026 capacity sensitivity. Repl. is $|\widehat{A}_p \setminus A_{\text{mean}}|$; disagreement mass uses the full symmetric difference. Depth is the average distance from the estimated membership-probability cutoff, with zero for empty disagreement. Agree is $|A_v \cap \widehat{A}_p|/K$. The main realized-capacity analysis uses $K = 5{,}353$.

This sensitivity grid excludes K_{actual}. At all four capacities, the approximate roster agrees with the estimated membership roster.

For all five capacities the first-half/full-draw roster agreement is one. The following audit uses the saved independent halves; split SEs are computed from paired roster differences in the 128 second-half batches. Identical rosters give exactly zero differences and SEs. The two nonzero sensitivity gains are small, with appreciable Monte Carlo uncertainty, particularly at $K = 1{,}947$. Gains and Split SE are reported as normalized overlap fractions; Max SE is in probability units.

K	Max SE	Half/full	Plug-in gain	Split gain	Split SE
1,947	0.0027	1.000	9.69×10^{-6}	7.87×10^{-6}	3.93×10^{-6}
3,895	0.0027	1.000	3.34×10^{-6}	3.00×10^{-6}	9.48×10^{-7}
5,353*	0.0027	1.000	0	0	0
5,842	0.0027	1.000	0	0	0
9,737	0.0028	1.000	0	0	0

Here $K = 5{,}353$ is marked by *. The supplementary `docs/implementation_details.md` (ICLR section) specifies the fitting, grouping, and tie rules; `docs/data_provenance.md` records pinned source hashes and the numerical anonymization. Complete diagnostics are provided in `results/diagnostics/iclr/diagnostics.csv`.

C.4 Jester data, posterior fitting, and external reference

Data construction. The pinned Jester snapshot (SeppeV, n.d.) contains 1,761,439 observed ratings from 59,132 users on 140 items. We use its original rating column on $[-10, 10]$, without standardization or imputation; missing user–item pairs are absent records. We retain the 100 items with the largest observed-rating counts, breaking count ties by lexicographic item ID, and retain users observed on all 100 items. This availability-based rule gives 3,091 complete-case users. A fixed random split assigns 128 to training and 2,963 to the independent reference pool. Reference rating values do not enter panel construction, masks, priors, fitting, or tuning. The target is this complete-case user population, which may differ from users with missing ratings.

Sparse training panels. Each of two sparsity designs contains 60 panels. For each panel, 32 of the 128 training users are sampled uniformly without replacement; users can recur across panels. In the 2/6 design, 50 items receive two ratings and 50 receive six; in the 4/12 design, the counts are four and twelve. The low- and high-count item groups are randomized within each panel. Each item's raters are sampled uniformly without replacement from its 32 panel users, and the rating mask is generated before fitting. All capacities within a fitted panel use the same observations and posterior.

Posterior model. We fit the Gaussian item-plus-user model

$$y_{ij} = q_i + b_j + e_{ij}, \qquad q_i \sim N(\mu, v_q), \quad b_j \sim N(0, v_b), \quad e_{ij} \sim N(0, v_e),$$

with mutually independent priors and errors. Each panel's hyperparameters (μ, v_q, v_b, v_e) are estimated separately by Gaussian marginal likelihood and held fixed for posterior analysis. Integrating out user effects gives $q \mid D \sim N(m, C_q)$. Because users can rate several items, C_q is generally

nondiagonal; Cholesky sampling retains this full covariance. The supplementary implementation notes give the posterior matrix equations, three optimization starts, parameter bounds, and stopping tolerances.

Membership probabilities and conditional evaluation. Each fitted panel uses 1,048,576 posterior draws at $K \in \{10, 20, 30, 50\}$, with actual $\alpha_n = K/100$. Full-draw membership frequencies give $\hat{p}$ and $\widehat{A}_p = \mathrm{TopK}(\hat{p})$, compared with $A_{\mathrm{mean}} = \mathrm{TopK}(m)$ under the same fitted posterior. Score ties use increasing candidate index in the saved item order. The count estimator, split-draw evaluation, and paired-batch standard errors follow Appendix C.1; the first and second halves each contain 512 disjoint batches of 1,024 draws.

Mass, depth, and model-conditional difference. For panel r and a fixed capacity, define

$$S_r = \widehat{A}_{p,r} \triangle A_{\mathrm{mean},r}, \qquad \hat{\delta}_r = \frac{|S_r|}{100}, \qquad \hat{\tau}_r = \frac{\hat{p}_{(K),r} + \hat{p}_{(K+1),r}}{2},$$

$$\hat{\bar{d}}_r = \frac{1}{|S_r|} \sum_{i \in S_r} |\hat{p}_{i,r} - \hat{\tau}_r|,$$

with depth zero when S_r is empty. Theorem 3, applied to the estimated membership scores, gives

$$\widehat{G}_r = \frac{\sum_{i \in \widehat{A}_{p,r}} \hat{p}_{i,r} - \sum_{i \in A_{\mathrm{mean},r}} \hat{p}_{i,r}}{K} = \frac{\hat{\delta}_r \hat{\bar{d}}_r}{\alpha_n}.$$

For each 60-panel summary, Table 3 reports mean mass and pooled depth:

$$\hat{\delta} = \frac{1}{60} \sum_r \hat{\delta}_r, \qquad d_{\mathrm{pool}} = \frac{\sum_r \hat{\delta}_r \hat{\bar{d}}_r}{\sum_r \hat{\delta}_r}, \qquad \frac{\hat{\delta}\, d_{\mathrm{pool}}}{\alpha_n} = \frac{1}{60} \sum_r \widehat{G}_r.$$

Pooled depth is disagreement-mass-weighted, with value zero if total mass is zero. Mass is displayed as $100\hat{\delta}$, and Model as $100 \cdot 60^{-1} \sum_r \widehat{G}_r$ percentage points.

Numerical diagnostics. The following summary uses all 60 panels in each row. Max. SE is the largest coordinatewise Monte Carlo standard error across these panels; Agreement is mean first-half/full-draw roster agreement. Plug-in and Split are mean normalized gains. Conditional on the fitted posteriors and first-half rosters, Split SE combines paired second-half variances from independent Monte Carlo streams as in Appendix C.1. All gains and the final SE column are in percentage points.

Ratings	K	Max. SE	Agreement	Plug-in (pp)	Split (pp)	Split SE (pp)
2/6	10	0.00052	0.972	0.1349	0.1309	0.0011
2/6	20	0.00052	0.971	0.0547	0.0513	0.0007
2/6	30	0.00052	0.974	0.0244	0.0204	0.0006
2/6	50	0.00052	0.986	0.0021	-0.0004	0.0003
4/12	10	0.00051	0.997	0.1726	0.1718	0.0010
4/12	20	0.00052	0.998	0.0848	0.0846	0.0006
4/12	30	0.00052	0.997	0.0270	0.0263	0.0004
4/12	50	0.00052	0.999	0.0005	0.0003	0.0001

Across all 480 panel–capacity cases, mean first-half/full-draw agreement is 0.9868, with a minimum of 0.40; 29 split estimates are negative and are retained. At $K = 50$, the 2/6 design's positive plug-in mean is not resolved as a positive split-draw gain. These are numerical checks of first-half selected rosters, not external validation; panel-level diagnostics and the saved half estimates are included in the supplement.

Independent reference comparison. Each reference score $\overline{y}_i^{\mathrm{ref}}$ is the mean rating from the 2,963 disjoint reference users. Let $T_K^{\mathrm{ref}} = \mathrm{TopK}(\overline{y}^{\mathrm{ref}})$, with the same candidate-index tie rule. For panel r, the independent-reference difference is

$$E_r = \frac{|\widehat{A}_{p,r} \cap T_K^{\mathrm{ref}}| - |A_{\mathrm{mean},r} \cap T_K^{\mathrm{ref}}|}{K}.$$

Ratings	α	Mass (%)	Depth	Model (pp)	External (pp), 95% CI
2/6	0.1	2.50	0.0054	0.135	-0.83 [-2.81, 1.14]
2/6	0.2	3.37	0.0033	0.055	-0.50 [-1.96, 0.96]
2/6	0.3	3.30	0.0022	0.024	-0.44 [-1.33, 0.44]
2/6	0.5	2.17	0.0005	0.002	-0.33 [-0.69, 0.02]
4/12	0.1	1.57	0.0110	0.173	0.00 [-1.35, 1.35]
4/12	0.2	2.03	0.0083	0.085	0.58 [-0.17, 1.34]
4/12	0.3	1.53	0.0053	0.027	0.39 [-0.21, 0.98]
4/12	0.5	0.23	0.0011	0.001	0.03 [-0.08, 0.15]

Table 3: Jester summaries over 60 panels per design and capacity. Mass is mean symmetric disagreement (%); Depth is the mass-weighted pooled disagreement depth. Model is the mean full-draw plug-in estimate of the model-conditional expected-overlap difference per selected position. External reports the independent-reference mean and approximate paired-panel 95% confidence interval. Model and External are in percentage points (pp).

Reference users never enter posterior fitting, so this comparison uses independent evidence. Across items, reference-mean standard errors range from 0.0762 to 0.103 rating units. The median ratio of fitted posterior scale to reference-mean standard error is 9.92, indicating substantially greater information in the reference sample. We nevertheless treat the resulting ranking as a high-information empirical reference rather than exact ground truth.

For each design–capacity setting, the approximate paired-panel 95% confidence interval in Figure 5 and Table 3 is

$$\overline{E} \pm t_{59,0.975} \frac{s_E}{\sqrt{60}},$$

where $\overline{E}$ and s_E are the mean and sample standard deviation of E_r across the 60 panels. This interval summarizes training-panel and sparse-mask variation conditional on the fixed training and reference pools. It is not a Monte Carlo error interval and does not include reference-user sampling uncertainty.

The supplementary README and `docs/implementation_details.md` specify pinned data acquisition, preprocessing, fixed splits and masks, fitting, and verification commands. The package contains fitted posterior objects, random seeds, and panel-level outputs; source ratings are acquired separately from the pinned public snapshot.

C.5 Additional diagnostics for boundary approximation and stabilization

Boundary-approximation diagnostics. Using the fixed radius grid and valid density lower bounds in Appendix C.2, Table 4 compares the uniform regret bound with Q/K. Although Q/K can be substantially tighter, no array satisfies the strict exact-roster certificate: empirical roster agreement is not itself an analytic certificate.

		Regular			Small scales			Low density		
n	α_n	e_n	Global/K	Q/K	e_n	Global/K	Q/K	e_n	Global/K	Q/K
64	0.203	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
64	0.500	1.000	1.000	0.994	1.000	1.000	1.000	1.000	1.000	1.000
256	0.199	0.944	1.000	0.896	1.000	1.000	1.000	0.765	1.000	0.857
256	0.500	0.375	0.751	0.095	1.000	1.000	1.000	1.000	1.000	1.000
1024	0.200	0.240	0.480	0.063	1.000	1.000	1.000	0.210	0.419	0.081
1024	0.500	0.136	0.273	0.011	1.000	1.000	1.000	1.000	1.000	1.000

Table 4: Finite boundary-approximation diagnostics for the 18 Experiment B arrays. The column e_n is the uniform approximation radius, Global/K is the capped uniform regret bound $\min(1, 2e_n)$, and Q/K is the computable interval bound from Appendix A.3.

Population stabilization. Figure 6 separates the mass, depth, and normalized difference in Theorem 4, complementing Figure 4(b). The four regimes illustrate stabilization, not monotone convergence or a convergence rate. No separate depth limit is asserted when $\delta = 0$.

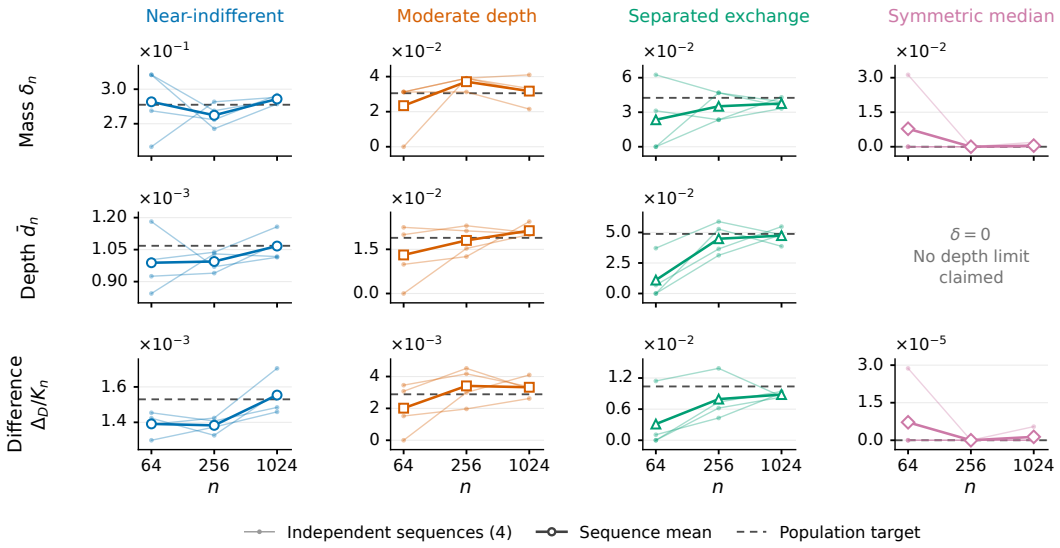


Figure 6: Finite-to-population stabilization: regimes by column; full-draw Monte Carlo estimates of mass δ_n, depth $\bar{d}_n$, and normalized difference Δ_D/K_n by row. Light paths show four independent sequences, markers their means, and dashes population targets. Vertical scales vary. No separate depth limit is claimed for the symmetric-median regime, where $\delta = 0$.

C.6 Practical interpretation of the real-data results

Posterior Top-K need not yield a practically substantial improvement over posterior-mean ranking under the same fitted posterior. In the ICLR analysis, posterior uncertainty was similar across papers, and posterior-mean ranking and the estimated posterior Top-K rule selected the same roster at the realized cohort capacity. In Jester, the rules sometimes disagreed, but the exchanged items had estimated membership probabilities close to the selection cutoff, yielding small estimated model-conditional gains. Evaluation against independent reference ratings did not establish a clear advantage for either rule. Practitioners should therefore examine both how many candidates a rule exchanges and how strongly the fitted posterior distinguishes them, as quantified by disagreement mass and depth, and assess whether any predicted improvement is supported by independent evidence.