PRECISION THRESHOLDS FOR SHARED EVALUATION CACHES

Anonymous authors
Paper under double-blind review

ABSTRACT

Model evaluation is often staged: a library is scored once on a shared set of examples, and later reports ask for more precise comparisons without rebuilding the entire evaluation table. We ask how far such a finite cache can be reused before fresh evaluation must scale with the size of the library. We consider h complete i.i.d. score vectors in $[0,1]^K$, arbitrary unknown cross-model dependence, and a fixed collection of mean differences that must satisfy a uniform marginal MSE target under a hard cap on fresh work. The key obstruction is residual candidate uncertainty: additional measurements of a shared reference improve every contrast, but cannot uniformly remove the uncertainty left in each candidate's finite cache. For a unit-price reference star with M candidates, this yields an exact threshold $d_0(h) = 1/[4(\sqrt{h}+1)^2]$ separating bounded, $\sqrt{M}$, and linear fresh work in M; matching lower bounds hold for fully adaptive acquisition, with critical work $(2+o(1))h\sqrt{M}$ and a $1/(h\sqrt{M})$ transition window. A stronger layout-wise accuracy contract changes the critical law to $\Theta(M + h\sqrt{M})$, with an integer crossover governed by M/h^2. For partial observations on general comparison graphs, we give an exact bounded-score reduction to edgewise union counts and asymptotically exact high-precision work limits.

1 INTRODUCTION

Model evaluation is increasingly cumulative rather than one-shot. A model library is often scored once on a shared benchmark or evaluation stream, and the resulting per-example table is retained for later analysis (Liang et al., 2023; Biderman et al., 2024). Subsequent decisions, however, need not ask the same statistical question as the original evaluation. A release may require tighter comparisons, a model-compression study may need a more precise comparison with its uncompressed reference, or a routing system may revisit a fixed set of candidate–reference differences under the same evaluation distribution. Re-running the full library is expensive, so the natural goal is to refine the report by reusing the existing table and evaluating only selected models on new examples.

The difficulty is that cache reuse is asymmetric. A new observation of a shared reference improves the same mean appearing in every candidate–reference contrast. The cached candidate scores, in contrast, remain finite. Without additional structural assumptions on cross-model dependence, observations of other coordinates cannot uniformly remove this candidate-specific uncertainty. A seemingly simple reuse decision therefore hides a statistical phase change:

> *when can a finite shared cache absorb a precision refinement, and when must the fresh evaluation cost begin to scale with the size of the model library?*

Existing viewpoints illuminate the two ends of this question, leaving the boundary between them unresolved. Classical optimal design and modern allocation methods describe how fresh measurements should be distributed once the relevant information model is fixed (Elfving, 1952; Kiefer, 1959; Pukelsheim, 2006; Cowen-Breen et al., 2026). Query-efficient and collaborative evaluation methods exploit selective measurement or structure in existing score tables (Kossen et al., 2021; Fogliato et al., 2025; Helm et al., 2026; Fisch et al., 2026). If the cache is treated as effectively exact, refining only the common reference looks sufficient; if the cache is ignored and one starts from a high-precision allocation problem, fresh information must be spread across active coordinates. The finite-cache

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.

regime lies between these limits. Locating it requires the *exact* residual risk left by the cache, rather than only a first-order $1/h$ approximation, and the conclusion must remain valid even when the policy reads the cache, adapts its queries, and stops from observed data.

We formalize this regime for a fixed collection of mean differences. The cache contains h complete i.i.d. score vectors in $[0,1]^K$; dependence among model scores within a row is unrestricted. A later, score-independent workload asks for the declared differences to meet a common marginal MSE target d, uniformly over all bounded score laws, while fresh work is bounded on every run. This contract captures the repeated-sampling accuracy of a fixed report and keeps the resource question separate from selection after seeing the reported values, simultaneous coverage, or conditioning on one realized numerical cache. The central insight is that the scalar minimax risk remaining after h cached observations becomes an operational threshold once the same finite cache must support many comparisons. Crossing that threshold changes the dependence of optimal fresh work on library size.

1.1 Overview of Our Results

Finite-cache phase transition. For a unit-price reference star with M candidates, the cache leaves bounded-mean minimax risk $d_0(h) = 1/[4(\sqrt{h}+1)^2]$ on each candidate side. This is the center of the acquisition transition. Writing $W^\star_{h,M}(d)$ for the minimum adaptive hard work, for fixed h,

$$W^\star_{h,M}(d) = \begin{cases} O_{h,d}(1), & d > d_0(h), \\ \Theta_h(\sqrt{M}), & d = d_0(h), \\ \Theta_{h,d}(M), & 0 < d < d_0(h). \end{cases}$$

Above the threshold, fresh reference measurements can absorb the refinement; below it, hard instances force nonvanishing new information into every candidate. At equality, randomized candidate refreshes balance the shared reference error and produce the intermediate $\sqrt{M}$ law. The transition is not an artifact of a nonadaptive construction: a single product beta prior, posterior factorization, and predictable posterior-variance reductions charge all candidates in one adaptive experiment. This gives matching lower bounds for cache-dependent querying, adaptive stopping, and arbitrary estimators. In joint limits, critical work is $(2+o(1))h\sqrt{M}$, and when $h\sqrt{M}\{d - d_0(h)\} \to t$ the normalized work converges to a piecewise limit. Hence the transition window has width $1/(h\sqrt{M})$. On this scale the correction in $d_0(h) = 1/(4h) - 1/(2h^{3/2}) + O(h^{-2})$ is operationally first-order: replacing the exact center by its leading approximation can place a target in the wrong acquisition regime.

Accuracy contracts and randomization. The preceding phase transition uses marginal MSE, which averages over the randomization of a precommitted layout. We compare it with a stronger layout-wise contract requiring every realized score-independent allocation to meet the same target. At critical precision, a marginal lottery may spread a sub-unit expected refresh across many candidates, whereas every certified layout must refresh each candidate at least once. The resulting critical work is $\Theta(M + h\sqrt{M})$. More sharply, when $M/h^2 \to \eta \in (0,\infty)$, the h^2-normalized work converges to $\min_{k \in \mathbb{Z}_{\geq 1}}(\eta k + 1/k)$. The scale M/h^2 therefore separates two effects that are easy to conflate: savings from sharing the reference mean, which remain available to deterministic layouts, and savings from randomizing fractional candidate refreshes, which depend on the accuracy contract.

Partial observations and graph allocation. We then leave the star and allow each fresh row to reveal an arbitrary subset of coordinates on a general comparison graph. For balanced precommitted layouts, the all-estimator bounded-score minimax risk of an edge depends exactly on the distribution of its union count U_e through a one-dimensional functional $g(\nu_e)$. The converse allows arbitrary within-row dependence and estimators using the entire observed table, so the reduction identifies the effective information in a randomized partial-observation layout rather than merely bounding a particular estimator. Building on this reduction, we obtain the exact adaptive request threshold and an asymptotically exact work law at high precision: for positive coordinate prices c, $\mathcal{W}^\star_{\mathrm{ad}}(\infty, d; c) = A(H,c)d^{-1}(1+o(1))$ as $d \to 0$ with $dh \to 0$, where $A(H,c)$ is a reciprocal-count allocation program. A multivariate van Trees converse allows fully adaptive sampling and arbitrary estimators, while a deterministic private-sampling construction attains the same leading constant. Thus the star theory identifies *when* fresh information must spread beyond the reference, and the graph theory identifies *where* that information must go once it does.

Related work. We give a more detailed discussion of related work in B.

2 PRELIMINARIES

The introduction isolates the finite cache as the source of the precision transition. We now formalize the statistical experiment, the resource ledger, and the accuracy contract used throughout the paper.

Let $[K] = \{1, \dots, K\}$ and let $H = ([K], F)$ be a fixed nonempty simple comparison graph. Fix an orientation of each edge $e = (i, j)$ and write $\theta_e = \mu_i - \mu_j$, where $\mu = \mathbb{E}_P Y$ and

$$\mathcal{P}_{\mathrm{bd}} = \{P : \mathrm{supp}(P) \subseteq [0,1]^K\}.$$

The vectors $Y_1, Y_2, \dots$ are i.i.d. from an unknown P, with arbitrary dependence among coordinates within each vector. The available cache is $\mathcal{C} = (Y_1, \dots, Y_h)$ for integer $h \geq 1$. The models, scoring rule, evaluation distribution, workload, prices, and target are fixed independently of all evaluation scores. The workload may be specified after cache collection provided that its definition uses external information rather than the cached scores. We study refinement under an unchanged population and treat the cost of constructing the complete cache as sunk.

Before fresh row Y_{h+t} is revealed, a policy π chooses $S_t \subseteq [K]$ using the cache, completed observations, and independent randomization. It observes only Y_{h+t,S_t} and may stop adaptively. Let r be the number of nonempty fresh requests and let $R < \infty$ be a hard request cap. Write $x_i = |\{t : i \in S_t\}|$. Every coordinate has a positive price c_i, and the additional work is

$$W_c = \sum_{i \in [K]} c_i x_i. \tag{1}$$

At unit prices we write $W = W_c$. A hard work reserve w requires $W_c \leq w$ on every transcript. Querying several coordinates on one row consumes one request and multiple evaluations. At request price $a \geq 0$, the joint charge is $W_c + ar$. We use $V_+ = \{i : \deg_H(i) > 0\}$ and $C_+ = \sum_{i \in V_+} c_i$; queried coordinates outside V_+ are charged as well.

For an arbitrary estimator vector $\widehat{\theta}$, define

$$\mathcal{R}_{\mathrm{marg}}(\pi, \widehat{\theta}) = \sup_{P \in \mathcal{P}_{\mathrm{bd}}} \max_{e \in F} \mathbb{E}_{\pi,P}(\widehat{\theta}_e - \theta_e)^2, \qquad \mathcal{R}_{\mathrm{bd}}(\pi) = \inf_{\widehat{\theta}} \mathcal{R}_{\mathrm{marg}}(\pi, \widehat{\theta}). \tag{2}$$

The expectation includes the cache, fresh scores, and acquisition randomness, so the risk is the repeated-sampling performance of the full protocol. For every fixed P and every fixed requested edge, the same bound applies. For $0 < d < 1$, let $\mathcal{W}^\star_{\mathrm{ad}}(\infty, d; c)$ be the infimum hard work reserve among feasible adaptive procedures, optimizing over finite request caps. The corresponding hard joint-charge optimum is $\mathcal{J}^\star_{\mathrm{ad}}(a, d; c)$.

A *precommitted* policy draws a finite layout $S = (S_1, \dots, S_R)$ from a law ν independently of all scores; empty sets may pad shorter lists. The marginal hard-work optima for such lotteries and for deterministic layouts are $\mathcal{W}^\star(R, d; c)$ and $\mathcal{W}^\star_{\mathrm{det}}(R, d; c)$. We also use $\mathcal{W}^\star_E(R, d; c)$ for precommitted expected work and $\mathcal{J}^\star(a, d; c)$ for precommitted hard joint charge. An argument ∞ denotes an infimum over finite request caps, and omitted prices mean unit costs. Section 4 defines the stronger layout-wise accuracy contract.

An admissible protocol reports the declared mean differences. For any fixed P, its largest marginal MSE satisfies

$$\max_{e \in F} \mathbb{E}(\widehat{\theta}_e - \theta_e)^2 = \sup_{\alpha \in \Delta(F)} \mathbb{E}\left[\sum_{e \in F} \alpha_e (\widehat{\theta}_e - \theta_e)^2\right], \tag{3}$$

where $\Delta(F)$ is the probability simplex and α is independent of the evaluation transcript. Thus the contract covers any externally requested entry, or any fixed weighted average of entrywise squared errors, on the fixed graph. A downstream rule that selects a contrast from the reported values, a simultaneous-error requirement, or risk conditional on the numerical cache induces a different statistical target; Appendix D separates these variants. Layout-wise accuracy below conditions on the acquisition layout while continuing to average over observed scores.

3 THE FINITE-CACHE PRECISION TRANSITION

We begin with the simplest geometry in which cache reuse can be shared across many comparisons: a reference star. This removes graph-allocation effects and isolates the central statistical tension—fresh

reference observations help all contrasts, whereas the residual uncertainty in each cached candidate can only be reduced by learning more about that candidate. The resulting model is already rich enough to reveal the exact point at which library size enters the fresh-work requirement.

Consider a unit-price star with vertices $\{0, 1, \dots, M\}$, reference 0, and contrasts $\theta_{i0} = \mu_i - \mu_0$. Set $W^\star_{h,M}(d) = \mathcal{W}^\star_{\mathrm{ad}}(\infty, d)$ and define

$$m(n) = (\sqrt{n} + 1)^{-2}, \qquad d_0(h) = \frac{m(h)}{4}, \qquad n_d = (d^{-1/2} - 1)^2. \tag{4}$$

The function m is defined for $n \geq 0$. With a complete cache of size h, $m(h)$ is the minimax risk for a bounded contrast, while $d_0(h) = m(h)/4$ is the scalar bounded-mean minimax risk for one coordinate. The associated shrinkage rule and beta equalizer are classical (Lehmann & Casella, 1998); our results use this finite-sample residual risk to characterize the additional acquisition required across a model library.

Theorem 1. *For fixed $h \geq 1$ and fixed $0 < d < 1$, as $M \to \infty$,*

$$W^\star_{h,M}(d) = \begin{cases} O_{h,d}(1), & d > d_0(h), \\ \Theta_h(\sqrt{M}), & d = d_0(h), \\ \Theta_{h,d}(M), & 0 < d < d_0(h). \end{cases} \tag{5}$$

If fresh acquisition is restricted to the reference, the infimum attainable uniform risk is $d_0(h)$; every finite reference-only reserve has strictly larger risk. The lower bounds allow cache-dependent queries, adaptive stopping, randomization, and arbitrary estimators. The upper bounds use precommitted layouts, so the same orders hold for $\mathcal{W}^\star(\infty, d)$.

No acquisition is needed for $d \geq m(h)$. Since $m(h) = 4d_0(h)$, a finite reference-only reserve can reduce the cache-only uniform contrast risk by any factor strictly below four; the boundary $d = d_0(h)$ is reached only in the limit of infinite reference-only sampling. Below $d_0(h)$, the common-prior lower bound assigns a positive expected fresh count to every candidate, so the hard reserve must scale with the library. At equality, vanishing per-candidate risk reductions can balance the remaining reference error and produce the intermediate $\sqrt{M}$ regime.

Theorem 2. *Let $h, M \to \infty$ at arbitrary relative rates. If $dh \to \rho \in (0, 1/4)$, then*

$$\frac{W^\star_{h,M}(d)}{Mh} \longrightarrow \frac{1}{4\rho} - 1. \tag{6}$$

If $h\sqrt{M}[d - d_0(h)] \to t \in \mathbb{R}$, then

$$\frac{W^\star_{h,M}(d)}{h\sqrt{M}} \longrightarrow \Psi(t), \qquad \Psi(t) = \begin{cases} 2 - 4t, & t \leq 1/4, \\ (4t)^{-1}, & t \geq 1/4. \end{cases} \tag{7}$$

In particular,

$$W^\star_{h,M}(d_0(h)) = (2 + o(1))h\sqrt{M}. \tag{8}$$

All these limits also hold for the precommitted marginal optimum $\mathcal{W}^\star(\infty, d)$.

The theorem characterizes the worst-case hard reserve, and fully adaptive acquisition has the same leading work in these regimes. Population-specific expected spending can be smaller; Appendix L gives a uniformly valid example. The term -1 in (6) keeps the cached observations as a first-order resource. By contrast, the high-precision law in Section 5.2 assumes $dh \to 0$, so its asymptotic expansion applies in a different precision regime.

For a candidate observed n times, use the scalar shrinkage estimate

$$\widehat{\mu}_i(n) = \tfrac{1}{2} + \frac{\sqrt{n}}{\sqrt{n} + 1}(\bar{Y}_i(n) - \tfrac{1}{2}),$$

whose uniform MSE is $m(n)/4$. Estimate the reference by the unbiased mean of an independent fresh block of $r_0 \geq 1$ rows, excluding its cached values. With no candidate refreshes this gives

$$\sup_P \mathbb{E}_P(\widehat{\mu}_i(h) - \bar{Y}_0^{\,\mathrm{fresh}} - \theta_{i0})^2 \leq d_0(h) + \frac{1}{4r_0}. \tag{9}$$

Hence $r_0 = \lceil [4d - m(h)]^{-1} \rceil$ suffices whenever $d > d_0(h)$.

To reach the threshold or finer precision, reserve $s \geq 0$ integer candidate evaluations. Write $s = kM + b$ with $0 \leq b < M$. Give every candidate k new observations and a uniformly selected b-subset one more each. A fixed candidate then receives $J_s = k + 1$ with probability b/M and k otherwise. Whenever $4d > \mathbb{E}m(h + J_s)$, choose

$$r_0(s, d) = \left\lceil [4d - \mathbb{E}m(h + J_s)]^{-1} \right\rceil, \qquad W = s + r_0(s, d). \tag{10}$$

Use realized-count shrinkage on each candidate and place all reference-block rows apart from the cache and new candidate rows. The cross-error expectation is zero, so the edge MSE is at most $\mathbb{E}m(h + J_s)/4 + 1/[4r_0(s, d)] \leq d$. Candidate observations can share rows; $r_0(s, d) + \lceil s/M \rceil$ requests suffice. At criticality, $s, r_0 \sim h\sqrt{M}$ attain the constant 2. The count lottery implements fractional refreshes while respecting a hard, rather than expected, work cap. The independent reference block makes sharing valid without learning a favorable candidate–reference covariance.

Give the Bernoulli mean of every coordinate an independent $\mathrm{Beta}(\sqrt{h}/2, \sqrt{h}/2)$ prior. This is one prior over admissible fixed populations. Conditional on the means, coordinates and rows are independent. An adaptive transcript has parameter-free decision kernels and coordinatewise Bernoulli likelihood factors, so its posterior remains a product. Let V_i be the final posterior variance, and let $\mathbb{E}_\Pi$ include the prior, data, and policy randomness. Write

$$s_h = h + \sqrt{h}, \quad b_h = s_h + 1, \quad \kappa_h = \mathbb{E}_{p \sim \mathrm{Beta}(\sqrt{h}/2, \sqrt{h}/2)} \sqrt{p(1-p)}.$$

Every feasible procedure satisfies $\mathbb{E}_\Pi V_i + \mathbb{E}_\Pi V_0 \leq d$. A new candidate observation reduces expected posterior variance by at most $1/(4b_h^2)$, while the reference obeys

$$d_0(h) - \mathbb{E}_\Pi V_i \leq \frac{\mathbb{E}_\Pi x_i}{4b_h^2}, \qquad \mathbb{E}_\Pi V_0 \geq \frac{\kappa_h^2}{s_h + \mathbb{E}_\Pi x_0}. \tag{11}$$

The first bound sums predictable variance decreases. The second follows from the beta posterior identity and Cauchy–Schwarz. Crucially, the counts are charged under the same prior for every candidate. With $\delta = d - d_0(h)$, $u = \mathbb{E}_\Pi x_0$, and $[v]_+ = \max\{v, 0\}$, summing gives

$$W^\star_{h,M}(d) \geq \inf_{u \geq 0} \left\{ u + 4Mb_h^2 \left[\frac{\kappa_h^2}{s_h + u} - \delta \right]_+ \right\}. \tag{12}$$

At $\delta = 0$ this implies $W^\star_{h,M} \geq [4b_h\kappa_h\sqrt{M} - s_h]_+$. Since $b_h/h \to 1$ and $\kappa_h \to 1/2$, it matches the critical allocation. Rescaling $u = zh\sqrt{M}$ in (12) gives the limiting objective $z + [z^{-1} - 4t]_+$, whose minimum is $\Psi(t)$. A count-dependent version of the posterior decrement gives (6). Appendix C provides the finite bounds and complete proofs.

The exact center matters on this scale:

$$d_0(h) = \frac{1}{4h} - \frac{1}{2h^{3/2}} + O(h^{-2}).$$

Replacing it by $1/(4h)$ shifts the center by asymptotically $\frac{1}{2}\sqrt{M/h}$ transition-window widths. Thus a first-order scalar-risk approximation can place a target in the wrong acquisition regime. A lower-order scalar correction becomes operationally first-order when it determines whether candidate-wide refreshes enter the reserve.

4 The Price of Certifying Each Layout

Section 3 shows that randomized candidate refreshes are first-order useful exactly at the finite-cache threshold. That conclusion uses the marginal accuracy contract, which averages over the randomization of a precommitted layout. We now ask whether the same resource law survives when each realized score-independent allocation must be certifiable on its own.

A marginal certificate averages over the randomization of a precommitted sampling protocol. A stronger operational requirement asks for the same error bound on every realized score-independent

allocation. Under this layout-wise contract, each supported layout must be certifiable on its own. For a precommitted layout law ν, define

$$\mathcal{R}_{\mathrm{lay}}(\nu) = \inf_{\widehat{\theta}} \sup_{S \in \mathrm{supp}\,\nu} \sup_{P \in \mathcal{P}_{\mathrm{bd}}} \max_{e \in F} \mathbb{E}_P[(\widehat{\theta}_e - \theta_e)^2 \mid S]. \tag{13}$$

Only the layout is conditioned on; cache and fresh scores remain random. Let $\mathcal{W}^\star_{\mathrm{lay}}(R, d; c)$ be the associated hard-work optimum. Every supported layout must itself be feasible, so choosing one such layout and conversely viewing a deterministic layout as a point mass gives

$$\mathcal{W}^\star_{\mathrm{lay}}(R, d; c) = \mathcal{W}^\star_{\mathrm{det}}(R, d; c). \tag{14}$$

This compares two accuracy requirements within the same precommitted class. For the star, Theorem 2 shows that the precommitted marginal optimum has the same leading critical work as fully adaptive acquisition. Write $W^{\mathrm{lay}}_{h,M}(d) = \mathcal{W}^\star_{\mathrm{lay}}(\infty, d)$.

Theorem 3. *At $d = d_0(h)$, every layout-wise feasible plan refreshes every candidate at least once. For fixed h, $W^{\mathrm{lay}}_{h,M}(d_0(h)) = \Theta_h(M)$. In joint limits $h, M \to \infty$,*

$$W^{\mathrm{lay}}_{h,M}(d_0(h)) = \Theta(M + h\sqrt{M}). \tag{15}$$

The leading behavior is

$$W^{\mathrm{lay}}_{h,M}(d_0(h)) \sim 2h\sqrt{M}, \qquad M = o(h^2), \tag{16}$$

$$\frac{W^{\mathrm{lay}}_{h,M}(d_0(h))}{h^2} \longrightarrow \Phi(\eta) := \min_{k \in \mathbb{Z}_{\geq 1}} \left\{ \eta k + \frac{1}{k} \right\}, \qquad M/h^2 \to \eta \in (0, \infty), \tag{17}$$

$$W^{\mathrm{lay}}_{h,M}(d_0(h)) \sim M, \qquad h^2 = o(M). \tag{18}$$

Set $\Delta_h(k) = m(h) - m(h+k) > 0$ and use $n_u = (u^{-1/2} - 1)^2$. Then, for every finite h, M,

$$\min_{k \geq 1} \left\{ Mk + \left(\lceil n_{\Delta_h(k)} \rceil - h \right)_+ \right\} \leq W^{\mathrm{lay}}_{h,M}(d_0(h)) \leq \min_{k \geq 1} \left\{ Mk + \lceil \Delta_h(k)^{-1} \rceil \right\}, \tag{19}$$

where both minima are over integers. To see necessity, fix a deterministic layout and a least-observed candidate with k new scores. Independent beta equalizer priors for this candidate and the reference imply edge risk at least $[m(h+k) + m(h+x_0)]/4$. If $k = 0$, this exceeds $d_0(h)$ for every finite x_0. For $k \geq 1$, feasibility forces $m(h+x_0) \leq \Delta_h(k)$ and total work is at least $Mk + x_0$. For attainment, refresh every candidate k times and use an independent unbiased reference block of $\lceil \Delta_h(k)^{-1} \rceil$ observations.

For fixed integer k and growing h, $h^2 \Delta_h(k) \to k$. Both sides of (19), divided by h^2, therefore approach $\eta k + 1/k$. When M/h^2 stays bounded away from zero, bounded normalized cost confines the minimizing k to a finite set, justifying the minimum and the limit in (17). The minimizing integer changes at $\eta = 1/[k(k+1)]$; it is $k = 1$ for $\eta \geq 1/2$. Appendix D gives the full argument, including the other two regimes.

At $M/h^2 \to \eta$, the ratio of layout-wise to precommitted marginal work is

$$\frac{W^{\mathrm{lay}}_{h,M}(d_0(h))}{\mathcal{W}^\star(\infty, d_0(h))} \longrightarrow \frac{\Phi(\eta)}{2\sqrt{\eta}}. \tag{20}$$

For $M = o(h^2)$, the critical count $h/\sqrt{M}$ per candidate diverges and integer rounding is negligible: the ratio tends to one, so reference sharing gives the leading saving without randomization. For $h^2 = o(M)$, that count is below one. A lottery can spread a fixed refresh budget over the library, but every certified layout still pays at least M; the ratio grows as $\sqrt{M}/(2h)$. Thus the comparison identifies when a fixed certifiable allocation preserves the leading resource law and when integer refreshes change it. For fixed h and $d \neq d_0(h)$, both contracts retain the bounded-work regime above the threshold and the linear-work regime below it. The precision threshold is a property of the finite-cache estimation problem, while randomization changes the critical constant and, in dense libraries, the critical order.

The contract gap can be large arbitrarily close to cache adequacy. For a complete workload on $K \geq 3$ models, Appendix D.6 constructs a target $d_K \in (m(h+1), m(h))$ for which a one-evaluation lottery is marginally feasible, whereas every layout-wise feasible plan needs work $K - 1$. Its effective additional sample deficit is $\Theta_h(1/K)$, making the source of the gap an accuracy-contract effect.

5 Partial Observations and General Comparison Graphs

The star analysis answers when candidate-side information must be refreshed, and Section 4 shows how the accuracy contract changes the critical cost. Neither question yet uses the full geometry of a shared evaluation table. We now restore that geometry: a fresh row may reveal an arbitrary subset of coordinates, so one request can serve several contrasts even though each observed coordinate still incurs work. This distinction introduces two new questions. First, what is the exact bounded-score information in a partially observed shared row? Second, once a target is sufficiently precise that fresh information must be distributed across endpoints, how should that information be allocated on a general comparison graph?

5.1 Exact risk of balanced partial observations

For a precommitted layout, let $B_i = \{t : i \in S_t\}$ and, for $e = (i, j)$, write

$$n_i = h + |B_i|, \qquad k_e = h + |B_i \cap B_j|, \qquad U_e = n_i + n_j - k_e = h + |B_i \cup B_j|. \tag{21}$$

Thus U_e counts distinct positions observing at least one endpoint. A layout is balanced on e when $n_i, n_j \geq U_e/2$, equivalently $|n_i - n_j| \leq k_e$. Let ν_e be the distribution of U_e and define

$$g(\nu_e) = \max_{t>0} \frac{t}{t+1} \mathbb{E}_\nu \frac{1}{U_e + t}. \tag{22}$$

The scalar random-sample-size minimax problem has classical antecedents (He, 1990; Amrhein, 1995). The reduction here connects its information count to shared, possibly incomplete observations of two bounded coordinates with unrestricted dependence. We write $\mathcal{R}_{\mathrm{bd}}(\nu)$ for the risk of the precommitted policy induced by ν.

Theorem 4. *For a precommitted lottery balanced on every requested edge,*

$$\mathcal{R}_{\mathrm{bd}}(\nu) = \max_{e \in F} g(\nu_e). \tag{23}$$

For arbitrary precommitted lotteries, including unbalanced ones, the right side remains a lower bound. If $U_e = N$ deterministically, its scalar value is $m(N)$; for nonconstant U_e, $m(\mathbb{E}U_e) < g(\nu_e) < \mathbb{E}m(U_e)$.

For attainment, fix a balanced edge and put $k = k_e$, $U = U_e$, $a_i = n_i - k$, $a_j = n_j - k$, and $\omega_e = (U/2 - a_i)/k \in [0, 1]$. Over all cache and fresh positions observing either endpoint, define

$$X_t^\circ = \begin{cases} Y_{t,i}, & i \text{ only}, \\ 1 - Y_{t,j}, & j \text{ only}, \\ \omega_e Y_{t,i} + (1 - \omega_e)(1 - Y_{t,j}), & \text{both}, \end{cases} \qquad Z_e = \frac{2}{U} \sum_t X_t^\circ - 1. \tag{24}$$

The weights give $\mathbb{E}(Z_e \mid S) = \theta_e$ and $\mathrm{Var}(Z_e \mid S) \leq (1 - \theta_e^2)/U_e$. Distinct positions are independent; the transformed observations need not be identically distributed. The estimator and its common lottery-calibrated shrinkage parameter are

$$\widehat{\theta}_e = \frac{U_e}{U_e + t_e} Z_e, \qquad \mathbb{E}_\nu \frac{U_e - t_e^2}{(U_e + t_e)^2} = 0. \tag{25}$$

The root is unique and positive. For necessity, take one fixed edge with $Y_i = X$, $Y_j = 1 - X$, $X \sim \mathrm{Bernoulli}(p)$ and all other coordinates constant. Each union position reveals one Bernoulli observation. A single $\mathrm{Beta}(t/2, t/2)$ prior chosen before the lottery yields $g(\nu_e)$, including for estimators using the entire observed table. The certificate depends on the layout and the public score range. Treating $\mathbb{E}U_e$ as a deterministic sample size is optimistic, while averaging fixed-layout minimax values is conservative; $g(\nu_e)$ is the exact lottery risk between these two summaries. The reduction provides a direct certificate for randomized partial-observation plans. Population-specific error may benefit from covariance and overlap structure. Appendix E proves the reduction.

5.2 Request complexity and high-precision work

The exact reduction above turns each balanced layout into an edgewise effective sample-size experiment. We now use that reduction to distinguish two resources: the number of independent fresh

positions and the total coordinate work needed to reach a target. The former has an exact finite threshold, while the latter converges to a graph-dependent allocation problem at high precision.

Let $q = (\lceil n_d \rceil - h)_+$ and define

$$A(H,c) = \min_{z_i>0,\ i\in V_+} \left\{ \sum_{i\in V_+} c_i z_i : z_i^{-1} + z_j^{-1} \le 4 \quad (ij \in F) \right\}. \tag{26}$$

Theorem 5. *The minimum adaptive hard request cap is q. On a unit-price star with $q \le h$, $W^\star_{h,M}(d) = q$ and $\mathcal{J}^\star_{\mathrm{ad}}(a,d) = (a+1)q$. For every nonempty finite graph with positive prices,*

$$\mathcal{W}^\star_{\mathrm{ad}}(\infty, d; c) = \frac{A(H,c)}{d}(1 + o(1)), \qquad d \to 0, \quad dh \to 0, \tag{27}$$

where $C_+/4 \le A(H,c) \le C_+/2$. More explicitly, with $A = A(H,c)$ and $C = C_+$, for $0 < d < A/(\pi^2 C)$,

$$\left[\left(\sqrt{A/d} - \pi\sqrt{C} \right)^2 - hC \right]_+ \le \mathcal{W}^\star_{\mathrm{ad}}(\infty, d; c) \le A/d + C. \tag{28}$$

A table of $h + R$ full rows can simulate any policy with hard request cap R. The complementary Bernoulli pair therefore gives the request lower bound $m(h + R)$. For a request-optimal construction at every precision, choose a vertex cover $C_0 \subseteq V_+$ and put $N = \max\{h, \lceil n_d \rceil\}$, $N_{\mathrm{full}} = \max\{h, \lceil N/2 \rceil\}$. Acquire $N_{\mathrm{full}} - h$ full active-coordinate rows and then $N - N_{\mathrm{full}}$ cover rows. Every edge is balanced with union count N, and the resources are

$$R = N - h, \qquad W_c = C_+(N_{\mathrm{full}} - h) + c(C_0)(N - N_{\mathrm{full}}), \tag{29}$$

where $c(C_0) = \sum_{i\in C_0} c_i$. For the low-deficit star, the cover is the reference without full refresh.

The high-precision upper bound instead gives coordinate i exactly $\lceil z_i^\star/d \rceil$ fresh singleton observations on disjoint rows, where $z^\star$ solves (26). Independent sample means then satisfy every edge constraint. The lower bound uses a product Bernoulli submodel and an adaptive van Trees argument (Gill & Levit, 1995). It limits arbitrary estimators. On a unit-price star, $A(H, \mathbf{1}) = (1 + \sqrt{M})^2/4$. At fixed total active price, the bounds on A limit variation in the high-precision work constant across graphs to a factor two. The singleton construction targets work and can use many more requests than (29). Work- and request-optimal schedules can differ substantially. With a request price a, executable designs must therefore be compared through the joint charge $W_c + ar$. Appendix F proves the converse, and Appendices G and J give finite-request packing and joint-price certificates.

6 Experiments

We evaluate the finite-scale resource laws in Sections 3–5. We use them as planning rules for a *locked report*: the library, distribution, requested contrasts, and target are fixed without using evaluation scores, and only the fresh acquisition is chosen later. This covers, for example, prespecified compression variants against a locked reference; selecting a contrast after seeing the report is a different contract. All costs below exclude the sunk cache.

Precision transition. Figure 1 evaluates the finite constructions and lower bounds from Section 3. Panel (a) fixes $h = 192$, $M = 127$, and varies precision; the candidate-refresh threshold is $m(h)/d = 4$. Panel (b) fixes $h = 192$ and uses $d/d_0(h) \in \{1.5, 1, 0.5\}$ as M grows, yielding the predicted bounded, square-root, and linear regimes. Panel (c) verifies the joint limits: along $h = M$, critical work normalized by $h\sqrt{M}$ approaches 2, while for $dh = 1/8$, work normalized by Mh approaches 1. At $h = 192$, $d_0(h) = 1.13269 \times 10^{-3}$ rather than $1/(4h) = 1.30208 \times 10^{-3}$; for $M = 4096$ this displacement is 2.08 transition-window widths $1/(h\sqrt{M})$. Appendix C.6 gives the integer checks.

Exact centering matters. At $h = 192$, $d_0(h) = 1.13269 \times 10^{-3}$ while $1/(4h) = 1.30208 \times 10^{-3}$. For the intermediate target $d = 1.21739 \times 10^{-3}$ and $M = 4096$, the exact threshold places the problem above the transition: reference-only acquisition certifies all contrasts with 2952 fresh evaluations. Using $1/(4h)$ instead would place the same target below the nominal threshold and suggest candidate refreshes. The two centers differ by 2.08 transition-window widths $1/(h\sqrt{M})$, enough to change the acquisition plan.

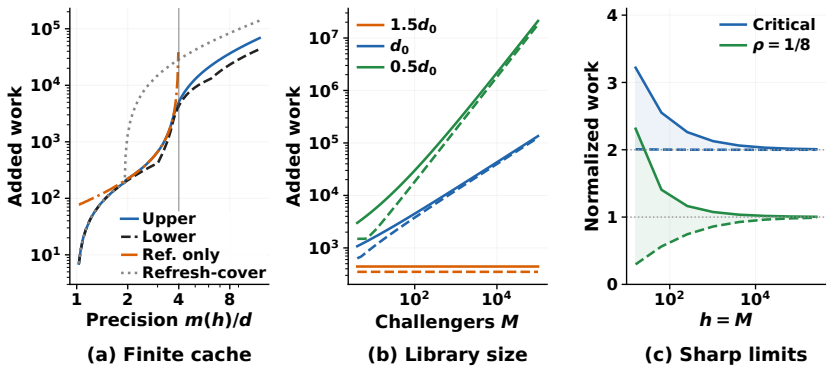

Figure 1: Finite costs across the precision transition.

Table 1: Finite critical (W, R) at $h = 192$.

M	Full-vector	Refresh–cover	Random split	Deterministic split
10	$(6963, 633)$	$(2843, 633)$	$(1573, 946)$	$(1573, 952)$
19	$(12660, 633)$	$(4832, 633)$	$(2083, 1183)$	$(2083, 1183)$

Reference sharing versus randomization. Table 1 uses $M = 10, 19$, matching the nonreference counts of the 11- and 20-model frozen pools. Relative to refresh–cover, the split construction reduces work by 44.7% and 56.9%; these are reference-sharing gains, not randomization gains, because the random and deterministic splits have identical work at these small M/h^2. To expose the contract effect, we also use dense finite stress tests at $h = 192, d = d_0(h)$: for $M = 4h^2$ the marginal split costs 164108 while the global layout-wise optimum is certified in $[192498, 193116]$; for $M = 8h^2$ the values are 232084 and $[339954, 340572]$. Thus the stronger contract costs at least 17.3% and 46.5% more in these two cases. Appendix D.7 gives the searches and wider crossover.

Request–work tradeoff. For the displayed random split and refresh–cover plans, minimizing $W + aR$ selects the split when $a < 1270/313 \simeq 4.06$ for $M = 10$ and $a < 2749/550 \simeq 5.00$ for $M = 19$.

General comparison graphs. The exact-risk frontiers in Appendix P show that, at $h = 3, d = m(4) = 1/9$, a three-edge star needs one evaluation under one fresh request whereas a three-edge matching needs three. On the complete $K = 2048$ workload, work drops from 2047 with one additional position to 1111 with unrestricted requests, separating requests from evaluations.

Population-specific audit. The frozen EuroSAT audit in Appendix O quantifies the conservatism of a range-uniform certificate. For $K_6, h = 192, d = 1/256$, cache-only paired means have actual MSE/$d = 0.3758$ (accuracy) and 0.0449 (Brier), while the uniformly certified bounded cover spends $W = 165$ and has ratios 0.5550 and 0.1771; this audit measures the price of uniformity, not population-wise superiority. RouterBench and LLMRouterBench separately study covariance-informed allocation (Appendix N); their principal target $(1/4096)/d_0(192) \simeq 0.21554$ is below the star threshold, so those gains are population-specific rather than evidence for the critical randomization law. Appendix L studies population-dependent spending under a uniform hard reserve.

7 Conclusion

We study precision refinement from a finite shared evaluation cache and characterize when reuse of a common reference ceases to suffice. The residual bounded-mean risk $d_0(h)$ induces an exact bounded, $\sqrt{M}$, and linear work transition in a reference star, which persists under fully adaptive acquisition and admits sharp critical scaling. Strengthening marginal to layout-wise MSE adds an integrality effect governed by M/h^2. For partial observations on general comparison graphs, edgewise union counts determine exact bounded-score risk, while $A(H, c)$ governs high-precision work and separates evaluation work from request complexity. Together, these results separate three mechanisms: residual uncertainty determines when fresh information must extend beyond the reference, the accuracy contract determines whether fractional randomized refreshes are usable, and graph geometry determines where new information should be allocated. They characterize when a cache remains sufficient and how to spend fresh evaluation once it is not.

AI USE STATEMENT

Generative AI assisted with manuscript editing, consistency checks, and bibliographic formatting. The authors remain responsible for the correctness, originality, attribution, and reproducibility of all scientific content.

REPRODUCIBILITY STATEMENT

Sections 2–5 specify the statistical experiment, acquisition classes, estimators, resource accounting, and theoretical results. Appendices C, D, E, and F contain the proofs of the main theorems. Appendix D.7 gives the finite critical calculations used in Section 6, while Appendices M and P specify the finite-layout and general-graph resource calculations. Appendix L gives the population-adaptive variance-screen construction and its fallback reserve. The frozen-score studies in Appendices N, O, and Q specify the empirical populations, score definitions, sampling schemes, graph workloads, resource ledgers, pilot/estimation splits, and risk calculations; Section N.1 gives the joint covariance-informed estimator and allocation. The supplementary material contains the code, frozen score matrices, model-column identities, preprocessing manifests, instantiated graph inputs, and numerical certificates used to reproduce the reported figures and tables.

REFERENCES

Peter Amrhein. Minimax estimation of proportions under random sample size. *Journal of the American Statistical Association*, 90(431):1107–1111, 1995.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnić. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnić. PPI++: Efficient prediction-powered inference. arXiv preprint arXiv:2311.01453, 2023b.

András Antos, Varun Grover, and Csaba Szepesvári. Active learning in multi-armed bandits. In *Proceedings of the 19th International Conference on Algorithmic Learning Theory (ALT)*, volume 5254 of *Lecture Notes in Computer Science*, pp. 287–302, 2008.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, second edition, 1985.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models. arXiv preprint arXiv:2405.14782, 2024.

Alexandra Carpentier, Alessandro Lazaric, Mohammad Ghavamzadeh, Rémi Munos, and Peter Auer. Upper-confidence-bound algorithms for active learning in multi-armed bandits. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory (ALT)*, volume 6925 of *Lecture Notes in Computer Science*, pp. 189–203, 2011.

Arun Tejasvi Chaganty, Stephen Mussmann, and Percy Liang. The price of debiasing automatic metrics in natural language evaluation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 643–653, 2018.

William G. Cochran. *Sampling Techniques*. John Wiley & Sons, third edition, 1977.

Charlie Cowen-Breen, Alekh Agarwal, Stephen Bates, William W. Cohen, Jacob Eisenstein, Amir Globerson, and Adam Fisch. Multiple-prediction-powered inference. In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026.

Matteo Croci, Karen E. Willcox, and Stephen J. Wright. Multi-output multilevel best linear unbiased estimators via semidefinite programming. *Computer Methods in Applied Mechanics and Engineering*, 413:116130, 2023.

Florian E. Dorner, Vivian Yvonne Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

Gustav Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2):255–262, 1952.

Adam Fisch, Daniel Deutsch, Joshua Maynez, Alekh Agarwal, Jonathan Berant, William Cohen, Amir Globerson, and Jacob Eisenstein. CollabEval: Statistically efficient collaborative model evaluation via matrix completion. arXiv preprint arXiv:2607.05046, 2026.

Riccardo Fogliato, Pratik Patil, Mathew Monfort, and Pietro Perona. A framework for efficient model evaluation through stratification, sampling, and estimation. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, volume 15146 of *Lecture Notes in Computer Science*, pp. 140–158. Springer, 2025.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Annual Conference on Learning Theory (COLT)*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, 2016.

Richard D. Gill and Boris Y. Levit. Applications of the van Trees inequality: A Bayesian Cramér–Rao bound. *Bernoulli*, 1(1–2):59–79, 1995.

Kun He. An ancillarity paradox in the estimation of multinomial probabilities. *Journal of the American Statistical Association*, 85(411):824–828, 1990.

Hayden Helm, Ben Johnson, and Carey Priebe. Query-efficient model evaluation using cached responses. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

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

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

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

Jack Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):272–304, 1959.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763, 2021.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, second edition, 1998.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

George L. Nemhauser and Leslie E. Trotter, Jr. Vertex packings: Structural properties and algorithms. *Mathematical Programming*, 8(1):232–248, 1975.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326, 2024.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006.

Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005.

Samuel Rosa. Optimal designs for treatment comparisons represented by graphs. *AStA Advances in Statistical Analysis*, 102(4):479–503, 2018.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51 of *Proceedings of Machine Learning Research*, pp. 1232–1240, 2016.

El Mehdi Saad, Gilles Blanchard, and Nicolas Verzelen. Covariance-adaptive best arm identification. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 73287–73298, 2023.

Guillaume Sagnol. Computing optimal designs of multiresponse experiments reduces to second-order cone programming. *Journal of Statistical Planning and Inference*, 141(5):1684–1708, 2011.

Daniel Schaden and Elisabeth Ullmann. On multilevel best linear unbiased estimators. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):601–635, 2020.

Abraham Wald. *Statistical Decision Functions*. John Wiley & Sons, 1950.

Tijana Zrnić and Emmanuel J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62993–63010, 2024.

Table 2: Principal notation used in the main text and appendices.

Symbol	Meaning
$H = ([K], F), V_+$	Fixed oriented comparison graph and its nonisolated vertices.
$P, \mathcal{P}_{\mathrm{bd}}, Y, \mu, \theta_e$	Score law, bounded class, score vector, mean vector, and edge contrast.
$h, \mathcal{C}$	Complete cache size and its observed score values.
π, ν, S_t, B_i	Adaptive policy, precommitted layout law, queried subset, and fresh observation positions.
x_i, n_i, k_e, U_e	Fresh and total coordinate counts, shared count, and edge union count.
c_i, C_+, a, r, R, W_c	Coordinate price, total active price, request price, used requests, request cap, and work.
$\mathcal{R}_{\mathrm{marg}}, \mathcal{R}_{\mathrm{bd}}, \mathcal{R}_{\mathrm{lay}}$	Procedure marginal risk, optimal-estimator marginal risk, and layout-wise risk.
$\mathcal{W}^\star_{\mathrm{ad}}, \mathcal{W}^\star, \mathcal{W}^\star_E$	Adaptive hard, precommitted hard, and precommitted expected work.
$\mathcal{W}^\star_{\mathrm{det}}, \mathcal{W}^\star_{\mathrm{lay}}$	Deterministic marginal and layout-wise hard-work optima.
$\mathcal{J}^\star_{\mathrm{ad}}, \mathcal{J}^\star$	Adaptive and precommitted hard joint-charge optima.
$m(n), n_d, q, d_0(h)$	Contrast equalizer risk, inverse target count, integer fresh deficit, and scalar cache risk.
$M, W^\star_{h,M}, W^{\mathrm{lay}}_{h,M}$	Star candidate count and marginal/layout-wise work functions.
s, J_s, r_0	Total candidate refresh budget, one candidate's random count, and independent fresh reference-block size.
$\rho, t, \Psi(t)$	Intermediate precision dh, transition-window coordinate, and limiting work.
$\Delta_h(k), \eta, \Phi(\eta)$	Scalar risk decrement, library–cache scale M/h^2, and integer-refresh limiting work.
$g(\nu_e), t_e$	Random-union minimax functional and lottery-calibrated shrinkage parameter.
$A(H, c), z^\star$	High-precision graph allocation value and optimizing allocation vector.
$m_{\mathrm{p}}, h_{\mathrm{est}}, v_e, \zeta, \lambda, \gamma$	Variance-gate pilot pairs, estimation rows, variance, cutoff, exponential parameter, and fallback risk fraction.
$\Sigma, G, \mathcal{S}, h_2, n_S, M_G, \ell_e$	Population covariance, planning covariance, query menu, estimation-cache size, subset counts, information matrix, and contrast vector.
$\tau_c, \tau_{f,c}, \ell, L$	Integral/fractional cover costs, real effective deficit, and integer deficit.

Appendix

A Main-text roadmap and notation

Tables 2 and 3 summarize notation used throughout the paper and the scope of the principal results. Appendix B gives an extended comparison with classical estimation and modern evaluation allocation. Appendices C–F contain the proofs for Sections 3–5. The later appendices collect finite-resource constructions, population-adaptive extensions, and frozen-population diagnostics used in Section 6.

A concrete operating protocol. Fix the model configurations, scoring rule, and evaluation distribution without using the evaluation scores. Collect the complete h-row table. An externally specified report then requests a fixed set of mean differences at target MSE d, for example all locked compression variants against the original reference. The acquisition policy can inspect the cache, choose which coordinates to evaluate on new independent examples, and stop within its hard reserve. The service returns every declared estimate. Its guarantee concerns each entry's repeated-sampling squared error, including the initial sampling stage. Cache-conditional accuracy and selection from the reported values define separate losses. This is the statistical experiment analyzed in Sections 2–5. The frozen-score studies later in the appendix evaluate fixed empirical populations under the same resource ledger.

What the cache and the workload contribute. Completeness gives every candidate the same initial sample count, while row alignment enables the balanced partial-observation transformation. Reference sharing uses the common target mean and requires no predictive relation between the reference and candidate scores. The star results isolate the residual candidate uncertainty that limits this reuse; the graph results characterize which fresh observations can serve multiple comparisons. All work laws count additional acquisition for an already cached model library. The workload is fixed independently of score values, whereas the acquisition policy may adapt to observed data in the main lower bounds.

Table 3: Information and accuracy scope of the principal results.

Result	Acquisition and information	Accuracy and scope
Theorems 1–2	Adaptive lower bounds; precommitted upper constructions; score range supplied.	Star marginal work, finite-cache orders, and sharp joint limits.
Theorem 3	Precommitted; each supported layout satisfies the target.	Layout-wise critical work and its M/h^2 crossover, averaging score randomness.
Theorem 4	Precommitted; balanced partial layouts.	Exact marginal MSE using the union-count distribution.
Theorem 5	Arbitrary adaptive policies and estimators.	Exact request threshold and high-precision work.
Equation (42)	Deterministic full active vectors.	A sufficient simultaneous error bound.
Theorem 14	Independent pilot and estimation/fallback data.	Uniform marginal risk and population-dependent spending.
Appendix N	Supplied or independently estimated covariance.	Frozen-population risk and resources of specified comparators.

The layout-wise contract conditions on score-independent acquisition. A layout selected from observed scores has a selection-dependent conditional score law; a contrast selected from the reported values likewise induces a different loss from the independent-demand criterion in (3).

B Related Work

Classical estimation ingredients and acquisition across a library. Bounded-mean shrinkage, beta equalizer priors, and minimax decision theory provide the scalar estimation ingredients used here (Wald, 1950; Berger, 1985; Lehmann & Casella, 1998). Our starting point is the residual finite-sample minimax risk they imply. The new resource question is how that residual risk propagates across a library after a complete cache has already been collected. Candidate-wise lower bounds with different difficult populations are insufficient for adaptive acquisition because the policy may distribute its counts unevenly. A single product prior, combined with predictable posterior-variance reductions, charges all candidate counts inside the same adaptive experiment and yields the dimension transition and its sharp joint limits.

Random sample size introduces a related conditioning issue (He, 1990; Amrhein, 1995): minimizing fixed-sample risk for each realized sample size and then averaging can differ from the minimax randomized experiment. Theorem 4 uses this principle after a balanced transformation of shared, partially observed bounded coordinates with arbitrary dependence. The resulting contribution is an exact reduction to the edgewise union-count distribution, together with a lower bound that permits estimators to use the entire observed table.

Measurement allocation and covariance information. Optimal design allocates observations to specified contrasts (Elfving, 1952; Kiefer, 1959; Pukelsheim, 2006), including graph and multiresponse formulations (Sagnol, 2011; Rosa, 2018). Multilevel estimation similarly optimizes partially observed sample groups (Schaden & Ullmann, 2020; Croci et al., 2023). Model-assisted evaluation (Chaganty et al., 2018) and prediction-powered inference (Angelopoulos et al., 2023a;b) exploit auxiliary predictions to improve estimation. These lines of work establish the broader allocation and information-reuse principles on which our constructions draw.

MultiPPI jointly chooses subset counts and estimator weights under hard budgets, proves known-covariance minimax optimality beyond linear unbiased competitors, and analyzes estimated covariance and reuse of initial observations (Cowen-Breen et al., 2026). Its scope already includes fixed contrasts, initial data, and inverse-information allocation. Our statistical experiment instead asks for a resource law under a range-uniform requirement while the finite cache remains first-order. Supplied-covariance minimaxity and range-uniform minimaxity with unknown dependence therefore describe different information classes. The reciprocal-count program reappears in our high-precision

regime; the adaptive converse identifies when it is asymptotically optimal, while the finite-cache threshold identifies when the cache changes that allocation law.

Efficient evaluation and complete initial caches. Active mean estimation (Antos et al., 2008; Carpentier et al., 2011), active testing (Kossen et al., 2021), and model evaluation (Fogliato et al., 2025) study selective measurement, while compact benchmarks choose representative examples (Polo et al., 2024). Query-efficient evaluation (Helm et al., 2026) and CollabEval (Fisch et al., 2026) exploit cross-model information in existing score tables. We focus on precision refinement for an already evaluated, fixed model library under an unchanged population. In this setting, a shared reference can be reused through its common target mean even without a predictive relation to candidate scores; Theorem 4 separately captures the value of aligned sample positions. Evaluation lower bounds such as Dorner et al. (2025) likewise make savings conditional on explicit information assumptions.

Accuracy requirements and randomization. Our primary loss is marginal squared error for a fixed collection of mean differences. Best-arm identification (Kaufmann et al., 2016; Garivier & Kaufmann, 2016), including covariance-adaptive methods (Saad et al., 2023), controls a selection objective. Active statistical inference (Zrnić & Candès, 2024), variance-sensitive bounds (Maurer & Pontil, 2009), time-uniform inference (Howard et al., 2021), simultaneous inference (Romano & Wolf, 2005), and adaptive reuse (Dwork et al., 2015; Russo & Zou, 2016) impose other forms of validity. Section 4 isolates one specific change of contract within the same precommitted acquisition class: marginal risk averages over layout randomization, whereas layout-wise risk certifies every supported allocation. The scale M/h^2 quantifies the resulting integer-refresh cost. The supplementary variance gate studies a separate axis—population-adaptive spending while preserving a uniform hard reserve.

C Proofs of the finite-cache precision transition

Put

$$m(u) = (\sqrt{u}+1)^{-2}, \quad d_0(h) = m(h)/4, \quad a_h = \sqrt{h},$$
$$s_h = h + \sqrt{h}, \quad b_h = s_h + 1, \quad \beta_h = s_h + \tfrac{1}{2}.$$

Here $h \geq 1$ is an integer and work counts additional observations. Policies have finite hard request caps, choose each subset before observing its fresh row, and otherwise may read all past values, randomize, stop, and use arbitrary biased estimators. Risk is fixed-edge marginal MSE over all bounded populations. Every upper construction below is precommitted; hence the adaptive lower bounds and the matching joint limits also apply to the precommitted marginal optimum.

C.1 One prior and the information gained by an observation

Give every active coordinate an independent $\mathrm{Beta}(a_h/2, a_h/2)$ prior for its Bernoulli mean. Conditional on these means, all coordinates and rows are independent. This is one prior over fixed admissible binary populations. Let $\mathbb{E}_\Pi$ include this prior, data, and policy randomness, x_i be the fresh count, and V_i the final posterior variance.

The posterior remains a product distribution after an adaptive transcript: its likelihood factors into parameter-free decision kernels and one Bernoulli likelihood for each coordinate. Actions and stopping decisions add no parameter-dependent factor after their observed histories have been included. Thus the Bayes risk for a contrast is at least $\mathbb{E}_\Pi V_i + \mathbb{E}_\Pi V_j$. Uniform contrast risk at most d implies

$$\mathbb{E}_\Pi V_i + \mathbb{E}_\Pi V_j \leq d \quad (ij \in F), \qquad \mathbb{E}_\Pi V_i \leq d \quad (i \in V_+). \tag{30}$$

Before fresh acquisition, the expected posterior variance equals

$$\mathbb{E}_\Pi V_i^{\mathrm{cache}} = \frac{a_h}{4(a_h+1)(h+a_h)} = d_0(h).$$

A $\mathrm{Beta}(A, B)$ posterior, with total mass $s = A + B$ and mean $p = A/s$, has variance $V = p(1-p)/(s+1)$. One new Bernoulli observation gives

$$\mathbb{E}[V_{\mathrm{next}} \mid \text{history}] = \frac{s}{s+1} V, \qquad V - \mathbb{E}[V_{\mathrm{next}} \mid \text{history}] = \frac{p(1-p)}{(s+1)^2} \leq \frac{1}{4(s+1)^2}.$$

Other coordinates' observations do not change this posterior. Telescoping predictable decreases, padding the bounded stopping time with empty steps, yields

$$d_0(h) - \mathbb{E}_\Pi V_i \le \mathbb{E}_\Pi G_h(x_i), \qquad G_h(n) = \frac{1}{4}\sum_{j=0}^{n-1}(b_h + j)^{-2}. \tag{31}$$

Extend G_h linearly between integers. It is increasing and concave.

Proposition 1 (A finite cost for every active candidate)**.** *For $0 < d < d_0(h)$, put $\Delta = d_0(h) - d$. Every uniformly feasible adaptive policy satisfies, under the same prior, $\mathbb{E}_\Pi x_i \ge G_h^{-1}(\Delta)$ for all $i \in V_+$. In particular, defining*

$$B_h(d) = \max\left\{4b_h^2\Delta,\ \frac{4\beta_h^2\Delta}{1 - 4\beta_h\Delta}\right\}, \tag{32}$$

its hard work cap is at least $C_+ B_h(d)$. The count bounds hold under the common prior and therefore imply the stated hard-cap bound.

Proof. Concavity in (31) gives $\Delta \le G_h(\mathbb{E}_\Pi x_i)$. The inverse exists since $G_h(\infty) \ge 1/(4b_h) > d_0(h)$. The bound $G_h(n) \le n/(4b_h^2)$ gives the first explicit term. For an integer n, convexity of u^{-2} on unit intervals gives

$$\sum_{j=0}^{n-1}(b_h + j)^{-2} \le \int_{\beta_h}^{\beta_h + n} u^{-2}\, du = \frac{n}{\beta_h(\beta_h + n)}.$$

The last expression is concave in n; therefore, with $\bar{x}_i = \mathbb{E}_\Pi x_i$, $\Delta \le \bar{x}_i/[4\beta_h(\beta_h + \bar{x}_i)]$. Inversion gives the rational term. Its denominator is positive: $4\beta_h d_0(h) = (h + \sqrt{h} + 1/2)/(\sqrt{h} + 1)^2 < 1$. Sum with the positive active candidate prices. The hard cap dominates this common-prior expected work. □

If $h \to \infty$ and $dh \to \rho \in (0, 1/4)$, then

$$B_h(d)/h \longrightarrow (4\rho)^{-1} - 1. \tag{33}$$

The rational term has this limit; the linear term has the smaller limit $1 - 4\rho$.

C.2 Reference-only acquisition and its exact risk floor

Consider a unit-price star with reference 0 and M leaves. For each leaf, use its h cached observations to set

$$\widehat{\mu}_i = \tfrac{1}{2} + \frac{\sqrt{h}}{\sqrt{h} + 1}(\overline{Y}_i^{\,\mathrm{cache}} - \tfrac{1}{2}).$$

Its MSE is at most $d_0(h)$ uniformly over bounded scores. Estimate the reference by the unbiased mean of $r \ge 1$ independent fresh observations. Its error is independent of every leaf estimate and has mean zero, so

$$\sup_P \max_i \mathbb{E}_P\left[\left((\widehat{\mu}_0 - \widehat{\mu}_i) - (\mu_0 - \mu_i)\right)^2\right] \le d_0(h) + \frac{1}{4r}. \tag{34}$$

All contrasts come from one bounded vector of estimated means. For $d > d_0(h)$,

$$\mathcal{W}^\star_{\mathrm{ad}}(\infty, d) \le \left\lceil [4(d - d_0(h))]^{-1} \right\rceil,$$

independently of M. Cache-only acquisition suffices if $d \ge m(h)$.

Let $\mathcal{R}_{\mathrm{ref}}(h, r)$ be the minimax maximum-edge risk when only the reference receives fresh observations, with hard reference count at most r, allowing adaptive stopping. For every finite r,

$$\mathcal{R}_{\mathrm{ref}}(h, r) \ge \max\{m(h + r),\ d_0(h) + m(h + r)/4\} > d_0(h). \tag{35}$$

The first term is the unrestricted adaptive request bound. For the second, use independent Bernoulli reference and leaf means with symmetric beta equalizer priors for sample sizes $h + r$ and h. Providing all $h + r$ reference observations simulates any allowed transcript. In that more informative experiment,

the posterior contrast variance is the sum of scalar posterior variances, and its Bayes risk is the displayed sum. Other coordinates can be independent nuisance parameters. Consequently

$$\inf_{r<\infty} \mathcal{R}_{\mathrm{ref}}(h,r) = d_0(h), \qquad \mathcal{R}_{\mathrm{ref}}(h,r) = d_0(h) + \frac{1}{4r} + o(r^{-1}) \quad (r \to \infty,\ h \text{ fixed}).$$

The floor is not attained at any finite hard reference budget. For $0 \le r \le h$, balance gives the exact identity $\mathcal{R}_{\mathrm{ref}}(h,r) = m(h+r)$.

C.3 A SHARP INTERMEDIATE LAW

Theorem 6 (A nonnegligible cache at intermediate precision)**.** *For unit-price stars with M leaves, if $h, M \to \infty$ and $dh \to \rho \in (0, 1/4)$, then*

$$\mathcal{W}^{\star}_{\mathrm{ad}}(\infty, d)/(Mh) \longrightarrow (4\rho)^{-1} - 1.$$

There is no restriction on the relative growth rates of h and M.

Proof. Use Proposition 1 with $C_+ = M+1$. For attainment, let $r = \lceil h\sqrt{M} \rceil$ fresh positions observe only the reference, put $d' = d - 1/(4r)$, and choose

$$\ell = \left\lfloor \left\lceil \left(\frac{1}{2\sqrt{d'}} - 1 \right)^2 \right\rceil - h \right\rfloor_+ .$$

Eventually $0 < d' < 1/4$. Each leaf receives ℓ new observations on positions disjoint from the reference's new positions; leaves may share their positions. Scalar shrinkage on its $h + \ell$ observations has leaf risk $d_0(h+\ell)$. Use the reference's unbiased fresh mean. The total edge risk is at most $d_0(h+\ell) + 1/(4r) \le d$. Work is $r + M\ell$. Since $hd' \to \rho$, $\ell/h \to (4\rho)^{-1} - 1$ and $r/(Mh) \to 0$. Integer rounding divided by h vanishes. □

For $dh \to \rho > 1/4$, reference-only work is $O(h)$. The threshold therefore occurs while the cache remains a first-order part of the target's information requirement.

C.4 CRITICAL PRECISION

At $d = d_0(h)$ define

$$\kappa_h = \mathbb{E}_{p \sim \mathrm{Beta}(a_h/2, a_h/2)} \sqrt{p(1-p)} = \frac{\mathrm{B}((a_h+1)/2, (a_h+1)/2)}{\mathrm{B}(a_h/2, a_h/2)}.$$

As $h \to \infty$, $\kappa_h \to 1/2$.

Proposition 2 (Finite critical bounds)**.** *For a unit-price star,*

$$\mathcal{W}^{\star}_{\mathrm{ad}}(\infty, d_0(h)) \ge [4b_h\kappa_h\sqrt{M} - s_h]_+. \tag{36}$$

For an integer $s \ge 1$, write $k = \lfloor s/M \rfloor$, $p = s/M - k$, and

$$D_h(s/M) = m(h) - \{(1-p)m(h+k) + p\, m(h+k+1)\}.$$

Then

$$\mathcal{W}^{\star}_{\mathrm{ad}}(\infty, d_0(h)) \le \min_{s \in \mathbb{Z}_{\ge 1}} \{ s + \lceil D_h(s/M)^{-1} \rceil \}. \tag{37}$$

Proof. Use the same product beta prior. The reference posterior obeys

$$(s_h + x_0)V_0 = \mathbb{E}[p_0(1-p_0) \mid \text{transcript}].$$

Conditional Jensen and Cauchy–Schwarz imply

$$\kappa_h \le \mathbb{E}_\Pi \sqrt{(s_h + x_0)V_0} \le \sqrt{(s_h + \mathbb{E}_\Pi x_0)\mathbb{E}_\Pi V_0}.$$

Thus $\mathbb{E}_\Pi V_0 \geq \kappa_h^2/(s_h + \mathbb{E}_\Pi x_0)$. At the critical target each leaf needs posterior-variance reduction at least $\mathbb{E}_\Pi V_0$. One leaf observation decreases expected variance by at most $\gamma_h = 1/(4b_h^2)$. If B is a hard work cap and $u = \mathbb{E}_\Pi x_0$, summing gives

$$\gamma_h(B-u) \geq M\mathbb{E}_\Pi V_0 \geq \frac{M\kappa_h^2}{s_h+u}.$$

The inequality $(B-u)(s_h+u) \leq (B+s_h)^2/4$ proves the lower bound. All expectations use one common prior.

For the upper bound, give every leaf k new observations, and one more to a uniformly selected subset of $s - kM$ leaves. Hard leaf work is s and every leaf has the same count mixture. Apply scalar minimax shrinkage for the realized leaf sample count, giving marginal leaf risk at most $\{(1-p)m(h+k)+pm(h+k+1)\}/4$. Independently observe the reference on $r = \lceil 1/D_h(s/M)\rceil$ fresh positions and use its unbiased mean. Reference and leaf error cross expectations vanish, so edge risk is at most $d_0(h)$. Hard work is $r+s$ and requests are at most $r+k+1$. All estimates have a common bounded mean-vector representation. □

Theorem 7 (Fixed-cache dimension laws and the critical constant)**.** *For fixed h, as $M \to \infty$,*

$$\mathcal{W}^\star_{\mathrm{ad}}(\infty, d) = \begin{cases} O_{h,d}(1), & d > d_0(h), \\ \Theta_h(\sqrt{M}), & d = d_0(h), \\ \Theta_{h,d}(M), & 0 < d < d_0(h). \end{cases}$$

The first line includes zero work when the cache already suffices. For arbitrary simultaneous $h, M \to \infty$,

$$\mathcal{W}^\star_{\mathrm{ad}}(\infty, d_0(h))/(h\sqrt{M}) \longrightarrow 2. \tag{38}$$

Proof. Above the floor, use reference-only acquisition. Below it, Proposition 1 gives a linear lower bound, and complete-vector completion gives a linear upper bound at fixed h, d. At the floor the critical lower bound is $\Omega_h(\sqrt{M})$. For fixed h, choose s of order $\sqrt{M}$ in the finite upper bound; eventually $s < M$ and $D_h(s/M) = (s/M)[m(h) - m(h+1)]$, giving $O_h(\sqrt{M})$.

For the simultaneous limit, the critical lower bound divided by $h\sqrt{M}$ tends to 2, since $b_h/h \to 1$, $\kappa_h \to 1/2$, and $s_h/(h\sqrt{M}) \to 0$. For an explicit upper bound put

$$T_h = (1+h^{-1/2})^{3/2}, \qquad c_{h,M} = \frac{T_h(1+1/h)}{1 - T_h/\sqrt{M}} \quad (\sqrt{M} > T_h)$$

and choose $r = s = \lceil c_{h,M} h\sqrt{M}\rceil$. Let $v = \lceil s/M\rceil$ and $u = h + v$. The function $-m'(z) = 1/[\sqrt{z}(\sqrt{z}+1)^3]$ is decreasing, so

$$D_h(s/M) \geq \frac{s/M}{\sqrt{u}(\sqrt{u}+1)^3}.$$

Write $c = s/(h\sqrt{M}) \geq c_{h,M}$. Both ceiling operations are absorbed by

$$\frac{u}{h} \leq 1 + \frac{c}{\sqrt{M}} + \frac{1}{h}, \qquad \frac{\sqrt{\sqrt{u}(\sqrt{u}+1)^3}}{h} = \frac{u}{h}(1+u^{-1/2})^{3/2} \leq T_h(1 + c/\sqrt{M} + 1/h) \leq c.$$

The last inequality follows from the definition of $c_{h,M}$; its coefficient $1 - T_h/\sqrt{M}$ is positive, so increasing s to the next integer preserves it. Hence $s^2 \geq M\sqrt{u}(\sqrt{u}+1)^3$ and $D_h(s/M) \geq 1/s = 1/r$, proving feasibility. Finally $c_{h,M} \to 1$ and $2s/(h\sqrt{M}) \to 2$. □

C.5 The transition window and its exact center

The same prior yields a finite lower bound across the threshold. Let $\delta = d - d_0(h)$ and $\gamma_h = 1/(4b_h^2)$. For every $d > 0$,

$$\mathcal{W}^\star_{\mathrm{ad}}(\infty, d) \geq \inf_{u \geq 0} \left\{ u + \frac{M}{\gamma_h}\left[\frac{\kappa_h^2}{s_h+u} - \delta\right]_+ \right\}. \tag{39}$$

Indeed, each leaf needs posterior-variance reduction at least $\mathbb{E}_\Pi V_0 - \delta$, while the sum of these reductions is at most $\gamma_h \sum_{i=1}^{M} \mathbb{E}_\Pi x_i$. Insert $\mathbb{E}_\Pi V_0 \geq \kappa_h^2/(s_h + u)$ with $u = \mathbb{E}_\Pi x_0$, sum the counts, and minimize over u. The displayed one-dimensional infimum is attained at

$$u_\star = \begin{cases} [\,\sqrt{M\kappa_h^2/\gamma_h} - s_h\,]_+, & \delta \leq 0, \\ \min\{[\,\sqrt{M\kappa_h^2/\gamma_h} - s_h\,]_+, [\,\kappa_h^2/\delta - s_h\,]_+\}, & \delta > 0. \end{cases}$$

This follows by differentiating the active branch, with the second case truncated at the point where the positive part becomes zero.

Corollary 1 (The dimension transition window)**.** *Let $h, M \to \infty$ at arbitrary relative rates, and suppose $h\sqrt{M}\{d - d_0(h)\} \to t \in \mathbb{R}$. Then*

$$\frac{\mathcal{W}_{\mathrm{ad}}^\star(\infty, d)}{h\sqrt{M}} \longrightarrow \Psi(t) = \begin{cases} 2 - 4t, & t \leq 1/4, \\ (4t)^{-1}, & t \geq 1/4. \end{cases} \tag{40}$$

Proof. Put $\Lambda = h\sqrt{M}$ and $t_{h,M} = \Lambda\{d - d_0(h)\}$. After substituting $u = z\Lambda$, the objective in (39), divided by Λ, is

$$f_{h,M}(z) = z + 4(b_h/h)^2 \left[\frac{\kappa_h^2}{s_h/\Lambda + z} - t_{h,M}\right]_+.$$

Here $b_h/h \to 1$, $\kappa_h \to 1/2$, and $s_h/\Lambda \to 0$. The minima remain in a bounded set because $f_{h,M}(z) \geq z$ and $f_{h,M}(1)$ stays bounded, and remain away from zero because the reciprocal term diverges there while $t_{h,M}$ stays bounded. Uniform convergence on the resulting compact interval gives

$$\liminf \frac{\mathcal{W}_{\mathrm{ad}}^\star}{\Lambda} \geq \inf_{z>0}\{z + [z^{-1} - 4t]_+\} = \Psi(t).$$

For $t < 1/4$, put $v = 1 - 4t > 0$ and choose an integer hard leaf budget $s = \lceil v\Lambda \rceil$, allocated by the balanced count lottery above. Since every leaf's extra count divided by h is at most $s/(Mh) + 1/h = O(M^{-1/2} + h^{-1})$, the derivative formula for m gives

$$\Lambda D_h(s/M) \longrightarrow v.$$

For completeness, $D_h(s/M)$ lies between $(s/M)(-m'(h))$ and $(s/M)(-m'(h + \lceil s/M \rceil))$, and both endpoints, after multiplication by Λ, tend to v. Set

$$r = \left\lceil [D_h(s/M) + 4\{d - d_0(h)\}]^{-1} \right\rceil.$$

The denominator is eventually positive, its product with Λ tends to $v + 4t = 1$, and hence $r/\Lambda \to 1$. The independent fresh reference mean and the leaf shrinkage estimates have edge risk at most d by construction. Thus $(r + s)/\Lambda \to 1 + v = 2 - 4t$. For $t \geq 1/4$, $\delta > 0$ eventually and reference-only acquisition with $r = \lceil 1/(4\delta) \rceil$ gives $r/\Lambda \to (4t)^{-1}$. This also covers the joining point $t = 1/4$. □

The exact finite-cache center expands as

$$d_0(h) = \frac{1}{4h} - \frac{1}{2h^{3/2}} + \frac{3}{4h^2} + O(h^{-5/2}).$$

Its leading correction divided by the window width $1/(h\sqrt{M})$ has magnitude asymptotic to $\frac{1}{2}\sqrt{M/h}$. Consequently the exact centering determines the transition when the challenger count grows faster than the cache size.

These laws concern a fixed score-independent graph, ordinary marginal MSE, a sunk full cache, and hard additional work with unrestricted finite requests. The lower bounds allow full adaptivity, while attaining constructions are explicit and nonadaptive. The critical lower bound uses independent binary scores.

C.6 Threshold figure and exact posterior checks

Figure 1 evaluates the finite constructions and lower bounds above. All costs count fresh coordinate evaluations, including the reference. Panel (a) uses $h = 192$, $M = 127$, and varies $m(h)/d$; panel (b) fixes $h = 192$ and uses $d/d_0(h) \in \{1.5, 1, 0.5\}$; panel (c) sets $h = M$ and normalizes the critical curves by $h\sqrt{M}$ and the $\rho = 1/8$ curves by Mh.

For the finite upper curves, an integer leaf budget s is split between counts $\lfloor s/M \rfloor$ and $\lceil s/M \rceil$, with a uniformly randomized remainder. With J_s denoting a fixed leaf's resulting count, put

$$Z_s = 4d - \mathbb{E}m(h + J_s).$$

When $Z_s > 0$, an independent reference block of $r = \lceil Z_s^{-1} \rceil$ observations gives feasible work $r + s$. The upper envelope also includes balanced reference-only completion when $q \le h$ and zero acquisition when the cache already suffices. The lower envelope is the maximum of the exact request lower bound q, the common-prior count bound $(M + 1)B_h(d)$ for $d < d_0(h)$, and (39). The reference-only curve uses $\lceil (4d - m(h))^{-1} \rceil$ on its feasible domain; refresh–cover uses $q + M \max\{0, \lceil (q - h)/2 \rceil\}$.

Exact arithmetic and exhaustive integer checks verify the plotted risk formulas, critical constructions, unequal-growth cases, and transition-window points. As a separate adaptive check, three-coordinate binary experiments for $(h, R) = (1, 5)$ and $(4, 3)$ are integrated over every cache and subsequent outcome. Their integrated squared contrast errors equal the sums of posterior variances, and the predictable variance decrements telescope according to (31).

D Proofs for the accuracy-contract comparison

We prove the layout-wise star result in Section 4, then separate it from simultaneous and cache-conditional accuracy. All layout-wise claims concern acquisition chosen independently of the score values.

D.1 Finite critical bounds and the layout-wise dimension law

Use $s_h = h + \sqrt{h}$, $b_h = s_h + 1$, and $\kappa_h = \mathbb{E}\sqrt{p(1-p)}$ for $p \sim \mathrm{Beta}(\sqrt{h}/2, \sqrt{h}/2)$, as in Appendix C. At the marginal critical target,

$$\max\{M, [4b_h\kappa_h\sqrt{M} - s_h]_+\} \le W^{\mathrm{lay}}_{h,M}(d_0(h)) \le \inf_{k \in \mathbb{Z}_{\ge 1}} \left\{ Mk + \left\lceil \frac{1}{m(h) - m(h+k)} \right\rceil \right\}. \quad (41)$$

Critical order and the regime $M = o(h^2)$ in Theorem 3. For a finite request cap, the possible subset lists form a finite set. A lottery meets the layout-wise risk bound only if each of its supported layouts admits an estimator meeting the bound. Selecting any such layout preserves its hard work and request caps. Conversely, a deterministic layout is a point-mass lottery. Taking infima proves (14), also under unrestricted finite request caps. This argument allows a different estimator on each observed layout.

Fix a deterministic layout and suppose leaf i has $x_i = 0$. Consider independent Bernoulli scores for coordinates i and 0, with independent symmetric beta priors of total masses $\sqrt{h}$ and $\sqrt{h + x_0}$, respectively. Give every other coordinate a fixed constant score. The entire observed table then contains exactly h independent Bernoulli observations for leaf i and $h + x_0$ for the reference. Independence of the coordinates makes shared request positions immaterial for this subfamily. The two equalizer priors yield scalar Bayes risks $d_0(h)$ and $d_0(h + x_0)$; the posterior variance of their difference is their sum. Every estimator therefore has worst-case contrast MSE at least $d_0(h) + d_0(h + x_0) > d_0(h)$. Every feasible layout must observe each leaf at least once, so its work is at least M. The adaptive marginal lower bound (36) applies to this subclass as well, proving the left side of (41).

For the upper bound, give every leaf $k \ge 1$ fresh observations, and acquire $r = \lceil [m(h) - m(h + k)]^{-1} \rceil$ reference-only observations on a disjoint set of fresh positions. Leaf observations may share positions with one another. Estimate each leaf with the scalar shrinkage rule based on its $h + k$ observations; estimate the reference with its unbiased fresh mean, discarding its cached values for

this estimator. The errors of these two estimates are independent and the reference error has mean zero. Hence each contrast has MSE at most

$$d_0(h+k) + \frac{1}{4r} \le d_0(h).$$

The deterministic work is $Mk + r$, proving the upper bound.

For fixed h, the choice $k = 1$ gives $M + O_h(1)$, matching the lower bound M in order. For a joint bound, let $k = \max\{1, \lceil h/\sqrt{M} \rceil\}$. Since

$$-m'(u) = \frac{1}{\sqrt{u}(\sqrt{u}+1)^3},$$

monotonicity of this derivative's magnitude and $h \ge 1$ give

$$m(h) - m(h+k) \ge \frac{k}{\sqrt{h+k}(\sqrt{h+k}+1)^3} \ge \frac{k}{8(h+k)^2}.$$

Here $k \le 2h$, $Mk \le M + h\sqrt{M}$, and $h^2/k \le h\sqrt{M}$. It follows that $r \le 72h\sqrt{M} + 1$, giving a universal upper multiple of $M + h\sqrt{M}$. For a lower bound, $\sqrt{p(1-p)} \ge 2p(1-p)$ implies

$$\kappa_h \ge \frac{\sqrt{h}}{2(\sqrt{h}+1)} \ge \frac{1}{4}.$$

Thus $4b_h\kappa_h\sqrt{M} - s_h \ge h\sqrt{M} - 2h$. For $M \ge 16$, the lower bound is at least $h\sqrt{M}/2$ as well as M, which proves the joint Θ statement.

Finally, suppose $h, M \to \infty$ and $M = o(h^2)$. Then $k = \lceil h/\sqrt{M} \rceil \sim h/\sqrt{M}$ and $k/h \to 0$. The derivative formula gives $m(h) - m(h+k) \sim k/h^2$. Therefore $Mk \sim h\sqrt{M}$ and $r \sim h\sqrt{M}$. The lower bound, using $b_h/h \to 1$, $\kappa_h \to 1/2$, and $s_h/(h\sqrt{M}) \to 0$, has the same constant 2. □

For fixed $h, d > d_0(h)$, the deterministic reference-only construction already has $O_{h,d}(1)$ work. For fixed $0 < d < d_0(h)$, the adaptive marginal lower bound is $\Omega_{h,d}(M)$, while a deterministic complete-vector acquisition with sufficiently many additional rows costs $O_{h,d}(M)$. Together with Theorem 3, these observations give the fixed-cache layout-wise dimension law: bounded work above $d_0(h)$ and linear work at and below it, over unrestricted deterministic layouts.

D.2 A FINITE SQUEEZE AND THE INTEGER-REFRESH CROSSOVER

We prove (19) and complete Theorem 3. All acquisition in this subsection is precommitted, and all score risks average over the numerical cache.

Fix a deterministic feasible layout. Let $k = \min_{1 \le i \le M} x_i$ and choose a candidate attaining this count. Give this candidate and the reference independent Bernoulli means, with symmetric beta priors of total masses $\sqrt{h+k}$ and $\sqrt{h+x_0}$, respectively. Make all other scores constant. This is a prior over admissible fixed populations. Since the layout is fixed, its likelihood contains exactly $h+k$ observations of that candidate and $h+x_0$ of the reference. Even co-queried coordinates are independent in this subfamily. The Bayes squared error of their difference is

$$\frac{1}{4}m(h+k) + \frac{1}{4}m(h+x_0).$$

The priors may depend on the fixed layout: the layout-wise requirement must hold for every population separately on that layout. They never depend on its realized scores. If $k = 0$, the displayed lower bound exceeds $d_0(h)$ for every finite x_0. Therefore $k \ge 1$ and

$$m(h+x_0) \le \Delta_h(k) := m(h) - m(h+k).$$

Since m is decreasing and $h + x_0$ is integer,

$$x_0 \ge \lceil n_{\Delta_h(k)} \rceil - h, \qquad W \ge Mk + x_0.$$

Taking the minimum over positive integer k proves the lower side of (19). Every supported layout obeys this argument, and (14) transfers it to the layout-wise optimum. For the upper side, give

all candidates exactly k fresh observations and use scalar shrinkage; estimate the reference from a disjoint fresh block of $r_0 = \lceil 1/\Delta_h(k) \rceil$ observations. Then

$$\sup_P \max_i \mathbb{E}_P(\widehat{\theta}_{i0} - \theta_{i0})^2 \le \frac{m(h+k)}{4} + \frac{1}{4r_0} \le d_0(h).$$

Both one-dimensional minima exist because their objectives are at least Mk.

Suppose now that $h, M \to \infty$ and $M/h^2 \to \eta \in (0, \infty)$. For each fixed positive integer k, the derivative

$$-m'(u) = \frac{1}{\sqrt{u}(\sqrt{u}+1)^3}$$

implies $h^2\Delta_h(k) \to k$. Hence

$$\frac{Mk + \lceil \Delta_h(k)^{-1} \rceil}{h^2} \longrightarrow \eta k + 1/k,$$
$$\frac{Mk + \lceil n_{\Delta_h(k)} \rceil - h}{h^2} \longrightarrow \eta k + 1/k.$$

For the second limit use $n_u = u^{-1} - 2u^{-1/2} + 1$; the $u^{-1/2}$ term is $O(h)$ for fixed k, and rounding and the subtraction of h are also negligible after division by h^2.

To justify the minimum, choose $k = 1$ in the upper bound. This gives a bounded normalized feasible cost. The lower-bound objective is at least Mk, and $M/h^2 \ge \eta/2$ eventually. Its minimizing k therefore belongs to a fixed finite set. The same reasoning applies to the upper minimizer. The displayed convergence is uniform on that finite set. Minimizing and squeezing yield

$$\frac{W^{\mathrm{lay}}_{h,M}(d_0(h))}{h^2} \longrightarrow \Phi(\eta) = \min_{k \in \mathbb{Z}_{\ge 1}} \{\eta k + 1/k\}.$$

Adjacent choices k and $k+1$ tie precisely when $\eta = 1/[k(k+1)]$. Thus $k = 1$ is optimal for $\eta \ge 1/2$; for $k \ge 2$, it is optimal when

$$\frac{1}{k(k+1)} \le \eta \le \frac{1}{k(k-1)}.$$

Endpoints allow both adjacent minimizers.

For $h^2 = o(M)$, the lower bound $W^{\mathrm{lay}}_{h,M} \ge M$ and the $k = 1$ construction give

$$1 \le \frac{W^{\mathrm{lay}}_{h,M}(d_0(h))}{M} \le 1 + \frac{\lceil \Delta_h(1)^{-1} \rceil}{M} \longrightarrow 1,$$

since $\Delta_h(1)^{-1} \sim h^2$. This proves (18). The $M = o(h^2)$ leading constant was proved above, completing the theorem.

The marginal precommitted optimum satisfies $\mathcal{W}^\star(\infty, d_0(h)) \sim 2h\sqrt{M}$ by Theorem 2. Division gives (20); in the dense regime it gives the asymptotic ratio $\sqrt{M}/(2h)$. The arithmetic–geometric mean inequality ensures $\Phi(\eta) \ge 2\sqrt{\eta}$, with equality exactly when $1/\sqrt{\eta}$ is a positive integer. This equality is asymptotic; the corresponding finite minimax costs may differ.

D.3 Expected maximum loss and simultaneous error

For $\varepsilon > 0$ and $0 < \alpha < 1$, collecting the following number of complete active-coordinate vectors suffices:

$$q_{\mathrm{sim}} = \left(\left\lceil \frac{2}{\varepsilon^2} \log \frac{2|F|}{\alpha} \right\rceil - h \right)_+ \quad \Longrightarrow \quad \sup_P \Pr\Big(\max_e |\widehat{\theta}_e - \theta_e| > \varepsilon \Big) \le \alpha. \tag{42}$$

The estimator averages the $h + q_{\mathrm{sim}}$ paired contrasts, and hard fresh work is $q_{\mathrm{sim}} C_+$. The argument below uses Hoeffding's inequality (Hoeffding, 1963).

For $E_e = (\widehat{\theta}_e - \theta_e)^2 \geq 0$, pointwise comparison and linearity of expectation give

$$\max_e \mathbb{E}E_e \leq \mathbb{E}\max_e E_e \leq \sum_e \mathbb{E}E_e \leq |F| \max_e \mathbb{E}E_e.$$

Markov's inequality then gives $\Pr(\max_e |\widehat{\theta}_e - \theta_e| > \varepsilon) \leq |F|d/\varepsilon^2$ under marginal MSE at most d. For the direct complete-vector estimator based on $N = h + q_{\mathrm{sim}}$ rows, each contrast observation lies in $[-1, 1]$ and has mean θ_e. Hoeffding's inequality and a union bound yield

$$\Pr\Big(\max_e |\widehat{\theta}_e - \theta_e| > \varepsilon\Big) \leq \min\{1, 2|F|e^{-N\varepsilon^2/2}\}.$$

The integer choice in (42) makes the last quantity at most α. Integrating the same tail with $u_0 = 2\log(2|F|)/N$ proves

$$\mathbb{E}\max_e E_e \leq u_0 + \int_{u_0}^{\infty} 2|F|e^{-Nu/2}\,du = \frac{2[\log(2|F|) + 1]}{N}.$$

These arguments average over the random initial cache. Fixed-budget simultaneous estimation and time-uniform stopping have different coverage requirements; the latter is addressed by confidence-sequence methods such as Howard et al. (2021).

D.4 Conditioning on the numerical cache

Define the cache-only conditional minimax risk by

$$\mathcal{R}^\star_{\mathcal{C}} = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}_{\mathrm{bd}}} \operatorname*{ess\,sup}_{\mathcal{C} \sim P^{\otimes h}} \max_{e \in F} \mathbb{E}_P[(\widehat{\theta}_e - \theta_e)^2 \mid \mathcal{C}].$$

The estimator may use an independent random seed but no fresh observations. For every finite h and every nonempty comparison graph,

$$\mathcal{R}^\star_{\mathcal{C}} = 1. \tag{43}$$

To prove the lower bound, fix an edge (i, j) and use independent Bernoulli coordinates with all means strictly between zero and one. The all-zero cache has positive probability under each such law. Conditional on that cache, the estimator's distribution is fixed independently of the means. Let A be its reported value for this edge. Clipping to $[-1, 1]$ can only improve squared error. By approaching $\mu_i - \mu_j = 1$ and -1 through interior Bernoulli means, the supremum conditional risk is at least

$$\max\{\mathbb{E}(A - 1)^2, \mathbb{E}(A + 1)^2\} = 1 + \mathbb{E}A^2 + 2|\mathbb{E}A| \geq 1.$$

The zero contrast vector has squared error at most one for every edge and population, proving equality. Thus cache-conditional minimax accuracy is a different target from the repeated-sampling criterion used in the main text.

D.5 Exact contract checks

For the packed K_6 example in Appendix Q, exact rational evaluation of all 45 supported layouts reproduces their union-count laws and the common lottery shrinkage calculation. The resulting marginal uniform-risk ratio is $0.999743746186\ldots$; layouts with positive probability exceed the corresponding layout-wise target on the designated complementary Bernoulli edge. The same arithmetic reproduces the star construction costs and the integer-refresh squeeze summarized in Appendix D.7.

D.6 The contract gap near cache adequacy

Write $\tau_c(H)$ for the minimum weighted vertex-cover cost and $\tau_{f,c}(H)$ for the fractional-cover value

$$\tau_{f,c}(H) = \min\left\{\sum_i c_i z_i : z_i \geq 0,\ z_i + z_j \geq 1\ (ij \in F)\right\}.$$

Let $\mathcal{W}^\star_{\mathrm{lay}}(R, d; c)$ denote the minimum hard work under the contract

$$\inf_{\delta} \sup_{S \in \operatorname{supp} \nu} \sup_{P \in \mathcal{P}_{\mathrm{bd}}} \max_e \mathbb{E}_{Y|S}(\delta_e(S, Y) - \theta_e)^2 \leq d.$$

Only the layout is conditioned on. Cached and future score observations are still random.

Theorem 8 (Layout certification and incremental precision)**.** *For every graph, target and request cap,*

$$\mathcal{W}^{\star}_{\mathrm{lay}}(R, d; c) = \mathcal{W}^{\star}_{\mathrm{det}}(R, d; c).$$

Suppose $0 < \ell = n_d - h \le h$*, put* $L = \lceil \ell \rceil$*, and allow unrestricted requests. Then*

$$\mathcal{W}^{\star}_{\mathrm{lay}}(\infty, d; c) = \begin{cases} L\tau_{f,c}(H), & L \text{ even}, \\ (L-1)\tau_{f,c}(H) + \tau_c(H), & L \text{ odd}. \end{cases}$$

For the ordinary marginal optimum,

$$1 \le \frac{\mathcal{W}^{\star}_{\mathrm{lay}}(\infty, d; c)}{\mathcal{W}^{\star}(\infty, d; c)} \le 1 + \frac{2}{\ell}.$$

At an aligned endpoint $\ell = L$*, the ratio is one for even* L *and at most* $1 + 1/L$ *for odd* L*. The odd aligned additive gap is at most* $\tau_c(H) - \tau_{f,c}(H)$*.*

Proof. Every supported layout under the layout contract is itself a deterministic feasible layout. Selecting one such layout preserves both hard caps, while a deterministic feasible layout is a point-mass lottery. Equivalently, the conditional minimax value of a finite-support lottery is the maximum of the individual layout minimax values: its estimator may be specified separately on each observed layout.

For a deterministic layout, the binary lower bound $m(U_e) \le d$ requires the integer union $U_e \ge \lceil n_d \rceil = h + L$. Thus its fresh marginal counts form an integer L-cover. Conversely, truncating such counts at L and giving them private positions produces a balanced layout because $L \le h$, attaining $m(h + L) \le d$. The deterministic optimum is therefore the integer L-cover value, giving the two exact formulas.

The ordinary marginal lower bound is $\ell\tau_{f,c}$. The deterministic upper value is at most $(L + 1)\tau_{f,c} \le (\ell + 2)\tau_{f,c}$, proving the general ratio. At $\ell = L$, the even values coincide. The odd formula differs from the marginal lower bound by $\tau_c - \tau_{f,c} \le \tau_{f,c}$, proving both the sharper ratio and additive bound. □

Proposition 3 (An exact separation next to cache adequacy)**.** *Let* δ_u *denote the point mass at* u*. For any integer* $h \ge 1$ *and* $K \ge 3$*, define*

$$d_K = g((1 - 2/K)\delta_h + (2/K)\delta_{h+1}). \tag{44}$$

Then $m(h + 1) < d_K < m(h)$*. On* K_K *with unit costs,*

$$\mathcal{W}^{\star}(1, d_K) = \mathcal{W}^{\star}(\infty, d_K) = 1, \qquad \mathcal{W}^{\star}_{\mathrm{lay}}(1, d_K) = \mathcal{W}^{\star}_{\mathrm{lay}}(\infty, d_K) = K - 1.$$

Moreover $n_{d_K} - h = \Theta_h(1/K)$*. All lower bounds remain valid for binary scores.*

Proof. Query one uniformly chosen candidate on one fresh position. Every supported layout has hard work and request count one. Each fixed edge has union $h + 1$ with probability $2/K$ and h otherwise, so the balanced exact risk reduction gives d_K. Since both probabilities are strictly positive, the scalar Bayes formula is strictly between $m(h + 1)$ and $m(h)$. Zero added work has minimax risk $m(h)$ and cannot meet the target.

For any deterministic layout with work below $K - 1$, at least two candidates receive no fresh observation. Their edge still has union count h and minimax risk at least $m(h) > d_K$, using its fixed binary subfamily. Hence every feasible realized layout needs work at least $K - 1$, independently of how many requests are available. One fresh request containing a minimum vertex cover of size $K - 1$ gives every edge union $h + 1$ and attains $m(h + 1) < d_K$. The preceding theorem equates layout-certified and deterministic optima.

Writing $\varepsilon = 2/K$, the scalar variational formula and evaluation at $t = \sqrt{h}$ give

$$\varepsilon\{m(h) - m(h + 1)\} \le m(h) - d_K \le \frac{\varepsilon m(h)}{h + \sqrt{h} + 1}.$$

The inverse $m^{-1}(d) = n_d$ has a finite positive derivative in magnitude on this fixed interval, so $n_{d_K} - h = \Theta_h(1/K)$. The fixed-edge lower-bound population uses an anticorrelated Bernoulli pair and zeros elsewhere. □

Thus large relative randomization gains can occur close to the cache-adequacy boundary. When the incremental effective sample deficit grows, the ratio bounds above limit the price of certifying each realized layout. Both contracts average over score randomness; they differ in whether layout randomization is averaged or certified layout by layout.

D.7 Finite construction comparison

For Table 1, set $h = 192$, $M \in \{10, 19\}$, and $d = d_0(h)$. The full-vector plan uses $q = \lceil n_d \rceil - h = 633$ fresh rows, while the refresh–cover plan in (29) has $N = 825$ and $N_{\text{full}} = 413$. The random candidate–reference split minimizes (10) over integer s, and the deterministic split minimizes the upper side of (19). The selected (s, r_0) are $(697, 876)$ and $(950, 1133)$; the deterministic choices (k, r_0) are $(69, 883)$ and $(50, 1133)$. Their work values are 1573 and 2083, with finite global layout-wise lower bounds 1322 and 1824.

For the dense-library comparisons in Section 6, at $M = 4h^2 = 147456$ the marginal split has work 164108, while (19) gives the layout-wise interval $[192498, 193116]$. At $M = 8h^2 = 294912$, the corresponding values are 232084 and $[339954, 340572]$. These finite intervals certify the contract separation used in the main text.

The integer minimizations have finite tail bounds: once an incumbent cost is U, budgets $s \geq U$ and $Mk \geq U$ cannot improve the corresponding construction. Rational interval arithmetic resolves square-root and ceiling boundaries. Evaluating both sides of (19) across $h \in \{16, 64, 256, 1024, 4096\}$ and $\eta \in \{1/25, 1/4, 1/2, 1, 2, 4\}$ gives the finite crossover values reported with $\Phi(\eta)$.

E Proofs for the bounded-score reduction

This proof extends the bounded-mean equalizer argument (Lehmann & Casella, 1998) to shared partial layouts.

E.1 Balanced contrasts as averages of bounded observations

Fix an edge (i, j) and write $k = k_e$, $a = n_i - k$, $s = n_j - k$, and $N = U_e = k + a + s$ for its shared count, solo counts, and union count. The balance inequalities are equivalent to $|a - s| \leq k$, so $\omega = (k + s - a)/(2k)$ lies in $[0, 1]$. Every transformed observation X_t° in (24) therefore lies in $[0, 1]$. Distinct request positions are independent. The identities

$$a + k\omega = N/2, \qquad s + k(1 - \omega) = N/2$$

give

$$\sum_t \mathbb{E}X_t^\circ = \frac{N}{2}\mu_i + \frac{N}{2}(1 - \mu_j) = \frac{N}{2}(1 + \theta_e).$$

Consequently $Z_e = 2N^{-1}\sum_t X_t^\circ - 1$ is unbiased.

Let $q_t = \mathbb{E}X_t^\circ$ and $\bar{q} = N^{-1}\sum_t q_t = (1 + \theta_e)/2$. Boundedness gives $\mathrm{Var}(X_t^\circ) \leq q_t(1 - q_t)$. Independence and concavity of $q(1 - q)$ yield

$$\mathrm{Var}(Z_e) = \frac{4}{N^2}\sum_t \mathrm{Var}(X_t^\circ) \leq \frac{4}{N^2}\sum_t q_t(1 - q_t) \leq \frac{4}{N}\bar{q}(1 - \bar{q}) = \frac{1 - \theta_e^2}{N}.$$

The X_t° need not be identically distributed. In the main-text notation, the resulting identities are

$$\mathbb{E}(Z_e \mid S) = \theta_e, \qquad \mathrm{Var}(Z_e \mid S) \leq \frac{1 - \theta_e^2}{U_e}. \tag{45}$$

Expanding the transformed observations shows that this Z_e has the following explicit coefficients: solo observations have absolute coefficient $2/N$, and paired coefficients are

$$p_i = \frac{k + s - a}{kN}, \qquad p_j = \frac{k + a - s}{kN}.$$

The coefficient of each j observation is negative, and its total absolute weight is one.

For comparison with a direct moment calculation, put $D = Y_i - Y_j$, $A = Y_i + Y_j$ and $\eta_e = (a + s + (a - s)^2/k)/N \in [0, 1]$. The same coefficients satisfy the exact identity

$$\mathrm{Var}(Z_e) = \frac{\mathrm{Var}(D) + \eta_e \,\mathrm{Var}(A)}{N}.$$

The pointwise inequality $(Y_i - Y_j)^2 + (Y_i + Y_j - 1)^2 \le 1$ provides an alternative derivation of the bound.

E.2 Deterministic layouts: shrinkage and the lower bound

For $\lambda_N = \sqrt{N}/(\sqrt{N} + 1)$, unbiasedness gives

$$\begin{aligned}\mathbb{E}(\lambda_N Z_e - \theta_e)^2 &= \lambda_N^2 \,\mathrm{Var}(Z_e) + (1 - \lambda_N)^2\theta_e^2 \\ &\le \frac{\lambda_N^2}{N}(1 - \theta_e^2) + (1 - \lambda_N)^2\theta_e^2 = \frac{1}{(\sqrt{N} + 1)^2}.\end{aligned}$$

This establishes the upper bound using only score range and independence between request positions.

For the matching lower bound, fix one target edge before drawing observations and consider the family

$$Y_i = X, \qquad Y_j = 1 - X, \qquad X \sim \mathrm{Bernoulli}(p), \quad 0 \le p \le 1.$$

Set all other coordinates to the constant 0. This is a valid population on $\{0, 1\}^K \subset [0, 1]^K$ for every p. At each request position revealing at least one endpoint, the observed score or its complement reveals one Bernoulli observation. Paired observations reveal that same variable twice, so the experiment has exactly N independent Bernoulli observations, not $n + m$. The other coordinates contain no information about p.

Let T be the number of successes and give p the proper prior $\mathrm{Beta}(\sqrt{N}/2, \sqrt{N}/2)$. The target is $\theta_e = 2p - 1$, and its posterior mean is

$$\frac{2T - N}{N + \sqrt{N}} = \lambda_N \left(\frac{2T}{N} - 1\right).$$

Under this subfamily every transformed observation X_t° equals the same Bernoulli variable at its position. Thus $Z_e = 2T/N - 1$, with variance $(1 - \theta_e^2)/N$. The posterior-mean estimator has frequentist risk $1/(\sqrt{N} + 1)^2$ for every p. It is Bayes and has constant risk, so its Bayes risk lower-bounds the maximum risk of every estimator using the full table. This proves exact all-estimator minimax optimality. The prior and population parameter are chosen before any data.

E.3 One prior across a randomized layout

Fix an edge and let $N = U_e(S)$ be its positive, finite-support random union count under an outcome-blind lottery. For $t > 0$ define

$$G(t) = \mathbb{E}\frac{N - t^2}{(N + t)^2}, \qquad B(t) = \frac{t}{t + 1}\mathbb{E}\frac{1}{N + t}.$$

Differentiating finite sums gives

$$G'(t) = \mathbb{E}\frac{-2N(t + 1)}{(N + t)^3} < 0.$$

Also $G(0) = \mathbb{E}(1/N) > 0$ and $\lim_{t\to\infty} G(t) = -1$. There is exactly one root $t_\star > 0$, lying between $\sqrt{\min N}$ and $\sqrt{\max N}$. Moreover

$$B'(t) = \frac{G(t)}{(t + 1)^2},$$

so $t_\star$ is the unique maximizer of $B(t)$.

For the upper bound, use the layout-dependent coefficient $\lambda_N = N/(N + t_\star)$ multiplying the unbiased contrast. Conditioning on the layout and then averaging,

$$\mathbb{E}(\lambda_N Z_e - \theta_e)^2 \le A(1 - \theta_e^2) + C\theta_e^2,$$
$$A = \mathbb{E}\frac{N}{(N + t_\star)^2}, \qquad C = \mathbb{E}\frac{t_\star^2}{(N + t_\star)^2}.$$

The root equation is $A = C$, so the risk is at most A for every bounded population. The algebraic identity

$$\mathbb{E}\frac{1}{N + t_\star} = A + t_\star \mathbb{E}\frac{1}{(N + t_\star)^2} = A + \frac{A}{t_\star}$$

shows that $A = B(t_\star)$.

For the lower bound, use the same fixed anticorrelated Bernoulli family as above, with the single proper prior $\mathrm{Beta}(t/2, t/2)$, where t is chosen before the lottery. The layout is independent of p and supplies no likelihood information about it. Conditional on a realized layout and success count T, the posterior mean of θ_e is

$$\frac{2T - N}{N + t}.$$

Since this prior has $\mathbb{E}\theta_e^2 = 1/(t + 1)$, the prior-average squared error conditional on N equals

$$\frac{N}{(N + t)^2}\frac{t}{t + 1} + \frac{t^2}{(N + t)^2}\frac{1}{t + 1} = \frac{t}{(t + 1)(N + t)}.$$

After averaging over the layout its Bayes risk is $B(t)$. Therefore every estimator has maximum risk at least $B(t)$ for every $t > 0$. Choosing $t = t_\star$ matches the upper bound.

All observations of the remaining coordinates are constant and uninformative in this lower-bound family. Thus the proof covers estimators using the entire table, the full layout rather than just N, and additional independent estimator randomness, covering arbitrary measurable estimators.

For a graph, the construction supplies one upper-bound estimator for each edge simultaneously. For necessity, $\sup_P \max_e r_e(P) = \max_e \sup_P r_e(P)$ on the finite edge set. Each edge may therefore use its own fixed Bernoulli family and fixed prior in its lower-bound argument. This proves (23) using one fixed population family for each edge.

E.4 Why averaging fixed-layout optima is inexact

For a nonconstant edge union count, the comparisons are

$$m(\mathbb{E}U_e) < g(\nu_e) < \mathbb{E}m(U_e). \tag{46}$$

For a fixed positive number u,

$$\max_{t>0} \frac{t}{(t + 1)(u + t)} = m(u), \qquad \text{with unique maximizer } t = \sqrt{u}.$$

Consequently

$$\max_t \mathbb{E}\frac{t}{(t + 1)(N + t)} \le \mathbb{E}m(N).$$

If N is nonconstant, no single t maximizes every positive-probability summand. At $t = t_\star$ the inequality is therefore strict.

Conversely, $u \mapsto 1/(u + t)$ is strictly convex for each $t > 0$. Jensen's inequality at $t = \sqrt{\mathbb{E}N}$ gives

$$\max_t \frac{t}{t + 1}\mathbb{E}\frac{1}{N + t} \ge \max_t \frac{t}{(t + 1)(\mathbb{E}N + t)} = m(\mathbb{E}N).$$

The inequality is strict for a nonconstant N. Both comparisons concern the same outcome-blind marginal risk contract. The difference from $\mathbb{E}m(N)$ comes from using a common prior and population across layouts.

F PRECISION LIMITS FOR FULLY ADAPTIVE COMPLETION

The converse uses a multivariate van Trees inequality (Gill & Levit, 1995). For $A = A(H, c)$ and $C = C_+$, this section proves the finite adaptive work sandwich (52), which sharpens the guarantee in Section 5.2, together with its exact request and joint-cost claims.

Let $H = ([K], F)$ be a fixed nonempty graph, let $V_+ = \{i : \deg_H(i) > 0\}$, and put $C_+ = \sum_{i \in V_+} c_i$. Every coordinate has price $c_i > 0$, and all fresh queried coordinates contribute to work; the constructions query only V_+. The cache contains $h \geq 1$ complete independent score vectors. An adaptive policy may read every cached value. Before each new independent sample is drawn, it selects its queried subset using all previously observed values and independent internal randomness. It may stop according to those observations. Let r be its number of nonempty requests, let x_i be its fresh observation count at candidate i, and write $N_i = h + x_i$. Its estimator may be biased and may use the entire transcript. The uniform requirement is

$$\sup_{P:\, \mathrm{supp}(P) \subseteq [0,1]^K} \max_{ij \in F} \mathbb{E}_P(\widehat{\theta}_{ij} - (\mu_i - \mu_j))^2 \leq d. \tag{47}$$

The expectation includes the cache, all later observations, and the policy's internal randomness. We consider finite hard request/work/total-cost caps. A finite hard work or total-cost cap itself bounds r, since each nonempty request costs at least $\min_i c_i$.

Write

$$m(n) = (\sqrt{n} + 1)^{-2}, \qquad n_d = (d^{-1/2} - 1)^2, \qquad q(d, h) = (\lceil n_d \rceil - h)_+.$$

Let $\mathcal{W}^\star_{\mathrm{ad}}(\infty, d; c)$ and $\mathcal{J}^\star_{\mathrm{ad}}(a, d; c)$ be the minimum hard work and hard total charge $ar + W_c$, respectively, among these policies, with $a \geq 0$. All caps hold on every realized transcript.

F.1 THE MINIMUM REQUEST COUNT REMAINS EXACT

Theorem 9 (Adaptive request complexity)**.** *For every integer $R \geq 0$, the best uniform maximum-edge MSE using at most R fresh requests, with unrestricted candidate work, is exactly $m(h + R)$. Hence the minimum hard request cap for target d is $q(d, h)$. Querying every vertex in V_+ on each of these $q(d, h)$ fresh positions attains the target, with work $q(d, h)C_+$.*

Proof. A table of $h + R$ complete score vectors, together with independent policy randomness, can simulate any allowed adaptive policy: provide its first h rows as the cache and reveal the chosen subset of the next complete row at each successive request. Stop revealing rows when the policy stops. This is a parameter-independent transformation, even when the policy reads all cache values or randomizes its stopping time.

For a fixed edge ij, restrict the population to $Y_i = B$, $Y_j = 1 - B$, with $B \sim \mathrm{Bernoulli}(p)$ and all other coordinates zero. The complete table supplies exactly $h + R$ Bernoulli observations about the contrast $2p - 1$. Their all-estimator minimax risk is $m(h + R)$, as follows from the symmetric beta equalizer argument. A simulated adaptive transcript cannot improve this lower bound.

For attainment, observe all active vertices on R fresh positions. There are $n = h + R$ complete active score vectors. For every edge, its paired sample mean D_{ij} is unbiased and has variance at most $(1 - \theta_{ij}^2)/n$. The common shrinkage $\widehat{\theta}_{ij} = \sqrt{n}/(\sqrt{n} + 1)\, D_{ij}$ therefore has risk at most $m(n)$. It is also induced by the bounded vector of estimates $\widehat{\mu}_i = 1/2 + \sqrt{n}/(\sqrt{n} + 1)(\bar{Y}_i - 1/2)$. □

F.2 ONE COMMON PRIOR CHARGES EVERY ACTIVE ENDPOINT

For $0 < \varepsilon < 1/2$, let Π_ε be the product prior on $p = (p_1, \ldots, p_K)$ whose marginal density is

$$\lambda_\varepsilon(p_i) = \frac{1}{\varepsilon} \cos^2\left(\frac{\pi(p_i - 1/2)}{2\varepsilon}\right) \mathbf{1}\{|p_i - 1/2| \leq \varepsilon\}.$$

Conditional on p, use the population $P_p = \bigotimes_i \mathrm{Bernoulli}(p_i)$. The notation $\mathbb{E}_\varepsilon$ averages over this prior on fixed population parameters, the conditionally independent cache and future rows, and policy randomness.

Define

$$F_\varepsilon = \frac{4}{1-4\varepsilon^2}, \qquad B_\varepsilon = \frac{\pi^2}{\varepsilon^2}.$$

The following is the multivariate van Trees argument (Gill & Levit, 1995), with the adaptive information calculated explicitly.

Lemma 1 (Adaptive information and edge risk)**.** *For every edge ij, any adaptive policy satisfying (47) obeys*

$$\frac{1}{F_\varepsilon \mathbb{E}_\varepsilon N_i + B_\varepsilon} + \frac{1}{F_\varepsilon \mathbb{E}_\varepsilon N_j + B_\varepsilon} \le d. \tag{48}$$

Consequently every $i \in V_+$ satisfies

$$\mathbb{E}_\varepsilon x_i \ge \left(\frac{d^{-1} - B_\varepsilon}{F_\varepsilon} - h \right)_+, \tag{49}$$

$$\mathbb{E}_\varepsilon r \ge \left(\frac{2d^{-1} - B_\varepsilon}{F_\varepsilon} - h \right)_+. \tag{50}$$

These inequalities charge prior-averaged counts under the same prior for every vertex and edge.

Proof. Include the policy's internal random seed in the transcript. On the product Bernoulli subfamily, the likelihood of a finite transcript factors into parameter-free policy terms and

$$\prod_i p_i^{T_i}(1-p_i)^{N_i - T_i},$$

where T_i is the total observed number of ones at coordinate i. The selected subsets and stopping indicators are functions of the observed history; their kernels add no parameter derivative.

For fixed p, the ith likelihood score is the sum of the cache and fresh increments $(Y_i - p_i)/[p_i(1-p_i)]$ whenever i is observed. The fresh increment is multiplied by a predictable query indicator. These are martingale differences, including after padding a bounded stopping time with empty requests. Their conditional variances add, so

$$\mathbb{E}_p D_i^2 = \frac{\mathbb{E}_p N_i}{p_i(1-p_i)}, \qquad \mathbb{E}_p D_i D_j = 0 \quad (i \ne j).$$

For two coordinates queried on the same row, independence under P_p makes the cross term zero. For different rows the martingale difference property makes the cross term zero. Cache scores and future increments are orthogonal for the same reason.

The prior has zero density at both endpoints of its support and

$$\int \frac{(\lambda_\varepsilon')^2}{\lambda_\varepsilon} = \frac{\pi^2}{\varepsilon^2}.$$

Thus the score Z of the joint parameter–transcript density has diagonal second-moment matrix with entries

$$I_i = \mathbb{E}_\varepsilon \frac{N_i}{p_i(1-p_i)} + B_\varepsilon \le F_\varepsilon \mathbb{E}_\varepsilon N_i + B_\varepsilon.$$

The prior–likelihood cross terms vanish because $\mathbb{E}_p D_i = 0$; different prior coordinates are independent.

Clipping a contrast estimate to $[-1, 1]$ cannot increase loss, so the estimator can be assumed bounded. Integration by parts over p_ℓ then gives

$$\mathbb{E}_\varepsilon[(\widehat{\theta}_{ij} - (p_i - p_j))Z_\ell] = \mathbf{1}\{\ell = i\} - \mathbf{1}\{\ell = j\}.$$

There is no assumption that the estimator is unbiased. Applying Cauchy–Schwarz to $Z_i/I_i - Z_j/I_j$ yields

$$\mathbb{E}_\varepsilon(\widehat{\theta}_{ij} - (p_i - p_j))^2 \ge \frac{1}{I_i} + \frac{1}{I_j}.$$

The prior Bayes risk is at most d under (47). Substituting the upper bounds on I_i, I_j proves (48).

Each endpoint belongs to an edge, so dropping the other positive reciprocal gives (49). Also $N_i, N_j \le h + r$ pathwise. Substitution into (48) gives $d \ge 2/[F_\varepsilon(h + \mathbb{E}_\varepsilon r) + B_\varepsilon]$, proving (50). Finite hard caps and the compact interior parameter support justify the score identities; the prior's finite Fisher information justifies the integration-by-parts step. □

F.3 AN ASYMPTOTICALLY EXACT ADAPTIVE WORK FRONTIER

The edgewise information bound also retains graph structure. Define the public convex allocation value

$$A(H,c) = \min_{\substack{z_i>0\ (i\in V_+)\\ z_i^{-1}+z_j^{-1}\le 4\ (ij\in F)}} \sum_{i\in V_+} c_i z_i. \tag{51}$$

Its objective concerns candidate work with unrestricted requests. The feasible choice $z_i = 1/2$ gives $A(H,c) \le C_+/2$. Every feasible z_i is at least $1/4$, so $A(H,c) \ge C_+/4$. Positive costs make the objective coercive on its nonempty feasible sublevel sets, hence a minimum is attained.

Theorem 10 (Adaptive high-precision work constant)**.** *Write $A = A(H,c)$ and $C = C_+$. For $0 < d < A/(\pi^2 C)$,*

$$\left[\left(\sqrt{\frac{A}{d}} - \pi\sqrt{C}\right)^2 - hC\right]_+ \le \mathcal{W}^\star_{\mathrm{ad}}(\infty, d; c) \le \frac{A}{d} + C. \tag{52}$$

In particular,

$$\frac{d\,\mathcal{W}^\star_{\mathrm{ad}}(\infty, d; c)}{A(H,c)} \longrightarrow 1 \qquad (d \to 0,\ dh \to 0).$$

This relative convergence is uniform over finite nonempty graphs and positive candidate costs, including sequences of graphs whose size grows. The upper bound uses an explicit deterministic private-sampling allocation; the lower bound permits reading the cache and fully adaptive sampling and stopping.

Proof. Fix the same prior Π_ε as above and put

$$z_i = d\left(\mathbb{E}_\varepsilon N_i + \frac{B_\varepsilon}{F_\varepsilon}\right).$$

Equation (48) gives $z_i^{-1} + z_j^{-1} \le F_\varepsilon$. Consequently $(F_\varepsilon/4)z$ is feasible for (51), and

$$\mathbb{E}_\varepsilon W_c \ge \frac{4A}{F_\varepsilon d} - \left(h + \frac{B_\varepsilon}{F_\varepsilon}\right) C.$$

For $s = \varepsilon^2 \in (0, 1/4)$, this expression is

$$\frac{A}{d} - \frac{4As}{d} - \frac{\pi^2 C}{4s} + (\pi^2 - h)C.$$

Its maximum occurs at $s = (\pi/4)\sqrt{dC/A} < 1/4$, giving the lower expression in (52). The hard cap dominates this common-prior expected work.

For the upper bound, let $z^\star$ attain A. Give each active candidate i exactly $n_i = \lceil z_i^\star/d \rceil$ new singleton observations, using disjoint positions for different candidates. Use the means of these fresh observations, without using the cache. Each bounded score has variance at most $1/4$, and different candidates' fresh means are independent because their positions are disjoint. Thus every edge has unbiased squared error at most

$$\frac{1}{4n_i} + \frac{1}{4n_j} \le \frac{d}{4}\left(\frac{1}{z_i^\star} + \frac{1}{z_j^\star}\right) \le d.$$

The hard work is at most $A/d + C$. Its request count is $\sum_i n_i$, since all observations are private singleton queries; this equals work only at unit prices.

For uniformity, division by A/d gives the lower expression

$$1 - 2\pi\sqrt{\frac{dC}{A}} + \pi^2\frac{dC}{A} - \frac{dhC}{A}$$

and upper expression $1 + dC/A$. Since $C/A \le 4$, both approach one whenever $d \to 0$ and $dh \to 0$, with errors bounded independently of graph size and the positive prices. □

For a unit-price star with $M = K - 1$ leaves, the allocation is

$$z_0^\star = \frac{1+\sqrt{M}}{4}, \qquad z_i^\star = \frac{1+1/\sqrt{M}}{4}\ (i \neq 0), \qquad A(H,\mathbf{1}) = \frac{(1+\sqrt{M})^2}{4}.$$

Indeed, for a fixed center value $z_0 > 1/4$, every leaf is minimized at $z_i = z_0/(4z_0-1)$; differentiation of $z_0 + Mz_0/(4z_0-1)$ gives the displayed optimum. For either a matching on K active vertices or a complete graph on K vertices, $A(H,\mathbf{1}) = K/2$. For a matching this follows by minimizing each pair separately. For a complete graph, symmetry and convexity reduce to $z_i = 1/2$. The range $C_+/4 \le A(H,c) \le C_+/2$ shows that graphs with the same total active candidate cost differ by at most a factor two in their high-precision leading work. This concerns work with unrestricted requests; request prices can still distinguish their executable schedules.

F.4 A STAR EXHIBITS THE PRECISION CHANGE INSIDE THE ADAPTIVE CLASS

Proposition 4 (Exact low-deficit star and high-precision cost)**.** *Let H be a unit-price star on $K \ge 2$ active vertices. If $q = q(d,h) \le h$, then*

$$\mathcal{W}^\star_{\mathrm{ad}}(\infty,d) = q, \qquad \mathcal{J}^\star_{\mathrm{ad}}(a,d) = (a+1)q.$$

Both are attained using q center-only fresh requests. For the same adaptive class,

$$\mathcal{W}^\star_{\mathrm{ad}}(\infty,d) \sim \frac{(1+\sqrt{K-1})^2}{4d} \qquad (d \to 0,\ dh \to 0),$$

while at integer deficits $1 \le L \le h$ the target $d = m(h+L)$ has exact work L, independently of K.

Proof. If $q = 0$, the cache-only bounded shrinkage rule suffices. For $1 \le q \le h$, query the center on q new positions. Each star edge has margins $h+q, h$, overlap h, and union $h+q$. The margins are balanced because $q \le h$. The bounded partial-layout estimator attains risk $m(h+q) \le d$ and work q.

Every nonempty request costs at least one candidate evaluation. A hard work cap below q would therefore imply a hard request cap below q, contradicting Theorem 9. More generally, every realized joint charge is at least $(a+1)r$. A hard total cap below $(a+1)q$ would again force $r < q$ on every transcript, which is impossible. These arguments permit arbitrary randomized stopping and sampling based on the cache and later scores. The high-precision statement follows from Theorem 10 and the star allocation computed above. □

The balanced center-only identity holds for $q \le h$. Appendix C locates the dimension threshold beyond this range and gives the sharp intermediate and critical work laws under the same adaptive policy class.

F.5 AN ENDPOINT BOUND AT ONE FIXED POPULATION

The preceding sharper constants use prior-averaged counts. For a complementary statement under one fixed population, let $P_0 = \bigotimes_i \mathrm{Bernoulli}(1/2)$. For every $0 < d \le 1/192$, any uniformly feasible adaptive policy has

$$\mathbb{E}_{P_0} N_i \ge \frac{1}{36d} \quad \text{for every } i \in V_+. \tag{53}$$

This complementary bound charges all endpoints under one fixed population. To prove it, fix a neighbor edge of i and perturb only its ith mean by $\epsilon = \sqrt{12d} \le 1/4$, forming P_i. The contrast changes by ϵ. If $\mathsf{Q}_0, \mathsf{Q}_i$ are the two transcript laws, integration of $(z-\theta_0)^2 + (z-\theta_i)^2 \ge \epsilon^2/2$ against their minimum density gives $1 - \mathrm{TV}(\mathsf{Q}_0,\mathsf{Q}_i) \le 4d/\epsilon^2 = 1/3$. Pinsker's inequality yields $\mathrm{KL}(\mathsf{Q}_0,\mathsf{Q}_i) \ge 8/9$. The adaptive likelihood chain rule gives

$$\mathrm{KL}(\mathsf{Q}_0,\mathsf{Q}_i) = \mathbb{E}_{P_0} N_i\ \mathrm{kl}(1/2, 1/2+\epsilon) \le 32d\,\mathbb{E}_{P_0} N_i,$$

since $-\frac{1}{2}\log(1-4\epsilon^2) \le 2\epsilon^2/(1-4\epsilon^2) \le (8/3)\epsilon^2$. This proves (53). The sequential change-of-measure method is classical (Kaufmann et al., 2016); its role here is to place every endpoint's lower bound at the same P_0.

G GRAPH STRUCTURE WITHIN THE CACHE PRECISION RANGE

The allocation viewpoint is related to classical optimal design (Pukelsheim, 2006). Write $c(C) = \sum_{i \in C} c_i$. Let $\tau_c(H)$ be the minimum weighted vertex-cover cost and

$$\tau_{f,c}(H) = \min \left\{ \sum_i c_i z_i : z_i \geq 0,\ z_i + z_j \geq 1\ (ij \in F) \right\}.$$

Write τ, τ_f for unit prices and $a_+ = \max\{a, 0\}$.

Theorem 11 (General graph resource laws)**.** *For every target $0 < d < 1$, the minimum number of fresh sample positions is $(\lceil n_d \rceil - h)_+$, and every feasible precommitted completion lottery, with any finite request cap, satisfies*

$$\mathbb{E}W_c \geq (n_d - h)_+ \tau_{f,c}(H). \tag{54}$$

At an integer endpoint $d = m(N)$, where $h < N \leq 2h$, let $L = N - h$. Then the minimum-request work is exactly

$$\mathcal{W}^\star(L, m(N); c) = \mathcal{W}^\star_E(L, m(N); c) = L\tau_c(H). \tag{55}$$

With unrestricted requests, deterministic acquisition has exact cost

$$\mathcal{W}^\star_{\mathrm{det}}(\infty, m(N); c) = \begin{cases} L\tau_{f,c}(H), & L \text{ even}, \\ (L-1)\tau_{f,c}(H) + \tau_c(H), & L \text{ odd}. \end{cases} \tag{56}$$

Randomized hard and expected work lie between $L\tau_{f,c}(H)$ and this deterministic cost. Consequently the fractional lower bound is exact for every graph when L is even, and for every L when the graph is bipartite.

The lower bound uses the same bounded populations as Theorem 4. For any precommitted layout distribution, Jensen's inequality gives $g(\nu_e) \geq m(\mathbb{E}U_e)$, so $\mathbb{E}x_i + \mathbb{E}x_j \geq n_d - h$. At the minimum-request integer endpoint, $U_e \leq N$; the fixed $\mathrm{Beta}(\sqrt{N}/2, \sqrt{N}/2)$ Bayes bound forces $U_e = N$ almost surely. Every fresh queried subset must therefore be a vertex cover.

For even L, a half-integral fractional cover gives integer counts Lz_i; private positions attain the lower bound. For odd L, the integer multicover identity follows from half-integrality and an integer threshold decomposition (Nemhauser & Trotter, 1975). Its proof is in Appendix H. Appendix J packs these constructions into finitely many shared positions.

What graph structure changes. A star with M edges needs L unit-cost evaluations, while a matching with the same number of edges needs LM. For either workload extra positions reduce neither cost. On a general graph, additional positions save at most a factor of two over the minimum-request endpoint. For a unit-price reference star, the optimum also has a common mean-vector implementation in $[0, 1]^K$ (Appendix H.5).

A completion at every precision. Let $N = \max\{h, \lceil n_d \rceil\}$, $H_0 = \max\{h, \lceil N/2 \rceil\}$, and let $C \subseteq V_+$ be any vertex cover. First acquire $H_0 - h$ full active vectors, then query C on $N - H_0$ fresh positions. Every edge has union N, and each endpoint has at least $H_0 \geq N/2$ observations. Theorem 4 certifies risk at most d. The resources are

$$r = N - h, \qquad W_c = (H_0 - h)C_+ + (N - H_0)c(C). \tag{57}$$

The request count is adaptively optimal. This construction extends the executable upper bound beyond $L \leq h$; choosing the center cover is exact in work for the low-deficit unit star. At higher precision the additional full-vector layer supplies information to every endpoint.

Finite-request colorings, clique lower certificates, and exact triangle and six-model examples are developed in Appendix J. For aligned $1 \leq L \leq h$, the cover fallback gives a factor-two joint guarantee against the precommitted optimum; coloring improves the feasible request–work choices. The complete-graph all-budget result appears in Appendix I.

H PROOF OF THEOREM 11

Throughout this section, $H = (V, F)$ is a finite graph with $F \neq \varnothing$, $c_i > 0$ is the cost of one fresh observation of vertex i, and $h \geq 1$ full-vector positions form the cache. All request positions are

independent draws from one fixed population on $[0,1]^V$. The completion lottery is independent of all cached and future outcomes. Risk averages over both the observations and the lottery and permits all measurable edge estimators. For a realized layout S, let $x_i(S)$ be the number of fresh observations of i, let $U_e(S)$ be the total number of positions observing at least one endpoint of e, and let

$$W_c(S) = \sum_{i \in V} c_i x_i(S).$$

A hard work budget W means $W_c(S) \le W$ almost surely. A hard request budget R bounds the number of fresh request positions almost surely. Empty requests can be deleted.

Write

$$\tau_{f,c}(H) = \min_{\substack{z_i \ge 0 \\ z_i + z_j \ge 1 \ (ij \in F)}} \sum_i c_i z_i,$$

$$\tau_c(H) = \min_{\substack{C \subseteq V \\ C \text{ covers } F}} \sum_{i \in C} c_i, \qquad m(u) = \frac{1}{(\sqrt{u}+1)^2}.$$

The fractional program has an optimum in $\{0, \frac{1}{2}, 1\}^V$. The results below use this standard weighted half-integrality property (Nemhauser & Trotter, 1975).

H.1 A LOWER BOUND FOR EVERY LAYOUT

Lemma 2 (A common-prior lower bound)**.** *For any outcome-blind completion lottery, every edge's all-estimator minimax risk is at least*

$$g_e = \sup_{a>0} \frac{a}{a+1} \mathbb{E} \frac{1}{U_e + a} \ \ge \ m(\mathbb{E}U_e). \tag{58}$$

No balance condition on the pair margins is needed. For nonconstant U_e, the second inequality is strict.

Proof. Fix the edge $e = (i, j)$ before the lottery and restrict to the bounded population family

$$Y_i = B, \quad Y_j = 1 - B, \quad B \sim \text{Bernoulli}(p), \qquad Y_\ell = 0 \quad (\ell \notin \{i, j\}).$$

A position observing either endpoint reveals one Bernoulli draw; observing both reveals that same draw twice. Other coordinates contain no information about p. Every score vector in this family is binary, so the same lower bound holds if the population class is restricted to all distributions on $\{0,1\}^V$. Thus, conditional on the layout, the informative experiment has U_e independent Bernoulli observations. The target is $\theta_e = 2p - 1$.

Give p one proper prior $\text{Beta}(a/2, a/2)$, chosen before the lottery. The layout is ancillary. For every realized $U_e = u$, the integrated posterior-mean squared error for θ_e is $a/[(a+1)(u+a)]$. Consequently every estimator using the complete observed table has worst-case risk at least $a(a+1)^{-1}\mathbb{E}(U_e+a)^{-1}$. Taking the supremum over $a > 0$ gives the first inequality.

For any fixed $a > 0$, the function $u \mapsto (u+a)^{-1}$ is strictly convex. Jensen's inequality at $a = \sqrt{\mathbb{E}U_e}$ therefore gives

$$g_e \ge \frac{\sqrt{\mathbb{E}U_e}}{(\sqrt{\mathbb{E}U_e}+1)(\mathbb{E}U_e + \sqrt{\mathbb{E}U_e})} = m(\mathbb{E}U_e),$$

with strict inequality when U_e is nonconstant. The argument never uses the relative sizes of the two margins. □

H.2 GENERAL TARGETS AND EFFECTIVE PRECISION

For $0 < d < 1$, write $n_d = (d^{-1/2} - 1)^2$ as in (4).

Expected and hard work.

Proof. Since m is decreasing and $m(n_d) = d$, Lemma 2 gives $\mathbb{E}U_{ij} \ge n_d$ for every edge. A fresh position contributes at most one to the union count, whereas its contributions to the two marginal counts can sum to two. Hence

$$U_{ij} \le h + x_i + x_j, \qquad \mathbb{E}x_i + \mathbb{E}x_j \ge n_d - h.$$

If $n_d > h$, the vector $z_i = \mathbb{E}x_i/(n_d - h)$ is a fractional vertex cover. Multiplying its cost lower bound by $n_d - h$ proves (54). If $n_d \le h$, the bound is zero and nonnegativity suffices.

For the cache-only construction, every edge has deterministic union count h and balanced margins. Shrinking the common sample means by $\sqrt{h}/(\sqrt{h}+1)$ about $1/2$ gives maximum edge risk $m(h) \le d$. □

Minimum request count at every target. Let R be a hard cap on fresh requests. The bound $U_e \le h + R$ and Lemma 2 show that any feasible procedure must satisfy $h + R \ge n_d$. Since R is a nonnegative integer,

$$R \ge (\lceil n_d \rceil - h)_+.$$

Conversely, query the full candidate vector at exactly $R = (\lceil n_d \rceil - h)_+$ independent fresh positions. Every edge then has equal margins and deterministic union count $h + R \ge n_d$. The common bounded-mean shrinkage estimator has edge risk $m(h + R) \le d$. This construction remains balanced for arbitrary R, so this request formula holds for every $0 < d < 1$.

H.3 The minimum-request endpoint

Let $h < N \le 2h$ be integers, set $L = N - h$, and use target $d = m(N)$. We prove (55), including both the hard-work and expected-work optima.

Proof. With at most R fresh requests, $U_e \le h + R$ on every layout. If $R < L$, Lemma 2 gives minimax edge risk at least $m(\mathbb{E}U_e) \ge m(h + R) > m(N)$.

Now suppose $R = L$. Fix one edge and use in the lemma's Bernoulli family the single prior $\mathrm{Beta}(\sqrt{N}/2, \sqrt{N}/2)$. Its Bayes risk is

$$\frac{\sqrt{N}}{\sqrt{N}+1}\mathbb{E}\frac{1}{U_e + \sqrt{N}}.$$

Since $U_e \le h + L = N$, this expression is at least $m(N)$. Equality holds only if $U_e = N$ almost surely. Feasibility at target $m(N)$ therefore forces this equality on every edge. Intersecting the finitely many probability-one events shows that every realized fresh request must hit every edge: if a request missed an edge, that edge would gain fewer than L new union positions. Each fresh request is consequently a vertex cover and has cost at least $\tau_c(H)$. There must be exactly L such requests, so $W_c(S) \ge L\tau_c(H)$ almost surely.

Conversely, repeat a minimum-cost vertex cover on L independent fresh positions. Every edge has union count $N = h + L$. If only one endpoint is in the cover, its margins are $h + L$ and h, with h shared positions; if both are in the cover, both margins and the shared count are N. Both cases are balanced because $L \le h$. The deterministic bounded-score estimator therefore attains risk $m(N)$ on every edge. The construction uses exactly L requests and work $L\tau_c(H)$. Since the lower bound on work holds almost surely, the same value is optimal under expected work. □

H.4 Unlimited requests and parity of the deficit

Let $\mathcal{W}^\star_{\mathrm{det}}(\infty, m(N); c)$ and $\mathcal{W}^\star(\infty, m(N); c)$ denote the minimum hard completion work with no request cap, respectively for deterministic layouts and outcome-blind lotteries. Estimators remain unrestricted in both definitions. The target throughout this subsection is $m(N)$ with $L = N - h$ and $1 \le L \le h$.

Lemma 3 (Deterministic completion is an integer multicover)**.** *Under these conditions,*

$$\mathcal{W}^\star_{\mathrm{det}}(\infty, m(N); c) = \tau_L(H, c) := \min_{\substack{x_i \in \mathbb{Z}_{\ge 0} \\ x_i + x_j \ge L \ (ij \in F)}} \sum_i c_i x_i. \tag{59}$$

Proof. For any deterministic feasible layout, the lower bound in Lemma 2 is $m(U_e)$. Thus $U_e \ge N$ on every edge, and $U_{ij} \le h + x_i + x_j$ gives $x_i + x_j \ge L$. Its work is therefore at least $\tau_L(H, c)$. This necessity applies to both balanced and unbalanced margins.

Conversely, truncate an optimal integer multicover coordinatewise at L. This preserves feasibility and cannot increase cost, so there is an optimum with $0 \le x_i \le L$. Give vertex i exactly x_i private singleton positions, using distinct fresh positions for different vertices. For edge ij this gives

$$k_{ij} = h, \qquad n_i = h + x_i, \qquad n_j = h + x_j, \qquad U_{ij} = h + x_i + x_j \ge h + L = N.$$

Moreover $|x_i - x_j| \le L \le h$, which is exactly the pairwise balance condition. The balanced bounded-score estimator attains risk $m(U_{ij}) \le m(N)$. The work equals $\tau_L(H, c)$. The total number of fresh requests can exceed h; pairwise balance still holds. □

Lemma 4 (Weighted parity formula)**.** *For every finite graph with positive vertex costs,*

$$\tau_L(H, c) = \begin{cases} L\,\tau_{f,c}(H), & L \text{ even}, \\ (L-1)\tau_{f,c}(H) + \tau_c(H), & L \text{ odd}. \end{cases} \tag{60}$$

Proof. If L is even, the fractional relaxation has value $L\tau_{f,c}(H)$. A half-integral optimum z of the unit-demand fractional cover gives an integer feasible vector $x = Lz$, attaining this value.

For odd $L = 2k + 1$, first prove the lower bound. Given any feasible integer vector x, the set $C = \{i : x_i \ge k + 1\}$ covers every edge. Put $y = x - \mathbf{1}_C$. For an edge with one endpoint in C, its sum decreases by one and remains at least $L - 1$. For an edge with both endpoints in C, the original sum was at least $2k + 2 = L + 1$, so decreasing it by two again leaves at least $L - 1$. When $L > 1$, $y/(L-1)$ is therefore a fractional cover. It follows that

$$\sum_i c_i x_i = c(C) + \sum_i c_i y_i \ge \tau_c(H) + (L-1)\tau_{f,c}(H).$$

For $L = 1$, the same conclusion follows directly because $C = \{i : x_i \ge 1\}$ is a cover.

For the upper bound, take a half-integral optimal fractional cover z and, independently, a minimum-cost integral cover C. The vector $x = (L-1)z + \mathbf{1}_C$ is integer-valued, satisfies $x_i + x_j \ge L$ on every edge, and has cost $(L-1)\tau_{f,c}(H) + \tau_c(H)$. Moreover $0 \le x_i \le L$. This matches the lower bound. □

Conclusion for unrestricted requests. We now establish (56) and its randomized and expected-work consequences.

Proof. The bound (54) gives the randomized lower bound because $n_{m(N)} = N$. Every deterministic schedule is an admissible lottery, so its attainable work is an upper bound on $\mathcal{W}^\star(\infty, m(N); c)$. Lemmas 3 and 4 give all the displayed deterministic formulas. For even L, the bounds coincide. The same lower bound applies to expected work, and every deterministic construction satisfies both budget contracts. Thus the expected-work optimum also lies between $L\tau_{f,c}(H)$ and the deterministic value, and it equals $L\tau_{f,c}(H)$ for even L. On a bipartite graph the cover LP is integral, so $\tau_c(H) = \tau_{f,c}(H)$ and the bounds coincide for odd L as well. The following example proves the possible strict inequality for odd L. □

An exact odd-deficit separation. Take $H = K_5$, unit costs, $h = 3$, and target $d = m(4) = 1/9$, so $L = 1$. The minimum-request endpoint uses one request and work $\tau_c(K_5) = 4$. Deterministic completion has the same minimum work even without a request cap.

For a randomized completion, select a uniform three-vertex subset. Select a uniform pair from those three vertices and query that pair on one fresh position; query the remaining vertex on another fresh position. Every realized layout uses two requests and three fresh candidate evaluations. For each edge, the union count U is 3, 4, or 5 with probabilities $1/10$, $7/10$, and $2/10$, respectively. All these pair layouts are balanced. For the unbiased bounded contrast Z, use $\widehat{\theta} = UZ/(U+2)$. Its MSE is at most

$$A(1-\theta^2) + C\theta^2, \qquad A = \mathbb{E}\frac{U}{(U+2)^2} = \frac{6074}{55125}, \quad C = \mathbb{E}\frac{4}{(U+2)^2} = \frac{12139}{110250}.$$

Both coefficients are less than $1/9$. This proves feasibility uniformly over all bounded populations with hard work three. The fractional lower bound is $L\tau_f(K_5) = 5/2$. Since realized unit-cost

work is integer, a hard budget below three would permit at most two evaluations on every layout, contradicting this lower bound. The complete hard-work frontier in this example is therefore

$$\mathcal{W}^\star(R, 1/9) = \begin{cases} +\infty, & R = 0, \\ 4, & R = 1, \\ 3, & R \geq 2. \end{cases}$$

This conclusion uses the same marginal MSE contract for both randomized and deterministic designs.

H.5 A COMMON BOUNDED VECTOR FOR REFERENCE COMPARISONS

Under unit prices, evaluate only the star reference in $L \leq h$ fresh positions. Let $N = h + L$ and

$$\widetilde{\mu}_{\mathrm{ref}} = \frac{h - L}{h + L} \bar{Y}_{\mathrm{ref}}^{\mathrm{cache}} + \frac{2L}{h + L} \bar{Y}_{\mathrm{ref}}^{\mathrm{new}}, \qquad \widetilde{\mu}_i = \bar{Y}_i^{\mathrm{cache}} \quad (i \neq \mathrm{ref}).$$

Set $\widehat{\mu}_i = \frac{1}{2} + \frac{\sqrt{N}}{\sqrt{N}+1}(\widetilde{\mu}_i - \frac{1}{2})$. This vector belongs to $[0, 1]^K$, and every requested reference contrast has risk at most $m(N)$. It attains the minimum work even when its competitors may use more requests and arbitrary contrast estimators.

H.6 A REQUEST-OPTIMAL CONSTRUCTION AT EVERY PRECISION

Proposition 5 (Request-optimal refresh followed by cover completion). *Let $0 < d < 1$, let $h \geq 1$ be the number of full cached rows, and let C be any vertex cover of a nonempty graph $H = (V, F)$. Write $V_+ = \{i : \deg_H(i) > 0\}$, $C_+ = \sum_{i \in V_+} c_i$, and assume $C \subseteq V_+$. Set*

$$n_d = (d^{-1/2} - 1)^2, \qquad N = \max\{h, \lceil n_d \rceil\}, \qquad H_0 = \max\{h, \lceil N/2 \rceil\}.$$

Query every active candidate on $H_0 - h$ new rows, then query C on $N - H_0$ new rows. The resulting edgewise estimator family has uniform marginal MSE at most d, and its supplemental resources are

$$R = N - h, \qquad W_c = C_+(H_0 - h) + c(C)(N - H_0).$$

The request count R is minimal among all uniformly valid policies with a hard request cap, including cache-adaptive policies.

Proof. The refresh produces H_0 shared rows. Each requested edge meets C, so its observed-row union has exactly N rows. Each endpoint has either H_0 or N observations, and thus at least $N/2$. The balanced bounded-pair estimator attains $m(N) = (\sqrt{N} + 1)^{-2} \leq d$. Counting the two kinds of requests gives the displayed resources.

A table of $h + r$ full independent rows can simulate every policy with hard cap r: reveal the original cache, then the policy's chosen coordinates from the next unused row at each step, with its own independent randomness. On the submodel $(Y_i, Y_j) = (B, 1 - B)$, $B \sim \mathrm{Bernoulli}(p)$, with every other coordinate fixed, the full table has exact minimax risk $m(h+r)$ for $2p-1$. Hence any uniformly valid policy requires $h + r \geq \lceil n_d \rceil$, and therefore $r \geq (\lceil n_d \rceil - h)_+ = N - h$. □

I COMPLETE WORKLOADS AT ARBITRARY WORK BUDGETS

The entire complete-graph frontier. For $H = K_K$, write any integer work cap $w \geq 0$ as $w = qK + r$, where $0 \leq r < K$. Define

$$p_0 = \frac{(K - r)(K - r - 1)}{K(K - 1)}, \quad p_1 = \frac{2r(K - r)}{K(K - 1)}, \quad p_2 = \frac{r(r - 1)}{K(K - 1)}.$$

Theorem 12 (Complete-graph completion at every work budget). *For every $h \geq 1$, $K \geq 2$, and integer $w \geq 0$, the smallest marginal risk among precommitted lotteries with hard work at most w and any finite number of requests is*

$$\mathcal{C}_{K,h}(w) = \max_{t > 0} \frac{t}{t + 1} \sum_{j=0}^{2} \frac{p_j}{h + 2q + j + t}. \tag{61}$$

Give every candidate q private observations and a uniformly chosen r-subset one more private observation each. This hard-budget lottery attains the value. Thus $\mathcal{W}^\star(\infty, d) = \min\{w : \mathcal{C}_{K,h}(w) \le d\}$.

The same balanced allocation minimizes every scalar Bayes bound; a uniform relabeling attains its edgewise union-count law. The proof below allows arbitrary competing precommitted layouts and finite request counts.

We prove Theorem 12 without imposing a request cap or balance condition on the competing acquisition lists.

Fix $t > 0$. The fixed-edge Bernoulli prior in Appendix E gives, for every estimator family,

$$\sup_P \max_e \mathbb{E}(\delta_e - \theta_e)^2 \ge \max_e \frac{t}{t+1}\mathbb{E}\frac{1}{U_e + t} \ge \frac{t}{t+1}\mathbb{E}\left[\binom{K}{2}^{-1}\sum_{i<j}\frac{1}{h + x_i + x_j + t}\right].$$

The second bound uses $U_{ij} \le h + x_i + x_j$. Each realized layout has integer $x_i \ge 0$ with $\sum_i x_i \le w$. Adding any unused work weakly decreases the expression, so it is enough to minimize over allocations with sum w.

Let $f(s) = 1/(h + s + t)$ and $\Delta(s) = f(s+1) - f(s)$. The function f is decreasing and strictly convex, hence Δ is strictly increasing. If $x_i \ge x_j + 2$, move one observation from i to j. The (i, j) term is unchanged. For each other vertex ℓ, the change is

$$f(x_i - 1 + x_\ell) + f(x_j + 1 + x_\ell) - f(x_i + x_\ell) - f(x_j + x_\ell) = -\Delta(x_i - 1 + x_\ell) + \Delta(x_j + x_\ell) \le 0.$$

Repeating these transfers leaves counts q or $q + 1$, where $w = qK + r$ and precisely r candidates have the larger count. The average over all unordered edges becomes

$$\frac{p_0}{h + 2q + t} + \frac{p_1}{h + 2q + 1 + t} + \frac{p_2}{h + 2q + 2 + t}.$$

This minimum is the same for every initial layout. It therefore lower-bounds its expectation under every hard-budget lottery. Maximizing over t proves the lower bound in (61). The balancing allocation is independent of t; no interchange between incompatible optimizing allocations is needed.

For attainment, choose the r candidates uniformly and give all additional observations private positions, shared by no other candidate. Work is deterministically $qK + r = w$; w requests suffice. The union-count distribution for every fixed edge is exactly the three-point law above. Each endpoint has count $h + q$ or $h + q + 1$ and overlap h, so its count difference is at most one and the layout is balanced for every $h \ge 1$. Theorem 4 attains the common risk in (61). This completes the proof.

For $K = 2$ the transfer inequalities involving other vertices are empty, but the sum $x_1 + x_2 = w$ already fixes the union bound. The same balanced construction and formula remain valid. As $w \to \infty$ the attained risk tends to zero, so the minimum integer budget for every $d > 0$ exists.

The one-request complete-graph frontier. For a single fresh position and integer $0 \le w \le K$, put $\pi_0 = (K - w)(K - w - 1)/[K(K - 1)]$. The exact risk is

$$\mathcal{R}^{(1)}_{K,h}(w) = \max_{t>0} \frac{t}{t+1}\left\{\frac{\pi_0}{h + t} + \frac{1 - \pi_0}{h + 1 + t}\right\}.$$

To prove necessity, any queried subset of size at most w misses at least $\binom{K-w}{2}$ edges. For each fixed t, averaging its edgewise Bayes bounds therefore gives at least the displayed two-point mixture. A uniform w-subset has exactly that union-count distribution on every edge, and all its layouts are balanced because $h \ge 1$. Theorem 4 proves attainment. For targets below $m(h + 1)$, one fresh position is infeasible regardless of work. This formula supplies the one-request curves later reported in Figure 5; see Appendix P.

Uniform comparison with variance-only risk. Every union count satisfies $U_e \ge h$. Choosing $t = \sqrt{h}$ in (23) and using $U_e + \sqrt{h} \le U_e(1 + 1/\sqrt{h})$ gives

$$\left(\frac{\sqrt{h}}{\sqrt{h} + 1}\right)^2 \mathbb{E}\frac{1}{U_e} \le g(\nu_e) \le \mathbb{E}\frac{1}{U_e}.$$

The upper bound follows by using the unshrunk contrast. Thus the range correction vanishes relatively as h grows, uniformly over balanced lotteries. At a finite precision threshold it can still eliminate many additional observations, as the exact cache threshold (4) shows.

J FINITE-REQUEST DESIGNS AND RESOURCE CERTIFICATES

Let $H = (V, F)$ be a nonempty finite graph, $c_i > 0$, and $c(A) = \sum_{i \in A} c_i$. Throughout this section $1 \le L \le h$ are integers and $d = m(h + L)$, where $m(u) = (\sqrt{u} + 1)^{-2}$. Write $r(S)$ for the number of nonempty fresh requests, $x_i(S)$ for fresh marginal counts, and $W_c(S) = \sum_i c_i x_i(S)$. The ordinary risk averages over data and outcome-blind layout randomness. The fixed-edge binary subfamily gives the necessary conditions

$$\mathbb{E}[U_{ij} - h] \ge L, \qquad \mathbb{E}x_i + \mathbb{E}x_j \ge L, \qquad \mathbb{E}r \ge L. \tag{62}$$

These are expectations; a feasible marginal lottery need not satisfy any of these inequalities on each realized layout.

For any deterministic layout with $x_i \le L$ and $|B_i \cup B_j| \ge L$ on every edge, $|x_i - x_j| \le L \le h$ guarantees balance. The bounded-score estimator therefore has edge risk at most $m(h + L)$. This elementary sufficient condition certifies all the deterministic constructions below.

J.1 COMPRESSING THE FRACTIONAL-COVER ALLOCATION

Take a half-integral optimal fractional vertex cover $z \in \{0, \frac{1}{2}, 1\}^V$ and put

$$A = \{z_i = 1\}, \qquad B = \{z_i = \tfrac{1}{2}\}, \qquad Z = \{z_i = 0\}.$$

There are no edges inside Z or between B and Z. Let $D \subseteq B$ be a set of upgraded vertices, put $B' = B \setminus D$, and color $H[B']$ properly using q color classes. Costs of a selected cover and a coloring are directly verifiable, regardless of how they were found.

Proposition 6 (Finite-request implementation)**.** *If $L = 2k$, there is a deterministically certified layout with*

$$W_c = L\tau_{f,c}(H) + \frac{L}{2}c(D), \qquad r \le \max\{L, kq\}.$$

If $L = 2k + 1$ and C is any vertex cover of the full graph H, there is such a layout with

$$W_c = (L-1)\tau_{f,c}(H) + \frac{L-1}{2}c(D) + c(C), \qquad r \le \max\{L, kq + 1\}.$$

At $D = \varnothing$, an exact minimum cover C attains the deterministic work optimum. For even L, no integral cover computation is needed to attain the fractional work optimum.

Proof. For even L, give each color class a separate block of k fresh positions, querying every member of that class on every position in its block. Adjacent vertices in B' have disjoint blocks, so their union has size $2k = L$. Query all vertices in $A \cup D$ on any common L positions of the full schedule; append positions if fewer than L were available. These full-count vertices may share positions with the color blocks. Every edge incident to one already has a union of size at least L. Edges incident to Z end in A. Thus all edges meet the required union bound, and all marginal counts are at most L. The count and cost formulas follow.

For odd L, construct an even-demand $L - 1$ layer from the same k-position color blocks, overlaying $L - 1$ observations of $A \cup D$. Append one new request containing the global cover C. Each edge has at least $L - 1$ union positions in the first layer and gains one in the bonus request. The total request count is at most $\max\{L - 1, kq\} + 1 = \max\{L, kq + 1\}$. Each marginal count is at most L. When $L = 1$, the even layer is empty and only the cover request is used.

The risk claim follows from the sufficient condition above. For $D = \varnothing$, the even work is $L\tau_{f,c}$. For odd L, the even layer costs $(L - 1)\tau_{f,c}$ and an exact cover contributes $\tau_c(H)$. The resulting counts are $(L-1)z + \mathbf{1}_C$, so this construction uses no compatibility assumption between the chosen optimal fractional and integral covers. □

A greedy proper coloring and the threshold-rounded global cover $C = \{i : z_i \ge 1/2\}$ are polynomially constructible. Since $c(C) \le 2\tau_{f,c}(H)$, at $D = \varnothing$ this gives

$$W_c \le (L + 1)\tau_{f,c}(H) \quad \text{for odd } L.$$

The resulting request count is controlled by the coloring of the half-valued subgraph, rather than by the total number of candidate observations.

For even L, a particularly simple finite-request portfolio retains q color classes of a fixed coloring and upgrades the other classes into D. At a fixed q, retaining the classes with largest total cost minimizes the displayed work. Recoloring the retained subgraph can further reduce requests; every resulting schedule is independently certified from its explicit signatures.

J.2 A JOINT-PRICE GUARANTEE AND STRONGER INSTANCE CERTIFICATES

For $a \geq 0$, define the hard joint-price objective by

$$\mathcal{J}^\star(a) := \mathcal{J}^\star(a, d; c) = \inf_{\nu:\, \mathcal{R}_{\mathrm{bd}}(\nu) \leq d} \operatorname*{ess\,sup}_{S \sim \nu} \{ar(S) + W_c(S)\}.$$

Requests and work may vary in opposite directions across the supported layouts. Separate hard request and work caps define the Pareto frontiers studied below.

Proposition 7 (A scalable guarantee)**.** *For every graph,*

$$\mathcal{J}^\star(a) \geq L\{a + \tau_{f,c}(H)\}.$$

Rounding a half-integral fractional cover at $1/2$ and repeating it L times gives a deterministic plan with

$$J_{\mathrm{cover}} \leq L\{a + 2\tau_{f,c}(H)\}, \qquad \frac{J_{\mathrm{cover}}}{\mathcal{J}^\star(a)} \leq \frac{a + 2\tau_{f,c}(H)}{a + \tau_{f,c}(H)} \leq 2.$$

Selecting the cheapest certified plan from a portfolio containing this cover plan preserves the guarantee.

Proof. Any hard joint cap is at least the expected charge. Equation (62) and fractional-cover duality give

$$J \geq a\mathbb{E}r + \mathbb{E}W_c \geq La + L\tau_{f,c}(H).$$

No pointwise request or work lower bound is used. The threshold-rounded cover has cost at most $2\tau_{f,c}(H)$ and meets every edge in each of its L requests, proving the upper bound. □

Proposition 8 (Finite-request clique cuts)**.** *Let C be any s-vertex clique in H. Every feasible ordinary marginal-risk design satisfies, for $j = 0, \ldots, s-2$,*

$$(s - j - 1) \sum_{i \in C} \mathbb{E}x_i + \frac{j(j+1)}{2} \mathbb{E}r \;\geq\; L\binom{s}{2}. \tag{63}$$

If $r \leq R$ is imposed separately, $\mathbb{E}r$ may be replaced by R on the left.

Proof. A request containing v vertices of C hits

$$f_s(v) = \binom{s}{2} - \binom{s - v}{2} = \frac{v(2s - v - 1)}{2}$$

of its edges. For integer $v \in \{0, \ldots, s\}$, discrete concavity gives the supporting-line inequality

$$f_s(v) \leq (s - j - 1)v + \frac{j(j+1)}{2}.$$

The line is exact at $v = j, j+1$; directly subtracting $f_s(v)$ gives $(v - j)(v - j - 1)/2 \geq 0$. Sum over the nonempty requests and take expectations. The left total number of edge hits is at least $L\binom{s}{2}$ by (62). □

For any supplied family of cliques, minimize $ar + \sum_i c_i x_i$ over nonnegative real r, x, subject to $r \geq L$, $x_i + x_j \geq L$ on every edge, and the cuts (63). This small linear relaxation lower-bounds $\mathcal{J}^\star(a)$ even when candidate layouts may be unbalanced. Overlapping cliques may be used: they share the same vertex-count variables and do not double-charge the objective.

For an independently checkable dual certificate, choose nonnegative weights η_e, $\lambda_{C,j}$, and ξ such that

$$\sum_{e \ni i} \eta_e + \sum_{\substack{C \ni i \\ j}} \lambda_{C,j}(s_C - j - 1) \le c_i, \qquad \xi + \sum_{C,j} \lambda_{C,j} \frac{j(j+1)}{2} \le a.$$

Then a valid joint-price lower bound is

$$L \left\{ \xi + \sum_e \eta_e + \sum_{C,j} \lambda_{C,j} \binom{s_C}{2} \right\}. \tag{64}$$

For rational costs and price, all inequalities can be verified with rational arithmetic. A feasible plan divided by this lower certificate yields an instance-specific approximation ratio.

A weighted single-clique certificate. Write $W_{c,C} = \sum_{i \in C} c_i x_i$ for work on clique C. For a clique of size s, sort its vertex costs increasingly and let C_v be the sum of its v smallest costs. The points $(f_s(v), C_v)$ for $v = 0, \dots, s-1$ have increasing successive slopes $c_{(v)}/(s-v)$. Their linear interpolation is a convex function ϕ_C of edge hits. With a hard request cap R,

$$\mathbb{E} W_{c,C} \ge R\, \phi_C \left(\frac{L\binom{s}{2}}{R} \right).$$

The expression is infeasible if $R < L$. It is exact for this aggregate relaxation, not necessarily for the original statistical design problem. Bounds from vertex-disjoint cliques may be added. For overlapping cliques, use the joint-count LP or a valid dual allocation instead.

Continuous expected-request optimization. For vertex-disjoint cliques $\mathcal{C}$, define the convex piecewise-linear functions

$$\psi_C(\rho) = \rho\, \phi_C \left(\frac{L\binom{|C|}{2}}{\rho} \right), \qquad \rho \ge L.$$

The weighted clique bound also holds at $\rho = \mathbb{E} r$, without a hard request cap. Indeed, expected numbers of requests selecting each clique cardinality are nonnegative real variables; their sum is ρ, their edge-hit expectation is at least $L\binom{|C|}{2}$, and convexity gives the same perspective bound. Consequently the joint-price lower bound is

$$\mathcal{J}^\star(a) \;\ge\; \inf_{\rho \in \mathbb{R},\ \rho \ge L} \left[a\rho + \max \left\{ L\tau_{f,c}(H),\ \sum_{C \in \mathcal{C}} \psi_C(\rho) \right\} \right]. \tag{65}$$

The request variable is real: averaging an integer request count over a lottery need not give an integer. Minimizing only over an integer grid can increase the right side and is therefore not a valid joint-price lower certificate.

For rational costs and price, an exact computation can enumerate the endpoints $\rho = L$, the clique breakpoints $\rho = L\binom{s}{2}/f_s(v)$ for $v = 1, \dots, s-1$, and any intersections of the summed affine pieces with $L\tau_{f,c}(H)$. The function is constant in work beyond the largest clique breakpoint; the remaining price term is nondecreasing. Each candidate lies in a known affine interval, so checking its interval membership and rational objective value certifies the infimum. The separate hard-cap bound instead fixes an integer R and evaluates the work lower bound at that R; these are distinct operations.

J.3 A COMPLETE FINITE-REQUEST EXAMPLE

Theorem 13 (Triangle request–work frontier)**.** *On K_3 with unit costs and integer request cap R,*

$$\mathcal{W}^\star(R, m(h+L)) = \begin{cases} +\infty, & R < L, \\ \max\{\lceil 3L/2 \rceil,\ 3L - R\}, & R \ge L. \end{cases}$$

The value is attained by a deterministic layout with every edge union exactly $h + L$. The lower bound permits randomized and unbalanced competing layouts.

Proof. The universal bounds give $\mathbb{E}r \geq L$ and $\mathbb{E}W \geq 3L/2$. A hard unit-work cap is an integer after taking its floor, so it is at least $\lceil 3L/2 \rceil$. A nonempty triangle request hits at most one plus its size edges. Summing expected hits yields $\mathbb{E}r + \mathbb{E}W \geq 3L$. Hard caps therefore give $W \geq 3L - R$.

For attainment, set $t = \min\{R - L, \lfloor L/2 \rfloor\}$ and $u = L - 2t$. Give each vertex t private singleton positions and query a fixed pair at u further positions. This uses $r = 3t + u = L + t \leq R$ and $W = 3t + 2u = 2L - t$. Every edge receives $2t + u = L$ fresh union positions. The largest count difference is at most $L \leq h$, so the bounded-score certificate applies. The displayed work equals the claimed lower bound. □

For even L, the hard joint-price optimum is also exact:

$$\mathcal{J}^\star(a) = \begin{cases} \frac{3}{2}L(a+1), & 0 \leq a \leq 1, \\ L(a+2), & a \geq 1. \end{cases}$$

Indeed, the expectation constraints $\mathbb{E}W \geq 3L/2$, $\mathbb{E}r \geq L$ and $\mathbb{E}(W + r) \geq 3L$ give these two linear lower bounds. The deterministic frontier attains them. For odd L, the exact separate-cap frontier supplies a certified portfolio for choosing a priced plan.

J.4 Two certified Pareto points on a six-model workload

Let $K = 6$, $h = 192$, $L = 33$, and $d = m(225) = 1/256$. The minimum-request endpoint is $(R, W) = (33, 165)$. We certify two additional Pareto points under separate hard request and unit-work caps.

For any K and any realized layout, if its nonempty request sizes are s_t, then

$$\frac{1}{\binom{K}{2}} \sum_{i<j} U_{ij} = h + \frac{2W}{K} - \frac{\sum_t s_t(s_t - 1)}{K(K-1)} \leq h + \frac{2W}{K} - \frac{2(W - r)}{K(K-1)}.$$

For $K = 6$ this gives

$$\frac{1}{15} \sum_{i<j} \mathbb{E}U_{ij} \leq 192 + \frac{4\mathbb{E}W + \mathbb{E}r}{15}. \tag{66}$$

The inequality uses $s(s-1) \geq 2(s-1)$ for every nonempty request. Together with required mean union at least 225, it implies $4\mathbb{E}W + \mathbb{E}r \geq 495$.

Deterministic work 101 needs exactly 91 requests. Choose five heavy vertices and one light vertex. Query each of the ten pairs among the heavy vertices once, each heavy vertex alone thirteen times, and the light vertex alone sixteen times. There are $R = 10 + 65 + 16 = 91$ requests and $W = 20 + 65 + 16 = 101$ candidate evaluations. Heavy counts are seventeen, the light count is sixteen, heavy pairs overlap once, and heavy–light pairs do not overlap. Every edge consequently has 33 fresh union positions. The layout is balanced and attains risk $1/256$. A cap $R \leq 90, W \leq 101$ contradicts (66), proving request optimality at this work cap.

Randomized work 100 needs exactly 96 requests. Choose a uniform four-vertex heavy subset and one of its three undirected Hamilton cycles uniformly. Query each of the four cycle edges once, each heavy vertex alone fifteen times, and each of the other two vertices alone sixteen times. These 45 equally likely layouts all have $R = 4 + 60 + 32 = 96$ and $W = 8 + 60 + 32 = 100$. On every fixed edge, the total union is 224, 225, or 226 with probabilities $1/15, 12/15, 2/15$. All layouts are balanced. Using the common shrinkage parameter $t = 15$, the variance and squared-bias coefficients are

$$A = \mathbb{E}\frac{U}{(U+15)^2} = \frac{12437979907}{3184939008960}, \qquad C = \mathbb{E}\frac{225}{(U+15)^2} = \frac{4144795201}{1061646336320}.$$

Both are strictly below $1/256$, which is an exact rational feasibility certificate.

If $R \leq 95, W \leq 100$, (66) makes the average mean union at most 225. Each edge must have mean union at least 225, so all means equal 225. Strict Jensen in the common-prior bound then forces $U_e = 225$ almost surely for every edge. Every supported layout must therefore satisfy $x_i + x_j \geq 33$ on K_6, whose integer multicover minimum is $32 \cdot 3 + 5 = 101$, contradicting the hard work cap 100. This proves that 96 is the minimum request cap at work 100.

Work below 99 is excluded by the fractional lower bound. At hard work 99, averaging $U_{ij} \le h + x_i + x_j$ over edges gives an upper bound of 225, and the same strict-Jensen and integer-multicover argument excludes feasibility, regardless of request cap. Thus work 100 is globally minimal. The three displayed points are Pareto efficient under separate resource caps.

For the portfolio containing precisely $(33, 165), (91, 101), (96, 100)$, minimizing $aR + W$ selects the randomized point below $a = 1/5$, the deterministic compressed point between $1/5$ and $32/29$, and the minimum-request point above $32/29$ (with ties at the thresholds). These thresholds select among the three supplied plans. The global hard joint-price optimization also permits lotteries that vary requests and work in opposite directions.

K Exact planning over the full finite layout space

At an integer endpoint with even deficit $1 \le L \le h$, a half-integral fractional-cover solution followed by private acquisition is work-optimal when requests are unrestricted. With only L requests, rounding any fractional cover at $1/2$ and repeating that cover gives work at most $2L\tau_{f,c}$. For targets between endpoints and finite request allowances, the full layout risk supplies a direct optimization program.

Fix hard caps R, w, with $R \le h$, and let $\mathcal{S}_{R,w}$ contain every padded request list satisfying those caps. Define $a_{eS}(t) = t/[(t+1)(U_e(S)+t)]$. The exact best marginal risk is the semi-infinite LP

$$\min_{p,z} z \quad \text{s.t.} \quad \sum_{S \in \mathcal{S}_{R,w}} p_S a_{eS}(t) \le z \quad (e \in F,\ t > 0), \qquad \sum_S p_S = 1, \quad p_S \ge 0. \tag{67}$$

The layout space is finite, whereas each edge has a scalar continuum of constraints. For a proposed p, its most violated constraint is found by the unique root (25). An exchange method solves a finite set of constraints, adds the maximizing scalar for each edge, and repeats.

A certificate independent of the scalar grid. For rational probabilities p_S and a positive rational shrinkage parameter t_e, the estimator has the rigorous upper bound

$$u_e = \max\left\{ \sum_S p_S \frac{U_e(S)}{(U_e(S)+t_e)^2},\ \sum_S p_S \frac{t_e^2}{(U_e(S)+t_e)^2} \right\}. \tag{68}$$

For nonnegative rational weights $y_{e,t}$ on finitely many scalar constraints, summing to one, a lower bound is

$$\min_{S \in \mathcal{S}_{R,w}} \sum_{e,t} y_{e,t} a_{eS}(t). \tag{69}$$

Both bounds can be checked with exact fractions. An integer work optimum is certified by an upper bound at most d at that budget and a strict lower bound above d at the preceding budget. Our small-instance experiments enumerate the entire layout space, up to request permutation and identical risk columns. On larger graphs a restricted library supplies an achievable cost, while the graph lower bounds remain valid.

L Supplementary population-adaptive acquisition

The main results concern the hard reserve required for uniform validity. Cache-dependent acquisition can have smaller population-specific expected spending while retaining such a reserve. This section gives a variance screen that couples a zero-fresh-work branch to a uniformly valid fallback, with the target and held-out cache size defined below.

We calibrate a cache-only stopping decision by its contribution to expected squared error, while reserving a uniformly valid fallback. Fix a nonempty comparison graph $H = (V, F)$ before using the cache values. The cache consists of h independent full vectors with an unknown common law P on $[0,1]^V$. For an oriented edge $e = (i, j)$, write

$$D_e = Y_i - Y_j, \qquad \theta_e = \mathbb{E}_P D_e, \qquad v_e = \mathrm{Var}_P(D_e) \le 1.$$

Risk is $\max_{e \in F} \mathbb{E}_P(\widehat{\theta}_e - \theta_e)^2$, with the expectation over the cache, future observations, and policy randomization. The graph and the tuning parameters are outcome-blind; the decision to acquire more data can depend on the pilot.

Choose an integer $m_{\mathrm{p}} \geq 1$ with $h_{\mathrm{est}} = h - 2m_{\mathrm{p}} \geq 1$. On the first $2m_{\mathrm{p}}$ cached rows form m_{p} independent paired variance observations

$$Z_{r,e} = \frac{(D_{2r,e} - D_{2r-1,e})^2}{2} \in [0,2], \qquad \overline{Z}_e = \frac{1}{m_{\mathrm{p}}} \sum_{r=1}^{m_{\mathrm{p}}} Z_{r,e}. \tag{70}$$

They satisfy $\mathbb{E}_P Z_{r,e} = v_e$. For a positive cutoff ζ, let

$$A_\zeta = \{\overline{Z}_e \leq \zeta \text{ for every } e \in F\}. \tag{71}$$

On A_ζ, return the paired contrast means from the remaining h_{est} cached rows and acquire nothing. On A_ζ^c, execute a fallback B with uniform marginal risk at most γd, where $0 < \gamma < 1$. The fallback uses only these h_{est} held-out rows, fresh observations, and independent randomization. It can share the held-out rows with the successful-branch estimator; both experiments must be independent of the first $2m_{\mathrm{p}}$ rows.

Theorem 14 (A risk-calibrated variance screen)**.** *Fix $0 < d < 1$, $\lambda > 0$, $0 < \gamma < 1$, and $m_{\mathrm{p}}, h_{\mathrm{est}}$ as above. Suppose $b = \gamma d h_{\mathrm{est}} < 1$, and define*

$$\psi_\lambda = \frac{1 - e^{-2\lambda}}{2}, \qquad v_\star = \min\left\{1,\ b + \frac{1 - \psi_\lambda b}{\psi_\lambda (m_{\mathrm{p}} + 1)}\right\}, \tag{72}$$
$$G_{m_{\mathrm{p}}}(\psi_\lambda, b) = (v_\star - b)(1 - \psi_\lambda v_\star)^{m_{\mathrm{p}}}.$$

If $\zeta > 0$ satisfies

$$e^{\lambda m_{\mathrm{p}} \zeta} G_{m_{\mathrm{p}}}(\psi_\lambda, b) \leq (1 - \gamma) d h_{\mathrm{est}}, \tag{73}$$

then the cache-adaptive policy satisfies uniform ordinary marginal risk at most d. A sufficient positive cutoff is any

$$0 < \zeta \leq \frac{\log((1-\gamma) d h_{\mathrm{est}} / G_{m_{\mathrm{p}}}(\psi_\lambda, b))}{\lambda m_{\mathrm{p}}}. \tag{74}$$

Its hard supplemental resource caps are the fallback's caps. For every nonnegative acquisition resource C_{B},

$$\mathbb{E}_P C_{\mathrm{gate}} = \Pr_P(A_\zeta^c)\, \mathbb{E}_P C_{\mathsf{B}}. \tag{75}$$

When $b \geq 1$, the held-out mean already has uniform risk at most γd, and the policy may return it directly.

Proof. For $Z \in [0,2]$, convexity places $e^{-\lambda Z}$ below its endpoint secant. Therefore

$$\mathbb{E}_P e^{-\lambda Z_{r,e}} \leq 1 - \psi_\lambda v_e.$$

Markov's inequality and independence of the paired pilot blocks give

$$\Pr_P(A_\zeta) \leq \Pr_P(\overline{Z}_e \leq \zeta) \leq e^{\lambda m_{\mathrm{p}} \zeta} (1 - \psi_\lambda v_e)^{m_{\mathrm{p}}}. \tag{76}$$

The h_{est}-row contrast mean is independent of the gate and has risk v_e / h_{est}. The fallback is also independent of the gate and has risk at most γd, so the unconditional edge risk is at most

$$\Pr_P(A_\zeta) \frac{v_e}{h_{\mathrm{est}}} + (1 - \Pr_P(A_\zeta)) \gamma d = \gamma d + \Pr_P(A_\zeta) \frac{v_e - b}{h_{\mathrm{est}}}. \tag{77}$$

For $v_e \leq b$, this is at most γd. For $v_e > b$, apply (76). Differentiation of $f(v) = (v - b)(1 - \psi_\lambda v)^{m_{\mathrm{p}}}$ yields a positive derivative up to $b + (1 - \psi_\lambda b)/(\psi_\lambda (m_{\mathrm{p}} + 1))$ and a negative derivative afterwards. Consequently its maximum on $[b, 1]$ is exactly $G_{m_{\mathrm{p}}}(\psi_\lambda, b)$. Condition (73) now gives risk at most d.

The proof applies separately to every fixed edge. The event requiring all edges to pass is contained in each individual passing event, so the risk bound has no graph-size multiplier. The resource identity follows by independence of the pilot and the fallback experiment; hard caps follow because the successful branch uses no fresh resources. If $b \geq 1$, $v_e / h_{\mathrm{est}} \leq 1 / h_{\mathrm{est}} \leq \gamma d$. □

This calibration uses the risk contribution of a mistaken stop directly. Its statistical ingredients are the bounded-variable Chernoff bound and an independent estimation split. The resource connection is explicit: the same uniformly valid acquisition design supplies a hard reserve while the screen determines whether that reserve is used.

Proposition 9 (Positive-variance and nonatomic stopping classes)**.** *Under Theorem 14, suppose* $0 < \zeta < 2$*. For* $0 \le v \le 1$*, define*

$$\psi_{m_{\mathrm{p}},\zeta}(v) = \begin{cases} 0, & v = 0, \\ \min\{v/\zeta,\, \exp[-m_{\mathrm{p}}\ \mathrm{kl}(\zeta/2 \,\|\, v/2)]\}, & 0 < v < \zeta, \\ 1, & v \ge \zeta, \end{cases}$$

where $\mathrm{kl}(x\|y) = x\log(x/y) + (1-x)\log((1-x)/(1-y))$*. Then*

$$\Pr_P(A_\zeta^c) \le \min\left\{1, \sum_{e\in F} \psi_{m_{\mathrm{p}},\zeta}(v_e)\right\}. \tag{78}$$

In particular the expected resource use tends to zero continuously as the edge variances tend to zero, for a fixed graph and fixed positive cutoff.

If each requested contrast is supported on an interval of width at most $\sqrt{2\zeta}$*, then* A_ζ *occurs almost surely. Thus the policy acquires nothing on this entire class, including continuous populations with positive contrast variances and unknown means.*

Proof. Markov's inequality gives $\Pr_P(\overline{Z}_e > \zeta) \le v_e/\zeta$. For $v_e < \zeta$, apply the upper-tail bounded-variable Chernoff bound to $Z_{r,e}/2 \in [0,1]$, whose mean is $v_e/2$. Optimizing its exponential parameter gives the displayed binary-KL bound. A union bound controls the event that any edge fails. This graph-size dependence concerns acquisition probability; the risk guarantee in Theorem 14 remains per edge.

A contrast of support width w_e has $Z_{r,e} \le w_e^2/2$. If $w_e^2/2 \le \zeta$ on all edges, every pilot block passes deterministically. □

For example, let U_i be independent uniform variables on $[-1,1]$ and set $Y_i = \mu_i + \sigma U_i$, with arbitrary unknown $\mu_i \in [\sigma, 1-\sigma]$. The joint law has a density when $\sigma > 0$, each contrast has variance $2\sigma^2/3$, and

$$8\sigma^2 \le \zeta \quad \Longrightarrow \quad \Pr_P(A_\zeta) = 1.$$

The returned contrasts still have sampling error $2\sigma^2/(3h_{\mathrm{est}})$, which the theorem controls. A shared bounded random component can also be added to all coordinates, with the means chosen to keep scores in $[0,1]$. It cancels from the contrasts and leaves the gate unchanged. Thus a positive tolerance can exploit similarity of candidate responses even when their common variation is substantial.

L.1 Finite hard reserves

Let $c_i > 0$, $C_+ = \sum_{i:\deg_H(i)>0} c_i$, and let C be a vertex cover with cost $c(C) = \sum_{i\in C} c_i$. Write $n_u = (u^{-1/2} - 1)^2$ and $m(N) = (\sqrt{N}+1)^{-2}$. The following fallback is valid for every cache size and precision:

$$N_{\mathsf{B}} = \max\{h_{\mathrm{est}}, \lceil n_{\gamma d}\rceil\}, \qquad H_{\mathsf{B}} = \max\{h_{\mathrm{est}}, \lceil N_{\mathsf{B}}/2\rceil\}.$$

Acquire $H_{\mathsf{B}} - h_{\mathrm{est}}$ full active-candidate vectors and then $N_{\mathsf{B}} - H_{\mathsf{B}}$ cover vectors. It has caps

$$R_{\mathsf{B}} = N_{\mathsf{B}} - h_{\mathrm{est}}, \qquad W_{\mathsf{B}} = C_+(H_{\mathsf{B}} - h_{\mathrm{est}}) + c(C)(N_{\mathsf{B}} - H_{\mathsf{B}}). \tag{79}$$

Every edge has union count N_{B}, and every endpoint has at least $N_{\mathsf{B}}/2$ observations. The balanced bounded-pair estimator therefore has risk $m(N_{\mathsf{B}}) \le \gamma d$. This gives a concrete reserve even when splitting the cache leaves the low-deficit regime. Equation (75) converts these caps to instance-dependent expected acquisition.

L.2 Logarithmic pilot overhead in the cache-rich regime

Corollary 2 (A positive cutoff with vanishing reserve overhead)**.** *Consider aligned targets* $d = m(N) \to 0$*, with integer* $N = h + L$*, and constants*

$$\frac{1}{2} < \beta_- < \beta_+ < 1, \qquad \beta_- \le hd \le \beta_+. \tag{80}$$

Fix $b_0 = \beta_-/2$ and any $0 < \zeta < b_0$. Choose $\lambda > 0$ small enough that

$$\kappa = -\log(1 - \psi_\lambda b_0) - \lambda\zeta > 0.$$

Set

$$\gamma = 1 - d, \qquad m_{\mathrm{p}} = \left\lceil \frac{\log(2/(\beta_- d))}{\kappa} \right\rceil, \qquad h_{\mathrm{est}} = h - 2m_{\mathrm{p}}. \tag{81}$$

For all sufficiently small d, the screen has uniform marginal risk at most d, and $h_{\mathrm{est}} \geq 1$. With

$$N_{\mathsf{B}} = \lceil n_{\gamma d} \rceil, \qquad s = 2m_{\mathrm{p}} + (N_{\mathsf{B}} - N), \qquad L_{\mathsf{B}} = L + s,$$

we have

$$N_{\mathsf{B}} - N = O(1), \quad s = O(\log(1/d)), \quad L_{\mathsf{B}} \leq h_{\mathrm{est}}, \quad \frac{s}{L} = O(d\log(1/d)). \tag{82}$$

The cutoff ζ remains a fixed positive constant as $d \to 0$.

These parameters permit the following reserve comparisons.

1. *Repeating the same cover C, the hard request and work caps increase by the exact factor $L_{\mathsf{B}}/L = 1 + s/L$.*
2. *With unrestricted requests, a polynomially constructible reserve has work at most $(L_{\mathsf{B}} + 1)\tau_{f,c}(H)$. Relative to the original optimal outcome-blind hard or expected work, its hard work ratio is at most $1 + (s+1)/L = 1 + O(d\log(1/d))$.*
3. *If an original deterministic planner is a ρ-approximation to the optimal outcome-blind hard total price $aR + W_c$, its cache-adaptive wrapper can retain $\rho + 2s/L = \rho + O(d\log(1/d))$. In particular, relative to the original optimal deterministic hard total price, a deterministic optimal layout admits a reserve factor at most $1 + 2s/L$.*

All these caps are reserves; they are used with probability $\Pr_P(A_\zeta^{\mathrm{c}})$. For $\max_e v_e \leq v_0 < \zeta$, this probability is at most

$$\min\{1,\ |F| \exp[-m_{\mathrm{p}}\,\mathrm{kl}(\zeta/2 \| v_0/2)]\}. \tag{83}$$

For the positive-width support class of Proposition 9, it is zero.

Proof. Since $\psi_\lambda/\lambda \to 1$ as $\lambda \downarrow 0$, $\kappa/\lambda \to b_0 - \zeta > 0$, so the stated λ exists. Since $m_{\mathrm{p}} d \to 0$, $hd \geq \beta_-$, and $\gamma \to 1$, eventually

$$h_{\mathrm{est}} \geq \frac{3\beta_-}{4d}, \qquad \gamma \geq \frac{2}{3}, \qquad b = \gamma d h_{\mathrm{est}} \geq b_0.$$

For an edge with $v_e \leq b$, (77) is at most γd. For $v_e > b$, the passing probability is at most

$$e^{\lambda m_{\mathrm{p}} \zeta}(1 - \psi_\lambda b_0)^{m_{\mathrm{p}}} = e^{-\kappa m_{\mathrm{p}}} \leq \frac{\beta_- d}{2} \leq d^2 h_{\mathrm{est}}.$$

The excess risk in (77) is therefore at most $d^2 = (1-\gamma)d$.

Next,

$$0 \leq n_{(1-d)d} - n_d = \frac{1}{1-d} - \frac{2}{\sqrt{d}}\big((1-d)^{-1/2} - 1\big) \leq \frac{1}{1-d}.$$

Rounding adds at most one, so $0 \leq N_{\mathsf{B}} - N \leq 3$ for $d \leq 1/2$. The interval in (80) yields $L = \Theta(d^{-1})$ and $h - L = \Theta(d^{-1})$. It follows that

$$h - L \geq 4m_{\mathrm{p}} + (N_{\mathsf{B}} - N)$$

eventually, which is exactly $L_{\mathsf{B}} \leq h_{\mathrm{est}}$. This proves (82).

For the reserve comparisons, every original valid deterministic layout must have new signature union at least L on every edge: the antithetic Bernoulli submodel gives risk at least $m(h + |B_i \cup B_j|)$. Clipping any signature larger than L to L positions retains this union property and weakly decreases resources. Append s fresh repetitions of a cover. Every new count is now at most $L + s = L_{\mathsf{B}} \leq h_{\mathrm{est}}$, and every union, including the held-out cache, is at least $h_{\mathrm{est}} + L + s = N_{\mathsf{B}}$. The balanced estimator reaches the fallback target.

Table 4: Finite gates with $h = 192$.

d	$2m_{\mathrm{p}}$	h_{est}	γ	λ	ζ	R_0	R_{B}
$1/256$	6	186	0.9	2	0.0773	33	66
$1/1024$	48	144	0.9	2	0.0172	769	928
$1/4096$	96	96	0.4	2	0.0013	3777	9943

In particular cover repetition costs increase by $1 + s/L$. A half-integral fractional cover implements even deficits at cost $L_{\mathsf{B}}\tau_{f,c}$; for an odd deficit, its even predecessor plus a cover of cost at most $2\tau_{f,c}$ costs at most $(L_{\mathsf{B}} + 1)\tau_{f,c}$. The original outcome-blind lower bounds are $W^\star \geq L\tau_{f,c}$ and $J^\star \geq L(a+\tau_{f,c})$, also for randomized layouts under ordinary marginal risk. Appending s repetitions of a two-approximation cover costs at most $s(a + 2\tau_{f,c}) \leq (2s/L)J^\star$. These establish the stated relative comparisons. Finally, binary KL is decreasing in its second argument below $\zeta/2$; (83) follows from Proposition 9. □

The finite inequality $L_{\mathsf{B}} \leq h_{\mathrm{est}}$ applies to the relative cover and fractional-cover reserve comparisons. The general refresh reserve (79) and the gate theorem remain available outside that region. The following finite computations check the risk condition itself and report the actual reserve costs, including configurations where the reserve is substantially larger than the original plan.

Table 4 reports positive rational cutoffs certified by exact rational exponential enclosures. Here $R_0 = \max\{h, \lceil n_d \rceil\} - h$ is the original refresh–cover request count, and R_{B} is its independent fallback reserve (79). For a unit-cost 32-vertex star the corresponding work pairs are $33/66$, $9728/13080$, and $59360/162587$; for a 32-clique they are $1023/2046$, $24128/29160$, and $118880/313157$. These are the request–work pairs of the specified refresh–cover constructions.

For the first row, the nonatomic family $Y_i = \mu_i + (491/5000)U_i$, with independent $U_i \sim \mathrm{Unif}[-1, 1]$ and unknown $\mu_i \in [491/5000, 1 - 491/5000]$, satisfies $8(491/5000)^2 < 0.0773$. Every requested edge has positive variance $241081/37500000$, yet the gate accepts with probability one. Its retained-cache MSE is $241081/6975000000 < 1/256$. The hard reserve for this row is twice that of the original refresh–cover plan. The other rows similarly have positive-width certain-stop classes, with contrast support widths at most 0.1854 and 0.0509, respectively.

M Finite-frontier planning and independent certificates

The finite study uses a common cache of $h = 3$, request caps $R \in \{1, 2, 3\}$, and targets $1/9$, $1/8$, and $1/4$. The workloads are a three-edge star and matching on six candidates, a six-cycle, a five-cycle, and complete graphs on four, five, and six candidates.

Complete layout space and exchange optimization. For K candidates, a request is a subset of $[K]$; padded lists of at most R requests are enumerated up to permutation of request positions. Layouts with the same edge-union vector have identical risk coefficients and are represented by one column. For a cut (e, t), $t > 0$, define

$$B_{e,t}(S) = \frac{t}{(t+1)(U_e(S) + t)}.$$

The finite linear program minimizes a common risk epigraph over layout probabilities. For each edge, the most violated scalar constraint is obtained from the unique equalizer root and added iteratively.

Rational upper and lower certificates. For rational layout probabilities and any rational $t > 0$, define

$$A_e = \mathbb{E}\frac{U_e}{(U_e + t)^2}, \qquad C_e = \mathbb{E}\frac{t^2}{(U_e + t)^2}.$$

The estimator with coefficient $U_e/(U_e + t)$ has risk at most $\max\{A_e, C_e\}$, giving an exact rational upper certificate. If $y_\ell \geq 0$ are rational weights on finitely many cuts (e_ℓ, t_ℓ) with $\sum_\ell y_\ell = 1$, then every completion lottery with hard work at most W has risk at least

$$\min_{S:\, W(S) \leq W} \sum_\ell y_\ell \frac{t_\ell}{(t_\ell + 1)(U_{e_\ell}(S) + t_\ell)}.$$

A feasible certificate at work W together with a strict lower certificate at $W-1$ establishes the minimum integer work. These rational checks give the finite frontiers used in Appendix P.

Complete-graph formulas beyond enumeration. For $w = qK + r$, $0 \le r < K$, a uniform choice of r candidates receiving $q+1$ private observations, with all others receiving q, gives edge unions $h+2q$, $h+2q+1$, and $h+2q+2$ with probabilities

$$\frac{(K-r)(K-r-1)}{K(K-1)}, \quad \frac{2r(K-r)}{K(K-1)}, \quad \frac{r(r-1)}{K(K-1)}.$$

For one additional request, a uniform w-subset gives union h with probability $(K-w)(K-w-1)/(K(K-1))$ and union $h+1$ otherwise. The exact complete-graph risk formula from Appendix I applied to these laws produces the large-K curves in Figure 5.

N Supplementary frozen-score information diagnostics

The allocation comparators use partial observations, a classical theme in sampling design (Horvitz & Thompson, 1952; Cochran, 1977). The following studies compare range-only and covariance-informed plans on fixed empirical populations, separating the range-uniform reserve, allocation based on population information, and the resulting population-specific error.

Data and evaluation distribution. The frozen RouterBench and LLMRouterBench score matrices contain 36,481 rows and 11 models, and 14,716 rows and 20 models, respectively. We use their fixed query-hash evaluation partitions of 18,242 and 7,362 rows. Preprocessing keeps published scores in $[0,1]$, removes duplicate normalized queries by a score-independent order, and requires complete model coverage. A reference star uses column zero in the frozen ordering; clique workloads use every pair. The cache and fresh requests are independent draws with replacement from the evaluation rows, with all candidates on one physical request sharing the same draw. Resources are unit candidate evaluations W and physical requests R.

Range-only plans and exact actual error. Targets are $d \in \{1/256, 1/1024, 1/4096\}$ with $h = 192$. For full and cover completion, let $N = \max\{h, \lceil (1/\sqrt{d}-1)^2 \rceil\}$ and $N_{\text{full}} = \max\{h, \lceil N/2 \rceil\}$. Full completion uses $N-h$ complete vectors. Refresh–cover uses $N_{\text{full}}-h$ complete vectors and $N - N_{\text{full}}$ queries of a fixed minimum cover: the center of a star or the first $K-1$ clique vertices. Both employ the balanced unshrunk contrast followed by shrinkage $\sqrt{N}/(\sqrt{N}+1)$, with uniform bound $m(N)$. Actual risk is evaluated from the shared and private coefficients together with the squared shrinkage bias from the population mean contrast. Private precision allocation discards the cache and obtains independent singleton observations with $n_i = \lceil z_i/d \rceil$. For a clique, $z_i = 1/2$; for a star with M leaves, $z_0 = (1+\sqrt{M})/4$ and $z_i = (1+1/\sqrt{M})/4$ on leaves. Its coherent unshrunk mean estimates satisfy $\frac{1}{4}(1/n_i + 1/n_j) \le d$ and have exact actual risk $\Sigma_{ii}/n_i + \Sigma_{jj}/n_j$; $R = W = \sum_i n_i$.

Covariance-informed paired allocation. Write $v_e = \mathrm{Var}(Y_i - Y_j)$. The oracle uses these variances for planning and retains either all 192 cache rows or only 128 to match the pilot comparison; population means are never supplied to the planner or estimator. Given an estimation cache of size h_2, planning precision d_{plan}, and a planning variance V_e, it requires fresh paired counts $q_e = [\lceil V_e/d_{\text{plan}} \rceil - h_2]_+$. The fixed menu contains the full vector, all requested two-coordinate pairs and all leave-one-out vectors, removing duplicates and entries that contain no requested pair. It solves $\min \sum_S |S| x_S$ subject to $\sum_{S \supseteq e} x_S \ge q_e$, $x_S \ge 0$, then rounds counts upward and checks integer feasibility. The final plan is the lower-work choice between this rounded plan and $Q = \max_e q_e$ full-vector requests, with fewer requests breaking a work tie. The LP is a continuous relaxation within the fixed menu; the rounded output is a feasible integer plan. Each edge uses only its $h_2 + c_e$ paired contrast observations, giving an unbiased estimator with actual MSE $v_e/(h_2 + c_e)$. The population variances are evaluation-only information for the practical pilot methods.

Pilot separation and a marginal certificate. The primary split fixes 64 cache rows for the pilot and 128 independent cache rows for estimation; the sensitivity split is 96/96. The plug-in planner uses unbiased pilot contrast sample variances, capped at the public maximum one, with $d_{\text{plan}} = d$. The certified planner partitions the m pilot rows into $q_p = m/2$ disjoint pairs and computes

$$Z_{e,r} = \tfrac{1}{2}(D_{e,2r} - D_{e,2r-1})^2 \in [0,2], \qquad \overline{v}_e = \min\{1, \bar{Z}_e + \sqrt{2\log(1/\delta)/q_p}\}.$$

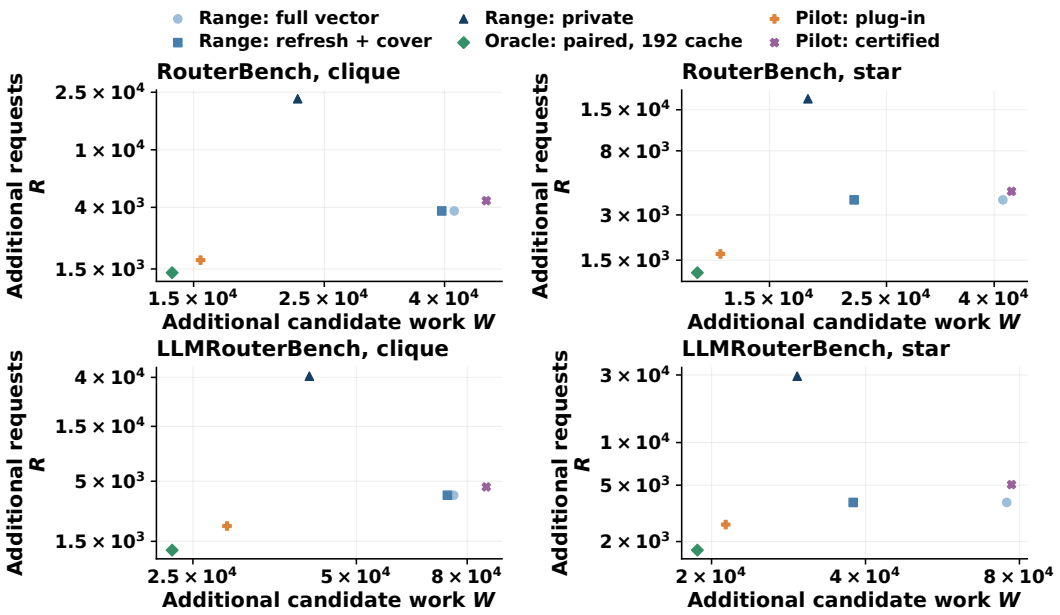

Figure 2: Work and requests on four fixed workloads.

Since $\mathbb{E}Z_{e,r} = v_e$, Hoeffding's inequality gives $\Pr(\overline{v}_e < v_e) \le \delta$. We fix $\alpha = 0.1$, $\delta = \alpha d h_2 < 1$, and plan with $V_e = \overline{v}_e$ and $d_{\text{plan}} = (1-\alpha)d$. All fresh queries are frozen after the pilot, whose observations never enter the final estimate. Conditional on any pilot, the estimator remains unbiased and its risk is $v_e/(h_2 + c_e) \le 1/h_2$. On $\overline{v}_e \ge v_e$ this risk is at most $(1-\alpha)d$, so for each edge

$$\mathbb{E}(\hat{\theta}_e - \theta_e)^2 \le (1-\alpha)d + \delta/h_2 = d.$$

Because the target is $\max_e \mathbb{E}\text{error}_e^2$, the edgewise certificate gives the required marginal bound directly. It averages the pilot, estimation cache, and fresh-score randomness for each fixed edge. Simultaneous and numerical-cache-conditional criteria are treated in Appendix D.

Random resources and numerical evaluation. For $V_e \le 1$, the full-vector fallback gives $W \le K\bar{Q}$ with $\bar{Q} = [\lceil 1/d_{\text{plan}} \rceil - h_2]_+$ and $R \le \lfloor K\bar{Q}/2 \rfloor$. At exact integer boundaries, quotas are rounded upward before ceilings are applied. For the certified 64/128 split, the three values of $\bar{Q}$ are 157, 1,010, and 4,424.

The primary analysis averages over 2,048 independent pilots; the 96/96 sensitivity analysis uses 512. Conditional on each pilot, holdout-cache and fresh-score risk is integrated analytically. Edgewise Monte Carlo standard errors quantify the remaining pilot integration error. Independent recomputation verifies pilot variances, integer feasibility, resource caps, and the fixed-design risk formulas to numerical precision.

Figure 2 reports work and requests at $d = 1/4096$, including the full request cost of private sampling. Pilot points average over the independent pilot draws. Figure 3 compares actual marginal MSE at all three targets, including bias and the randomness of pilot-selected plans.

N.1 Joint covariance-informed allocation and estimation

We compare acquisition and actual precision under a common query menu. Table 5 reports the full clique comparison, while Figure 4 adds the reference-star workloads. The same pilot split and true-covariance risk integration are used below.

We retain the score populations, query-hash evaluation split, graph identities and unit resource definitions of the preceding experiment. The primary comparison uses exactly the same full-vector, requested-pair and leave-one-out query menu, including its exclusion of subsets containing no requested pair. An additional oracle diagnostic augments the menu with all singletons and every leave-one-out subset, including the all-leaves query for a star. The optimization domain is the stated finite query menu.

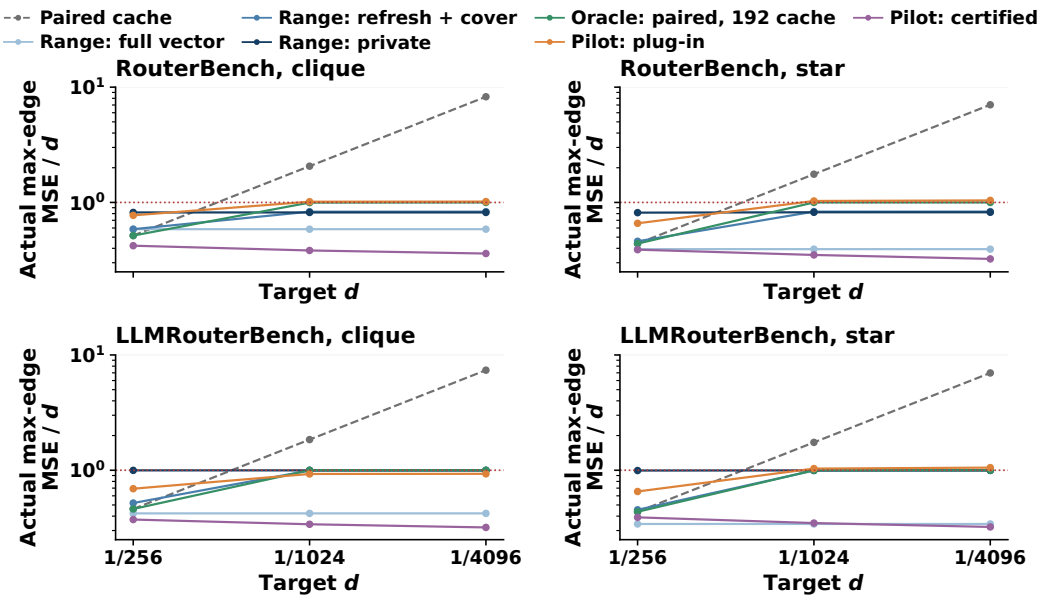

Figure 3: Actual marginal MSE across three precision targets.

Table 5: Clique refinement at $h = 192$ and $d = 1/4096$.

Plan	RouterBench, 11 models W	R	MSE/d (SE)	LLMRouterBench, 20 models W	R	MSE/d (SE)
Range: refresh + cover	39563	3777	0.832	73556	3777	0.996
Range: private	22528	22528	0.819	40960	40960	0.998
Paired oracle, 192	13792	1413	1.000	22888	1258	1.000
GLS oracle, 192	11224	3660	0.998	21092	2981	0.996
GLS oracle, 128	11972	3730	0.995	22392	3019	0.995
Pilot: LW paired	14964	1619	1.049 (0.006)	28825	1945	0.926 (0.004)
Pilot: LW GLS	12299	4169	1.076 (0.004)	24400	5857	1.068 (0.006)

Relation to MultiPPI. MultiPPI (Cowen-Breen et al., 2026) jointly optimizes subset sample sizes and linear weights, with minimax guarantees for known covariance and finite-sample analysis for estimated covariance. The comparator here applies this covariance-informed allocation principle to the fixed graph and query menu used in our study. Its pilot split estimates covariance independently of the final mean estimator; the reported pilot results refer to this comparator.

Common mean estimator. Let $G \succ 0$ denote the planning covariance, P_S select the coordinates in S, and h_2 denote the number of complete cached vectors reserved for estimation. For integer counts n_S, define

$$M_G = h_2 G^{-1} + \sum_S n_S P_S^\top G_{SS}^{-1} P_S,$$
$$\widehat{\mu} = M_G^{-1} \left[G^{-1} \sum_{r=1}^{h_2} Y_r + \sum_S P_S^\top G_{SS}^{-1} \sum_{r=1}^{n_S} Y_{S,r} \right]. \tag{84}$$

Here $G_{SS} = P_S G P_S^\top$, and $Y_{S,r}$ is the score vector observed on subset S at its rth fresh request. Different physical requests are independent. Every comparison uses $\widehat{\theta}_e = \ell_e^\top \widehat{\mu}$ with $\ell_e = \mathbf{e}_i - \mathbf{e}_j$, where $\mathbf{e}_i$ is a standard basis vector. For fixed G and counts, the coefficient matrices sum to the identity. Conditional on an independent pilot, $\widehat{\mu}$ is therefore unbiased for every population mean. For the true covariance Σ, its covariance is $M_G^{-1} B_G M_G^{-1}$, where

$$B_G = h_2 G^{-1} \Sigma G^{-1} + \sum_S n_S P_S^\top G_{SS}^{-1} \Sigma_{SS} G_{SS}^{-1} P_S.$$

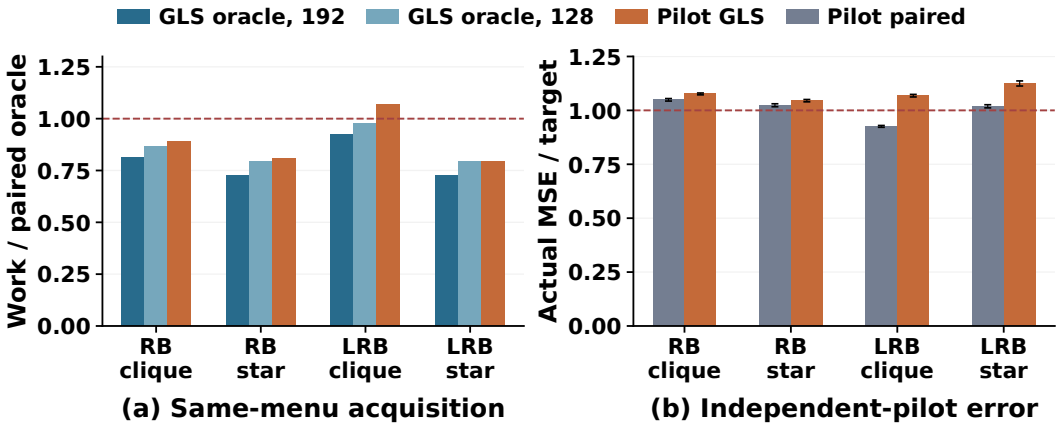

Figure 4: Joint weighting changes acquisition and finite-pilot error.

These expressions integrate the independent estimation cache and all fresh scores exactly. Only the oracle sets $G = \Sigma$, in which case $B_G = M_G$ and GLS minimizes every linear contrast variance among unbiased linear estimators for that layout. In particular, it cannot increase the risk of the paired-mean estimator on the same layout. For estimated G, the same comparison is with respect to the planning covariance G; actual risk is evaluated separately under the true Σ.

Joint allocation and numerical certificates. We minimize additional work $\sum_S |S| n_S$ subject to $\ell_e^\top M_G^{-1} \ell_e \le d$ for every edge, initially allowing continuous $n_S \ge 0$. Write $x_S = d n_S$ and $J = d M_G$. A single semidefinite constraint

$$\begin{pmatrix} J & I \\ I & T \end{pmatrix} \succeq 0, \qquad \ell_e^\top T \ell_e \le 1 \quad (e \in F)$$

gives the continuous convex formulation. The continuous problem is solved to 10^{-8} feasibility tolerance and the resulting primal is verified by direct inverse-matrix checks. Counts below 10^{-5} are dropped, the remaining counts are rounded upward, and full-vector observations are added if needed to restore planning feasibility. The final layout is the lower-work choice between this candidate and the paired integer plan under the same G; ties favor fewer requests. Weights are recomputed after rounding.

We independently retain numerical lower bounds for the menu-restricted continuous problem. For $f_e(x) = \ell_e^\top J(x)^{-1} \ell_e$, convexity gives supporting affine lower bounds. Any nonnegative λ with $|S| + \sum_e \lambda_e \partial_S f_e(x_0) \ge 0$ certifies

$$L = \frac{1}{d} \sum_e \lambda_e \left[f_e(x_0) - 1 - \nabla f_e(x_0)^\top x_0 \right].$$

Nonnegative dual multipliers are scaled to satisfy these inequalities and provide numerical lower bounds on the menu-restricted continuous work optimum. The reported interval pairs a feasible primal value with this lower witness.

Information and sample separation. Oracle runs use the exact empirical-population covariance and either all 192 cached vectors or 128 to match the pilot comparator. Neither planner receives population means. The practical method fits Ledoit–Wolf covariance to 64 pilot rows and adds εI, where $\varepsilon = 10^{-8} \max\{\mathrm{tr}(\widehat{\Sigma}_{\mathrm{LW}})/K, 10^{-3}\}$. This rule is fixed before evaluation and uses no true moments. It ensures positive definiteness even for singular pilot covariances; it regularizes covariance, without shrinking the estimated mean. The remaining 128 cached vectors and all fresh requests are independent of the pilot. The oracle population covariances are already positive definite and require no ridge.

We use 512 independent pilots per population, each evaluated at targets $1/1024$ and $1/4096$ on both graphs. For each method, exact conditional MSE is averaged separately for every edge before taking the maximum; the edge attaining that maximum is accompanied by its edgewise Monte Carlo standard error. Under the independent split, actual MSE equals unbiased-estimator variance, so evaluation under the population covariance isolates the effect of covariance substitution.

Table 6: Primary and expanded covariance-oracle menus ($h_2 = 192$).

Population	Graph	$1/d$	Primary continuous interval	Expanded continuous interval	Expanded (W, R)
RB	clique	1024	[1253.410, 1253.414]	[1251.348, 1251.355]	(1269, 686)
RB	clique	4096	[11190.621, 11190.646]	[11190.171, 11190.186]	(11234, 3691)
RB	star	1024	[564.347, 564.349]	[563.512, 563.515]	(573, 302)
RB	star	4096	[7927.580, 7927.591]	[7926.708, 7926.728]	(7974, 2381)
LRB	clique	1024	[2390.342, 2390.348]	[2390.346, 2390.348]	(2463, 679)
LRB	clique	4096	[20964.445, 20964.473]	[20964.447, 20964.473]	(21092, 2981)
LRB	star	1024	[695.512, 695.516]	[695.513, 695.516]	(706, 353)
LRB	star	4096	[13547.601, 13547.618]	[13547.486, 13547.618]	(13622, 2850)

Table 7: Pilot weights and allocation at $d = 1/4096$.

Population	Graph	Paired MSE/d	GLS, paired layout	Joint GLS
RB	clique	1.049	0.998	1.076
RB	star	1.024	0.939	1.045
LRB	clique	0.926	0.879	1.068
LRB	star	1.019	0.836	1.125

Work and request price for the reported oracle plans. For the two clique plans in Table 5, compare the charge $W + aR$ rather than work alone. The RouterBench GLS plan costs less than the paired plan precisely for $a < 2568/2247 = 8/7$. For LLMRouterBench the analogous threshold is $a < 1796/1723$. These are pairwise comparisons of the displayed integer plans. The full joint-price design problem is treated separately by the resource-planning analysis. The resource unit is a model-score evaluation, with a separate price for an independent request.

Resource caps. Every primary-menu query has at least two coordinates. For bounded pilot scores, the Ledoit–Wolf contrast variance plus the stated ridge is at most $1 + 5 \times 10^{-9}$. At $d = 1/D$ with integer D, the full-vector paired fallback gives $W \le K(D+1-128)$ and $R \le \lfloor K(D+1-128)/2 \rfloor$. We report these analytical hard caps together with the empirical resource distributions from the pilot study.

Menu and estimator diagnostics. Table 6 compares the primary menu with its singleton/leave-one-out expansion. The reported intervals bracket each menu's continuous work optimum numerically, using the checked supporting-hyperplane lower witness and feasible primal, with outward-rounded displayed endpoints. The integer plans are feasible rounded solutions. Enlarging the menu can only weakly lower the continuous optimum; at the displayed precision several intervals overlap. Separately rounded integer outputs can have either ordering.

The same-pilot comparison also separates weight optimization from reallocation. Table 7 first keeps the paired allocator's layout fixed and changes only its estimator to GLS, then allows joint reallocation. All three procedures use the same fitted covariance and independent estimation split. Their actual errors are evaluated under the population covariance. The change in layout spends less work but can raise the error when the covariance used for planning is estimated.

O Range-information baselines and real-score audit

This study compares public-range completion rules with their exact risk on a fixed empirical population, including shrinkage bias.

Input and workload. The input is the archived singleton-inference score table for 5,400 held-out EuroSAT test images. Six fixed pipelines use frozen ResNet-18 and ViT-tiny encoders, each followed by three logistic heads with $C \in \{10^{-5}, 10^{-3}, 10^{-1}\}$. The heads were fit on 2,400 separate training images. Columns $0, 1, 2$ use ResNet-18 and $3, 4, 5$ use ViT-tiny, each in increasing C order. Scores are accuracy and the Brier reward $1 - \frac{1}{2}\|p - \mathbf{e}_y\|_2^2$, both in $[0, 1]$. The frozen score table and its model/dataset generation manifest identify the empirical population used below.

The four score-independent edge sets are $\{01, 02, 03\}$, $\{01, 23, 45\}$, $\{01, 02, 03, 04, 05\}$, and all pairs of six candidates. The six-candidate library is unchanged, including the two isolated vertices of the three-edge star. We regard the score table as an empirical distribution and independently sample cache and fresh row positions with replacement. The target is its population mean contrast. Results are marginal over cache and fresh observations from this empirical population. Mean and covariance computed from the whole table are exclusively evaluation information.

Compared estimators and hard resource ledgers. At each $b \in \{2, 8, 32, 128, 512, 2048\}$, set $h = 3b/2$ and $d = 1/(2b)$. Cached paired means use no new observations. Their common-shrinkage variant is $\widehat{\theta}_e = \lambda_h \overline{D}_e$, with $D_e = Y_i - Y_j$ and $\lambda_n = \sqrt{n}/(\sqrt{n}+1)$. It corresponds to a coherent mean vector $\widehat{\mu}_i = 1/2 + \lambda_h(\overline{Y}_i - 1/2)$. Since $\mathrm{Var}(D_e) \le 1 - \theta_e^2$, its bound is $m(h) = (\sqrt{h}+1)^{-2}$ using only the same public range.

Joint shrinkage supplements every requested endpoint at each new position until

$$N = \max\{h, N_{\min}\}, \qquad N_{\min} = \left\lceil (\sqrt{2b} - 1)^2 \right\rceil = 2b + 1 - \lfloor \sqrt{8b} \rfloor.$$

Integer square-root evaluation avoids floating-point ceiling errors. Its resources are $R = N - h$ and $W = |V(F)|(N - h)$, where $V(F)$ is the set of endpoints of requested edges. The bounded cover repeats a minimum-cardinality cover for $N - h$ positions and shrinks its balanced edge estimates by λ_N; the cover is computed exactly over the six vertices. The variance-only cover repeats it $b/2$ times. Planning uses no realized scores, covariances, or pilot information. We report hard incremental W, R; existing cache work is sunk. For every tested precision, $2b$ is a perfect square and $d = m(N_{\min})$. Whenever completion is needed, $0 < N_{\min} - h \le h$, so the bounded cover's work $(N_{\min} - h)\tau(H)$ is exactly optimal at the minimum request budget $R = N_{\min} - h$. For $b = 2, 8$, the optimal added work and request count are both zero. The other baseline costs are feasible construction costs. For targets between endpoints, Appendix J supplies the general finite-layout optimization.

Explicit weights, bias, and independent analytic evaluation. For a fixed edge (i, j) and completion, let a_i, a_j be its two new solo counts, k its cache-plus-new paired count, and $U = a_i + a_j + k$. The balanced unshrunk estimate weights each first/second solo observation by $2/U$ and $-2/U$, and each joint observation by $(U - 2a_i)/(Uk)$ and $-(U - 2a_j)/(Uk)$. The endpoint weights sum to one with their respective signs. All tested plans have nonnegative block weights and a common sum of absolute coefficients $2/U$ at every informative request. The bounded plan multiplies the estimate by λ_U. For actual covariance Σ, its unshrunk variance is

$$V_e = \frac{4a_i\Sigma_{ii} + 4a_j\Sigma_{jj}}{U^2} + \frac{(U - 2a_i)^2\Sigma_{ii} + (U - 2a_j)^2\Sigma_{jj} - 2(U - 2a_i)(U - 2a_j)\Sigma_{ij}}{U^2 k}.$$

For any chosen coefficient λ, we evaluate the full risk $\lambda^2 V_e + (1 - \lambda)^2\theta_e^2$. For common paired observations, this simplifies to

$$\mathrm{MSE}(\lambda_n \overline{D}_e) = \frac{\mathrm{Var}(D_e) + \theta_e^2}{(\sqrt{n}+1)^2} = \frac{\mathbb{E}D_e^2}{(\sqrt{n}+1)^2}.$$

All reported risks are maxima of the analytically integrated edgewise MSEs. Direct blockwise recomputation from transformed score rows agrees with the covariance formula to numerical precision; endpoint-Bernoulli calculations likewise reproduce the common-shrinkage bound.

Certificate thresholds and empirical risk. At $h = 1.5b$, cache shrinkage meets its public target exactly when $b \le 14 + 8\sqrt{3} \approx 27.856$. The larger tested precisions therefore need new observations for the uniform guarantee over the bounded class. Across all 15 pairs, accuracy has maximum contrast variance 0.281846914 and maximum second moment 0.493518519; Brier reward has corresponding maxima 0.033650833 and 0.173357373. For every $h \ge b$, cached paired risk divided by d is consequently at most $2\max_e \mathrm{Var}(D_e)$, namely 0.563694 and 0.067302. Shrunk-cache risk divided by d is strictly below $2\max_e \mathbb{E}D_e^2$, namely 0.987038 and 0.346715. Thus the cached estimates meet the empirical-population target throughout this parameter region.

The worst shrunk-cache accuracy edge is $(3, 5)$: $\theta = -0.462778$, $\mathrm{Var}(D) = 0.279355$, and squared bias contributes 43.395% of MSE. The worst Brier edge is $(2, 3)$: $\theta = 0.378039$, $\mathrm{Var}(D) = 0.030444$, and the squared-bias share is 82.439%. These shares are independent of the paired sample count. The risk evaluation includes both terms. Table 8 lists the public-range request thresholds, and

Table 8: Minimum-request endpoints from public-range certificates.

b	h	$m(h)/d$	$N_{\min}$	Joint/cover R	Variance cover R
2	3	0.5359	1	0	1
8	12	0.8029	9	0	4
32	48	1.0182	49	1	16
128	192	1.1599	225	33	64
512	768	1.2421	961	193	256
2048	3072	1.2865	3969	897	1024

Table 9: Empirical-population risk and work at $b = 128$, $h = 192$.

Workload	Method	W	Accuracy MSE/d	Brier MSE/d
Star 3	Paired cache	0	0.3202	0.0392
Star 3	Shrunk cache	0	0.3789	0.1841
Star 3	Joint shrink	132	0.3267	0.1587
Star 3	Variance cover	64	0.4427	0.0410
Star 3	Bounded cover	33	0.3912	0.1647
Matching 3	Paired cache	0	0.3730	0.0406
Matching 3	Shrunk cache	0	0.5681	0.2011
Matching 3	Joint shrink	198	0.4898	0.1734
Matching 3	Variance cover	192	0.4013	0.0414
Matching 3	Bounded cover	99	0.5525	0.1790
Star 5	Paired cache	0	0.3380	0.0392
Star 5	Shrunk cache	0	0.3974	0.1841
Star 5	Joint shrink	198	0.3426	0.1587
Star 5	Variance cover	64	0.4427	0.0410
Star 5	Bounded cover	33	0.4036	0.1647
K_6	Paired cache	0	0.3758	0.0449
K_6	Shrunk cache	0	0.5724	0.2011
K_6	Joint shrink	198	0.4935	0.1734
K_6	Variance cover	320	0.3986	0.0625
K_6	Bounded cover	165	0.5550	0.1771

Table 9 reports exact empirical-population risks and work at $b = 128$. At the minimum request budget $R = 33$, bounded cover has minimum completion work; both cache-only rules use zero additional work.

P Additional finite graph diagnostics

The finite frontiers below make the partial-observation certificate operational on small workloads. The scalable study then applies the finite-request constructions and lower certificates to larger graphs.

Certified finite resource frontiers. We fix a shared cache of $h = 3$ samples and enumerate the complete layout space for up to six candidates and $R \in \{1, 2, 3\}$ new sample positions. The planning program uses the exact bounded-score lottery risk. Rational primal and dual certificates establish the minimum integer work at each target.

Figure 5(a) uses $d = m(4) = 1/9$. A three-edge star requires one evaluation, whereas a three-edge matching requires three, despite their equal number of comparisons. Increasing the request cap from one to two reduces complete-workload cost from four to three evaluations on five candidates and from five to four on six candidates. The five-cycle still requires three. All these minima remain unchanged when the cap increases to three.

The target matters as well as the workload. At $d = 1/8$, the six-candidate complete workload needs only two additional evaluations for each tested request cap. At the coarser target $d = 1/4$, the bounded-score cache-only estimator already certifies every workload with zero additional work. Figure 5(b) uses the exact complete-graph formulas beyond the enumerated instances. For $K = 2048$

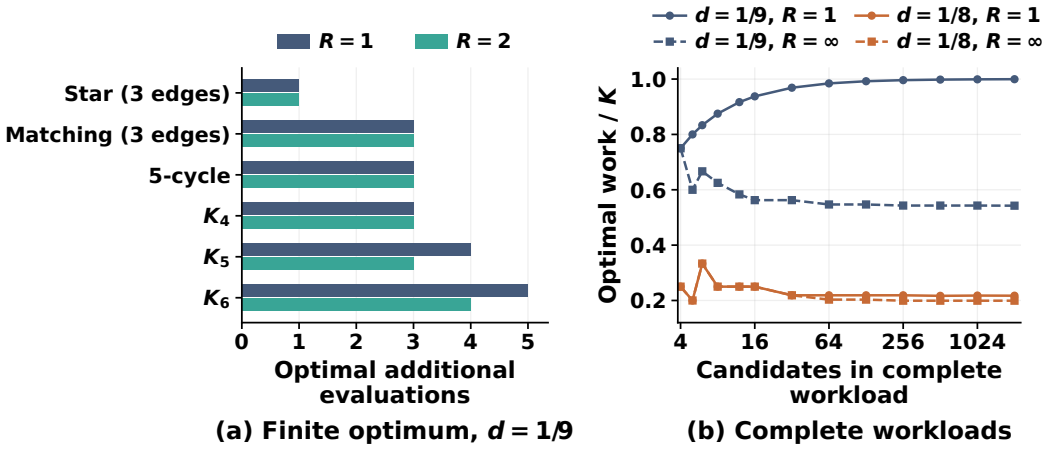

Figure 5: Finite-work optima and complete-workload limits.

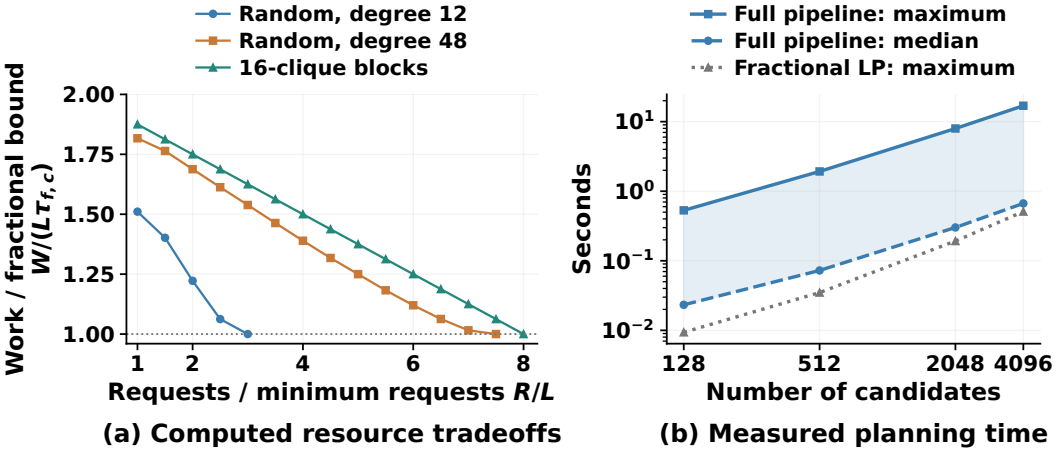

Figure 6: Finite-request graph schedules and planning time.

and $d = 1/9$, one additional position requires 2047 evaluations, while an unrestricted request allowance needs 1111. At $d = 1/8$, the corresponding costs are 445 and 408. Thus extra sample positions can change the optimal work substantially, and the attainable gain depends on the requested accuracy.

P.1 Scalable general-graph planning protocol

We evaluate feasible schedules and their lower certificates on larger graphs. Figure 6 reports resource tradeoffs and planning time; Figure 7 reports certified joint-price ratios across request prices and graph families. The feasible library is built from one fractional-cover solution and three proper colorings; every retained layout is checked against its exact integer resources and edge-union conditions.

Instances and public inputs. We use four library sizes, $K = 128, 512, 2048, 4096$, and eight graph families: a star, a matching, a random bipartite graph, an odd cycle on $K-1$ vertices with one isolated vertex, two uniform fixed-edge-count random graphs with average degrees 12 and 48, preferential-attachment graphs with three added edges per new vertex, and connected blocks of 16-cliques. The bipartite graph has equal parts; each left vertex selects 12 distinct right neighbors. The clique blocks are linked through one designated vertex in each consecutive pair of blocks. Candidate prices are either one or independently drawn from $\{4, 5, 6, 7, 8\}$. Each graph–price instance uses both $L = 8$ and $L = 9$, $h = 36 - L$, and target $1/49$. The instantiated graphs and prices are included in the supplementary artifacts.

A library with explicit schedules. We solve the weighted fractional vertex-cover LP and round the numerical primal and dual values to half-integral representatives, followed by exact checks of primal

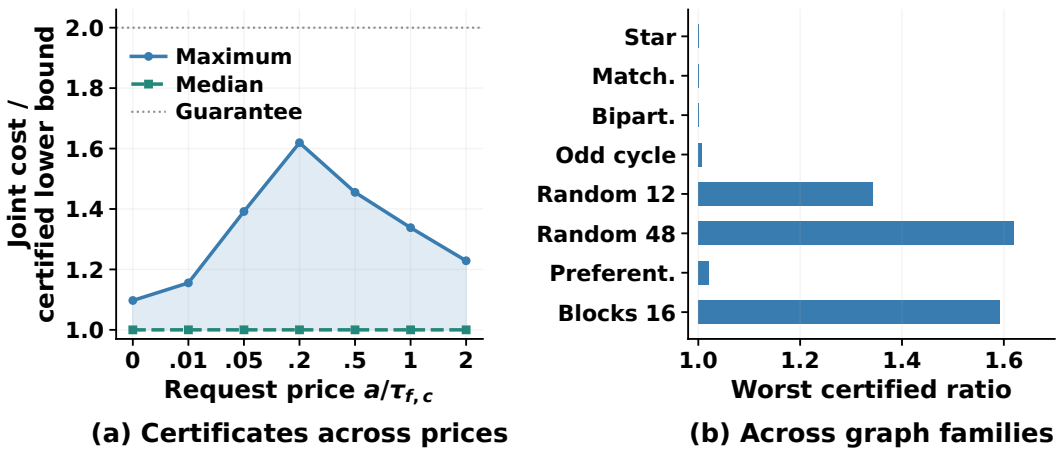

Figure 7: Certified joint-cost ratios across request prices and graph families.

feasibility, dual loads, and objective agreement. Largest-first, smallest-last, and DSATUR greedy colorings are applied to the half-valued induced graph. For each coloring, the program orders color classes by their total candidate price and keeps the most expensive q classes at half counts, upgrading the others to full counts. A block of fresh positions is allocated to each retained class; full-count vertices share the same initial positions. For odd L, the even construction for $L-1$ is followed by one cover request.

The cover candidate is the cheapest of the threshold cover, two greedy cleanups, and covers formed by omitting one entire color class. Cleanup removes a vertex only when all its neighbors remain selected. These procedures preserve feasibility and the factor-two cover bound. All feasible library layouts are checked; at equal request count only the cheapest is retained, and dominated resource pairs are removed. The resulting frontier is the computed library frontier. The lower certificates apply to the entire design class.

Instance lower certificates. Besides $L\tau_{f,c}$, we use a greedily constructed collection of vertex-disjoint cliques. For a clique of size s, let $f_s(r) = r(2s-r-1)/2$ be the number of edges touched by r selected vertices. A feasible marginal-risk design must touch clique edges at least $L\binom{s}{2}$ times in expectation. Order its vertex prices as $c_{(1)} \le \cdots \le c_{(s)}$ and write $C_r = \sum_{j=1}^{r} c_{(j)}$. For every $r = 1, \ldots, s-1$, each request contributes cost $c(S_t \cap C)$ to that clique of at least

$$\frac{c_{(r)}}{s-r} f_s(|S \cap C|) + C_{r-1} - \frac{c_{(r)}}{s-r} f_s(r-1).$$

Summing over at most R padded request positions gives a rational lower bound. The maximum over r is used for each clique; vertex-disjoint cliques can be added without double-counting work. Taking the maximum of this bound and $L\tau_{f,c}$ gives $B(R)$.

For request price a, the selected deterministic plan minimizes $aR + W$ over the feasible library. The competing class has a hard total-cost cap and may vary request count and work jointly across its lottery support. Its lower certificate is therefore $\min_{r \ge L,\ r \in \mathbb{R}} \{ar + B(r)\}$, with $r = \mathbb{E}[r(S)]$. We evaluate every rational clique breakpoint $r = L\binom{|C|}{2}/f_{|C|}(j)$ and every crossing of the summed clique bound with $L\tau_{f,c}$. The objective is affine between these knots. Beyond $\max_C L|C|/2$, the lower bound is constant and nonnegative request prices cannot improve the objective. The tested price grid is

$$a/\tau_{f,c} \in \{0, 1/100, 1/20, 1/5, 1/2, 1, 2\}.$$

Each upper and lower resource value is computed with integers or rational arithmetic.

Checks and timing. Exact feasibility checks cover the fractional primal–dual certificates, colorings, covers, retained layouts, and clique lower witnesses. Across the tested price grid, the largest certified joint-cost ratio is 1.61930, and the largest cost reduction from the repeated-cover candidate is 46.67%. At $K = 4096$, the median and maximum planning times per graph–price instance are 0.668 and 16.947 seconds; the maximum fractional-LP time is 0.510 seconds.

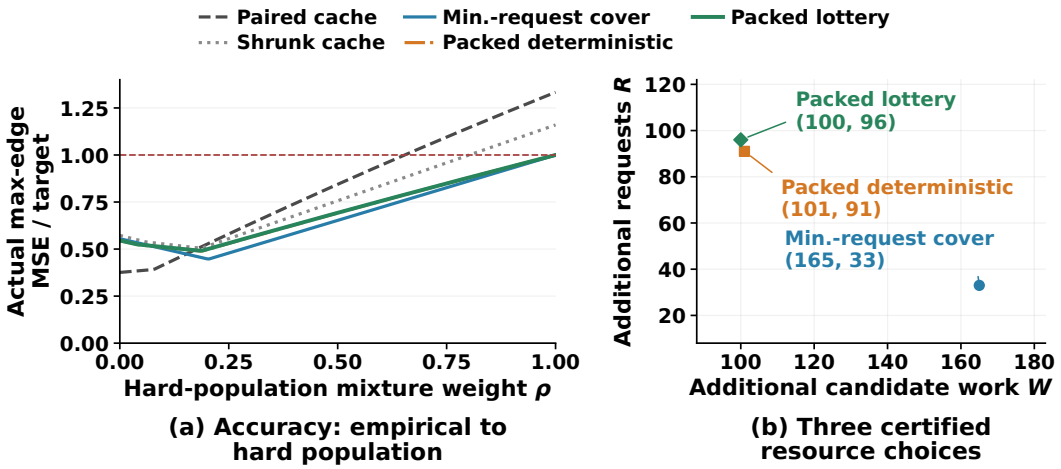

Figure 8: Controlled population sensitivity and executable resource choices.

Table 10: Resource and risk comparison on K_6.

Plan	R	W	Uniform risk	Accuracy MSE	Brier MSE	Hard-pair MSE
Paired cache	0	0	1.333333	0.375796	0.044868	1.333333
Shrunk cache	0	0	1.159878	0.572421	0.201073	1.159878
Full-vector completion	33	198	1.000000	0.493519	0.173357	1.000000
Min.-request cover	33	165	1.000000	0.554996	0.177103	1.000000
Packed deterministic	91	101	1.000000	0.545987	0.178035	1.000000
Packed lottery	96	100	0.999744	0.545446	0.178064	0.999744

Q Exact-moment resource and population tradeoffs

We hold each acquisition plan fixed while varying the score population. Table 10 compares resources and MSE divided by d; Figure 8 displays the accuracy-mixture and resource comparisons. Figure 9 gives the Brier-mixture and unknown-mean binary-family calculations. All risk calculations include bias.

Population and estimator. We use the six-model EuroSAT empirical population, score definitions, and sampling-with-replacement protocol of Appendix O, with $h = 192$ and $d = 1/256$. The acquisition plans depend only on the requested graph, the public range, h, and d; population moments enter the risk evaluation. For each realized layout we use the balanced estimator and exact covariance expression given in Appendix O. Completion plans use the common shrinkage parameter $t = 15$, so for union count U and unshrunk contrast variance V_{ij},

$$\mathrm{MSE}(S) = \left(\frac{U}{U+15}\right)^2 V_{ij} + \left(\frac{15}{U+15}\right)^2 (\mu_i - \mu_j)^2.$$

For $r_{ij}(S) = \mathbb{E}_P[(\widehat{\theta}_{ij} - \theta_{ij})^2 \mid S]$, the reported marginal risk is $\max_{ij} \mathbb{E}_{S\sim\nu} r_{ij}(S)$. The uniform marginal envelope on each edge is

$$\max\left\{\mathbb{E}_S \frac{U}{(U+15)^2},\quad \mathbb{E}_S \frac{225}{(U+15)^2}\right\}.$$

Paired cache uses $t = 0$, and shrunk cache uses $t = \sqrt{192}$.

Packed resource points. The three completion plans are the K_6 constructions of Appendix J.4. Minimum-request cover has $(R, W) = (33, 165)$ and $U = 225$ on every edge; the packed deterministic layout has $(91, 101)$ and the same deterministic union count. The packed lottery has $(96, 100)$, with each fixed edge satisfying

$$\Pr(U = 224, 225, 226) = (1, 12, 2)/15.$$

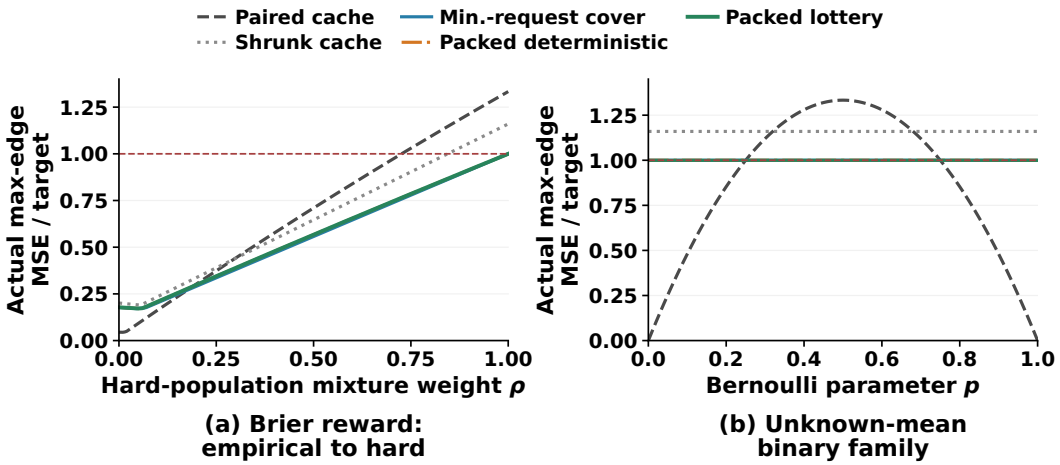

Figure 9: Fixed-plan risks for Brier mixtures and the binary family.

Exact evaluation gives

$$R_{\mathrm{var}} = \frac{12437979907}{3184939008960}, \qquad R_{\mathrm{bias}} = \frac{4144795201}{1061646336320}, \qquad \frac{\max\{R_{\mathrm{var}}, R_{\mathrm{bias}}\}}{d} = 0.9997437461861266.$$

For the $U = 224$ layouts, the public layout-conditional envelope exceeds d; for the complementary Bernoulli contrast at $p = 1/2$, the corresponding conditional risk is $1.003904d$. Averaging over the lottery yields the marginal risk shown in Table 10. The resource optimality arguments for $(91, 101)$ and $(96, 100)$ are given in Appendix J.4.

The main three-plan portfolio has additive costs $165 + 33a$, $101 + 91a$, and $100 + 96a$. Its switching request prices are $a = 1/5$ and $a = 32/29$; these thresholds refer to this three-plan portfolio.

Controlled populations. For a requested star, set the center to B and the leaves to $1 - B$ with $B \sim \mathrm{Bernoulli}(p)$; unused coordinates equal $1/2$. For the matching, each pair uses an independent copy. For the clique, only the prespecified pair $(0, 1)$ has complementary Bernoulli scores and all other coordinates equal $1/2$. The designated clique contrast has exact risk

$$4p(1-p)\,\mathbb{E}\frac{U}{(U+15)^2} + (2p-1)^2\,\mathbb{E}\frac{225}{(U+15)^2}.$$

For each real-score metric we also evaluate $P_\rho = (1-\rho)P_{\mathrm{empirical}} + \rho P_{\mathrm{hard}}$ using the fixed $p = 1/2$ hard population. Its covariance is

$$\Sigma_\rho = (1-\rho)\Sigma_0 + \rho\Sigma_1 + \rho(1-\rho)(\mu_1 - \mu_0)(\mu_1 - \mu_0)^\top.$$

All plans remain fixed as p and ρ vary, and the cache and fresh samples are drawn from the same chosen population. The empirical EuroSAT endpoint is cache-sufficient, while the complementary Bernoulli endpoint reaches the range-uniform boundary; Figures 8 and 9 trace the exact interpolation.

Exact moment calculations from the frozen score table and finite-support binary families reproduce the displayed risks and rational resource certificates to numerical precision.