POLICY LEARNING FROM REWARD TOTALS

Anonymous authors
Paper under double-blind review

ABSTRACT

A reward total records how much a group received, but not which interaction produced it. We study policy learning from N interactions released as at most M disjoint, unweighted reward totals. An encoder may use all contexts, actions, and logging propensities to form groups before seeing rewards; the learner receives the totals, metadata, and assignments. In a fixed two-policy family, data-independent grouping has worst-case regret at least of order $M^{-1/2}$, whereas two comparison-aware totals attain $N^{-1/2}$. For larger finite libraries, randomized block encoders trade the number of totals against estimation variance, and lower bounds cover every outcome-blind partition fitted to the complete metadata batch. The bounds match in specified polynomial-budget and large-sample linear-budget regimes. For policies generated by r fixed gates and A actions, a shared value-moment table requires at most $rA+1$ totals per reporting window and supports optimization over all stochastic heads after release. The reward mean need not be realizable by the gates; under i.i.d. logging and coverage, the binary-action rate matches individual feedback. Simulations and a retrospective click-log study distinguish policy-selection error, same-log acquisition loss, and later-period return: shared totals reduce acquisition loss on the observed log, while later-period intervals do not establish return dominance.

1 INTRODUCTION

Suppose a recommendation log retains the context and displayed action of every interaction, but its feedback consists only of group reward totals. A policy's value depends on which interactions received those rewards. The learning problem is therefore to decide what the totals should combine:

How should N interactions be grouped into M totals to learn a good policy?

This choice matters even for two policies. On a fixed i.i.d. family, data-independent grouping has worst-case regret at least $c/\sqrt{M}$, while comparison-aware grouping uses two totals to attain $N^{-1/2}$ (Theorem 4.1). Well-chosen groups can therefore use additional interactions without increasing the total budget.

A documented aggregation interface in AWS Clean Rooms motivates this distinction between permitted grouping operations and permitted aggregate expressions (Amazon Web Services, n.d.a). AWS identifies a disclosure risk behind restricting expressions: they can isolate small subgroups inside a larger aggregate (Amazon Web Services, n.d.c). A documented query groups partner-owned measures using the querying party's own dimension (Ramachandran & Masciola, 2023). This motivates precomputing policy-aware keys from metadata. Under our single-release protocol, the encoder assigns keys before seeing rewards, the collector sums each reward once by key, and the learner retains all metadata and assignments. Exact weighted-contrast or head-moment queries would instead provide our empirical targets directly. Section 2 formalizes the budget; Appendix H distinguishes the statistical protocol from the service's broader capabilities.

Importance-weighted evaluation pairs rewards with known policy weights (Horvitz & Thompson, 1952; Swaminathan & Joachims, 2015; Zhou et al., 2023). Our encoder puts those weights into assignment probabilities; the learner combines the resulting totals to estimate policy-value contrasts. Uniform error control then justifies selecting a policy from these estimates. The two-total example in Section 2 develops this idea before the block construction in Section 3.

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.

Contributions. Table 1 organizes the guarantees. For arbitrary finite libraries, we construct encoders with an explicit total-budget–variance tradeoff and prove lower bounds for every outcome-blind batch partition and measurable randomized selector. The bounds match at polynomial budgets and, for linear budgets, at sufficiently large sample sizes. For a fixed-gate policy class, all head values depend on the same moment table, so we encode those moments instead of separate contrasts. This requires $rA + 1$ totals per window for r gates and A actions. Both constructions use the same protocol and allow arbitrary bounded reward means. The gates constrain the policy class, not the reward function. Existing moment-based and policy-learning methods optimize once suitable statistics are available (Patrini et al., 2014; Kitagawa & Tetenov, 2018; Athey & Wager, 2021); our task is to acquire them through reward totals and control the error uniformly over the declared class.

Table 1: Results under the disjoint-total protocol; $p = K - 1$, L is the window count, and b the block size. Matching rows give worst-case minimax expected regret over libraries and all outcome-blind batch partitions under i.i.d. logging, for $p \geq 4$, one window, $W \leq 2$, and no minimum group size. $M_{\log}$ is defined in Corollary 4.4; Section 4 gives the remaining ranges and full conditions.

Object	Budget and guarantee	Result
Finite library	At most $L\lceil p/b \rceil 2^b$ totals; uniform contrast error controls policy regret.	Thm. 3.1, Cor. 3.2
Linear-budget matching	$M = 2p$, $N \geq Cp^{5/2} \log(2p)$; $\Theta(\sqrt{p \log p / N})$ minimax regret.	Thm. 4.3
Polynomial-budget matching	$M_{\log} \leq M \leq p^2$, $N \geq pM/\log p$; $\Theta(\sqrt{K/N})$ minimax regret.	Cor. 4.4
Fixed-gate heads	$rA + 1$ totals per window estimate shared moments; the binary-action rate matches individual feedback.	Thm. 5.1

Learning from label proportions provides estimators and learning guarantees for group feedback (Kück & de Freitas, 2005; Busa-Fekete et al., 2023; Li et al., 2024; Busa-Fekete et al., 2025; Applebaum et al., 2026). Curated bags and adaptive aggregation exploit how groups are formed (Chen et al., 2023; Javanmard et al., 2024; Agarwal et al., 2024). Our lower bounds include grouping fitted to the whole metadata batch; reward-driven updates such as PriorBoost use a different, multistage interface. Appendix H relates these models to stochastic quantization and communication-limited estimation.

The experiments follow the same distinction between acquiring separate comparisons and acquiring shared moments. Simulations measure reconstruction and population regret. The click-log study compares each aggregate head with the same log's individual-feedback optimum, then evaluates later-period return separately.

2 Problem Setup

2.1 The observation and the decision

Each of N rows has public context x_i, logged action a_i, correct propensity $\mu_i > 0$, and reward $y_i \in [0, 1]$ hidden from the encoder and learner. Before rewards are inspected, the encoder uses the complete metadata batch and parameter-independent coins to assign each row to exactly one group. The collector observes rewards and releases only the unweighted sums $S_g = \sum_{i:g_i=g} y_i$, at most M in total. The learner receives these sums, all metadata, and every assignment; group counts n_g are determined by the assignments. This *outcome-blind* protocol allows joint metadata-fitted grouping but excludes weighted reward queries, overlapping releases, and reward-driven regrouping. The budget measures reward totals, independently of metadata bandwidth.

For a declared library $\Pi = \{\pi_0, \ldots, \pi_{K-1}\}$, let $p = K - 1$ and define the comparison weights and their individual-feedback estimates:

$$d_{ij} = \frac{\pi_j(a_i \mid x_i) - \pi_0(a_i \mid x_i)}{\mu_i}, \qquad T_j = \frac{1}{N} \sum_i d_{ij} y_i, \quad T_0 = 0. \tag{1}$$

A positive d_{ij} means that policy j assigns the logged action more probability than the reference policy; a negative weight favors the reference. Policy evaluation needs each reward paired with this signed weight. The weights may be nonlinear in x_i, but are known before rewards are observed. A bound $B \geq \max_{i,j} |d_{ij}|$ can be computed from metadata. Zero contrasts need no coding.

Two totals explain the construction. Consider one comparison, with weights $d_i = d_{i1}$. If $d_i \in \{-B, B\}$, grouping by sign gives $T_1 = B(S_+ - S_-)/N$ exactly. For arbitrary $d_i \in [-B, B]$, independently place row i in the positive group with probability $(1 + d_i/B)/2$ and otherwise in the negative group. Then

$$\mathbb{E}_{\text{routing}}\left[\frac{B(S_+ - S_-)}{N}\right] = \frac{1}{N}\sum_i d_i y_i = T_1.$$

For each fixed reward array, the estimator is unbiased over the random assignments. Policy selection also requires an error bound that holds simultaneously for all contrasts. Section 3 provides that guarantee for the multi-comparison construction.

2.2 Population policy regret

Population statements assume independent, identically distributed logs: X is sampled, $A \sim \mu(\cdot \mid X)$, and $Y \in [0, 1]$ has arbitrary mean $m(X, A)$. Define

$$V(\pi) = \mathbb{E}_X \sum_a \pi(a \mid X) m(X, a), \qquad \text{Reg}(\widehat{\pi}) = \sup_{\pi \in \Pi} V(\pi) - V(\widehat{\pi}).$$

The library or shared representation is fixed independently of this log. Upper bounds may condition on independent pilot training; lower-bound experiments supply only the N counted outcome-bearing rows. Finite-library population guarantees require

$$0 \leq \pi_j(a \mid x)/\mu(a \mid x) \leq W < \infty \quad \text{on the target support.} \tag{2}$$

Shared-head guarantees use a known logging floor $\mu(a \mid x) \geq \kappa > 0$ for every action in that class. These conditions specify action coverage; the reward mean remains arbitrary.

The lower bounds allow every *outcome-blind batch partition* and every measurable randomized selector, including output mixtures. All metadata and assignments remain visible. The matching comparisons use one window and no positive minimum group size. Upper constructions also allow known reporting windows $w_i \in [L]$; minimum-size constraints are handled before reward release by the explicit merging rule in Appendix E.4. Conditional reconstruction fixes the complete reward array before the routing coins are drawn.

3 Policy learning with randomized groups

The two-total construction uses one known comparison weight per row. We now encode several comparisons in the same group label and control the error of selecting a policy from the decoded values.

3.1 From one comparison to a policy library

To serve several comparisons, a row selects a block, randomizes its weights into signs, and uses the block and signs as its group key. The collector sums rewards by key; the learner applies the key-specific coefficients. Formally, choose $b \in \{1, \ldots, p\}$ and split the p coordinates into $J = \lceil p/b \rceil$ blocks, padding the last with zeros. Each row independently draws U_i uniformly in $[J]$, then a sign $\xi_{ij} \in \{-1, 1\}$ for each coordinate j in that block with

$$\mathbb{P}(\xi_{ij} = 1 \mid U_i) = \frac{1 + d_{ij}/B}{2}. \tag{3}$$

The group key is (w_i, U_i, ξ_i), so there are at most $LJ2^b$ possible totals. The coefficient for contrast j in that group is

$$A_{gj} = JB\xi_j \mathbf{1}\{j \text{ belongs to block } U\}. \tag{4}$$

Algorithm 1 combines the totals using these coefficients and a variance-control center c. Stochastic quantization is applied to the known comparison weights to determine assignments (Suresh et al., 2017; Alistarh et al., 2017; Gandikota et al., 2021).

Algorithm 1 Reward-blind routing and policy selection

Require: Known d, w, bound $B > 0$, block size b, center $c \in \{0, 1/2, 1\}$.
1: Partition contrast coordinates into $J = \lceil (K-1)/b \rceil$ blocks.
2: **for** each row i **do**
3: Independently draw U_i and signs using (3); assign key (w_i, U_i, ξ_i).
4: **end for**
5: Collector releases one count n_g and unweighted reward sum S_g per nonempty key.
6: $\widetilde{T}_j \leftarrow \dfrac{c\sum_i d_{ij} + \sum_g A_{gj}(S_g - cn_g)}{N}$; set $\widetilde{T}_0 = 0$.
7: **return** $\widehat{\pi} = \pi_{\arg\max_j \widetilde{T}_j}$ and radius (5).

3.2 Finite-log and population guarantees

First fix the metadata and reward array; the only randomness is the group assignment. Let $S = \sum_g S_g$, $v_0 = S$, $v_{1/2} = N/4$, $v_1 = N - S$, and $D_c = \max(c, 1-c)$. The value S is invariant under grouping and becomes available after release. For $t = \log(2p/\delta)$, define

$$r_c(\delta) = \frac{B\sqrt{2Jv_c t}}{N} + \frac{2(J+1)BD_c t}{3N}. \tag{5}$$

Theorem 3.1 (Reconstruction for arbitrary bounded responses)**.** *For every fixed (d, y, w) with $|d_{ij}| \le B$, $y_i \in [0,1]$, and $N \ge 1$, Algorithm 1 uses at most $LJ2^b$ totals. For each fixed center $c \in \{0, 1/2, 1\}$ it is unbiased: $\mathbb{E}_U \widetilde{T}_j = T_j$. With probability at least $1-\delta$ over independent row routing,*

$$\max_{0 \le j < K} |\widetilde{T}_j - T_j| \le r_c(\delta).$$

Proof idea. For each row, $\mathbb{E}_U A_{g_i j} = d_{ij}$ and $\mathbb{E}_U A_{g_i j}^2 = JB^2$. Thus $N(\widetilde{T}_j - T_j)$ is a sum of independent, centered terms $(A_{g_i j} - d_{ij})(y_i - c)$. Their total variance is at most $JB^2 v_c$ and their magnitudes are at most $(J+1)BD_c$. Bernstein's inequality and a union bound give (5). Appendix A gives the full proof, including the probability quantifiers. □

Center $c = 1/2$ gives a response-independent radius and is used in the growing-library experiment. Simultaneous control of all three centers also permits a post-release choice (Appendix A). Population policy regret additionally includes ordinary sampling error:

Corollary 3.2 (Selection under i.i.d. logging)**.** *Under Section 2.2, use $B \le W$ and a fixed center. With probability at least $1-\delta$ over the log and routing,*

$$V(\pi^\star) - V(\widehat{\pi}) \le 2W\sqrt{\frac{\log(4K/\delta)}{2N}} + 2r_c(\delta/2). \tag{6}$$

In particular, with $c = 1/2$ this is $O(W\sqrt{J\log(K/\delta)/N} + JW\log(K/\delta)/N)$.

The first term is ordinary sampling error; the second is the additional error introduced by grouping. A given comparison uses, on average, a $1/J$ fraction of the rows. The multiplier J restores its mean but increases variance. Larger blocks let each row serve more comparisons, reducing this variance at the cost of more possible sign patterns. At $b = 1$, $2L(K-1)$ totals suffice.

Choosing the code size. Table 2 shows the tradeoff. Logarithmic blocks retain a polynomial key alphabet; large blocks approach the individual-feedback rate at an exponential key cost.

Computation. Scoring uses $N(K-1)$ queries and routing costs $O(Nb)$. Decoding costs $O(LJ2^b b)$ with sparse coefficients, or $O(LJ2^b p)$ work and storage in our dense implementation. Experiments use logarithmic or smaller blocks.

Table 2: Code size and statistical cost for $p = K - 1 \geq 4$, bounded W, and center $1/2$. The leading orders omit the additive $J \log K/N$ term. Theorem 4.3 matches the first row in its sample-rich regime; Corollary 4.4 matches the middle row.

Block choice	Total cap ($L = 1$)	Leading expected regret bound
$b = 1$	$2p$	$\sqrt{p \log K/N}$
$b = \lfloor \log_2 p \rfloor$	$M_{\log} := \lceil p/b \rceil 2^b$	$\sqrt{p/N}$
$b = p$	2^p	$\sqrt{\log K/N}$

4 Lower bounds and matching regimes

4.1 What data-independent groups lose

For independent uniform signs X, A, let

$$Z = XA, \qquad Y \mid X, A \sim \text{Bernoulli}(1/2 + \theta Z), \quad |\theta| \leq 1/4. \tag{7}$$

Policies π_+ and π_- choose $A = X$ and $A = -X$, with values $1/2 + \theta$ and $1/2 - \theta$.

Theorem 4.1 (A fixed-population separation)**.** *For $N \geq M \geq 1$, every data-independent partition into at most M nonempty groups and every selector satisfy*

$$\sup_{|\theta| \leq 1/4} \mathbb{E}_\theta \operatorname{Reg}(\widehat{\pi}) \geq (16\sqrt{M})^{-1}. \tag{8}$$

On the same model, grouping by $Z = \pm 1$ uses two totals and attains worst-case regret at most $\sqrt{2/N}$. Even individual feedback has minimax regret at least $(16\sqrt{N})^{-1}$.

Thus a fixed number of well-chosen totals can use all N interactions, whereas adding rows to data-independent groups need not improve the worst-case decision. Appendix C proves this by computing the information in one count, including its observed metadata.

4.2 A lower bound for arbitrary outcome-blind grouping

A metadata-fitted partition may favor useful comparisons over a data-independent partition. A lower bound must therefore control the best grouping of the realized batch, while retaining the learner's access to its metadata.

Theorem 4.2 (Many-policy cost under batch-adaptive grouping)**.** *Let $p = K - 1 \geq 4$ and $M \geq 2$. There is a fixed K-policy library with importance ratios at most two such that every outcome-blind partition into at most M groups and every selector over the library (including mixtures) incur, on some i.i.d. population,*

$$\mathbb{E} \operatorname{Reg}(\widehat{\pi}) \geq c \min\left\{1, \sqrt{\frac{p \log p}{pM + N \log M}}\right\}, \tag{9}$$

where $c > 0$ is absolute. The partition may jointly depend on the entire metadata batch and random coins; all metadata and assignments remain visible to the selector.

Write $\mathfrak{R}^{\text{batch}}_{N,M}$ for worst-case minimax expected regret over libraries and populations with $W \leq 2$, optimizing over these encoders and selectors.

The proof first bounds the information in a count conditional on its fixed metadata composition. It then controls the total comparison energy uniformly over partitions of the realized batch, so choosing a favorable partition is already accounted for. A multiple-hypothesis test converts the remaining ambiguity about the best policy into regret. Appendix D gives the count and partition bounds.

At $M = 2p$ and $N \geq p^2/\log p$, Theorem 4.2 gives an order-$\sqrt{p/N}$ lower bound, short of the one-coordinate code's $\sqrt{p \log p/N}$ rate. Closing this gap at large N requires rewards with additional variation that leaves policy-value contrasts unchanged but obscures them in the totals.

Theorem 4.3 (Linear total budget with unrestricted rewards)**.** *There are absolute constants $c, C > 0$ such that, for $p = K - 1 \geq 4$, $M = 2p$, and $N \geq Cp^{5/2}\log(2p)$, the worst-case minimax expected regret over K-policy libraries and all outcome-blind batch partitions is*

$$c\sqrt{\frac{p\log p}{N}} \;\leq\; \mathfrak{R}^{\mathrm{batch}}_{N,2p} \;\leq\; C\sqrt{\frac{p\log p}{N}}.$$

This uses one window, importance ratios at most two, and no minimum-group-size constraint. The lower bound holds for a fixed finite metadata law and library, against every partition fitted to the complete metadata batch and every randomized selector. The upper bound is the $b = 1$ code of Section 3 and holds for arbitrary bounded responses.

Proof overview. We fix the metadata law, library, and reward prior before sampling or encoder randomization. Reward means contain a component determining the policy contrasts and a *nuisance* component orthogonal to them. The nuisance leaves those contrasts unchanged, but each total combines it with the value-relevant component.

We first construct a metadata law whose geometry limits how many contrast directions a nonlinear grouping channel can estimate precisely despite nuisance variation. The spectral restriction holds uniformly over channels and admissible perturbations of the law (Appendices J.1–J.2). Conditional on the complete metadata and encoder randomness, every fitted partition induces an empirical channel through its within-type assignment frequencies, so the same restriction applies (Appendix J.5).

A joint latent shift then bounds information about the best-policy index in mean-matched binomial counts; multi-hypothesis testing converts this bound into regret (Appendix J.3). Gaussian variables generate the hard reward means; the observed rewards remain Bernoulli. To return to the original experiment, we compare heterogeneous group counts with the mean-matched counts (Appendix J.4). The counts are independent conditional on the reward mean, metadata, and partition; joint integration of the shared prior preserves the resulting dependence. Finally, sufficiently large N controls prior truncation, empirical metadata frequencies, and count-comparison error. These steps give the stated sufficient sample range, rather than a necessary transition threshold.

4.3 Where the upper and lower bounds meet

Corollary 4.4 (A matched polynomial-budget regime)**.** *Let $p = K - 1 \geq 4$, $b = \lfloor \log_2 p \rfloor$, $M_{\log} = \lceil p/b \rceil 2^b$, and use one window with $W \leq 2$. For every cap $M_{\log} \leq M \leq p^2$ and $N \geq pM/\log p$, the worst-case minimax expected regret over K-policy libraries and all outcome-blind batch partitions is $\Theta(\sqrt{K/N})$. Algorithm 1 with this b and center $1/2$ attains the upper bound. In particular, $M = M_{\log} = \Theta(p^2/\log p)$ suffices.*

Thus a polynomial budget attains $\Theta(\sqrt{K/N})$, while the linear-budget code is optimal in Theorem 4.3's sufficient sample range. Both comparisons use one window and no minimum-size constraint. Singleton releases settle $N \leq p^2/\log(2p)$ (Appendix J.6); the intermediate range retains the general $\Omega(\sqrt{p/N})$ and $O(\sqrt{p\log p/N})$ bounds. Proposition D.3 concerns the fixed code only.

These are worst-case finite-library costs. A structured class can instead share a small set of value moments, which the encoder acquires once for all its policies.

5 Learning shared policy heads

5.1 From context groups to shared policy heads

Consider first r fixed context groups. Choosing an action distribution in each group creates a large policy class, yet its values depend only on the group–action reward moments. Soft groups allow the same idea with overlapping contexts: fix measurable weights $\phi : \mathcal{X} \to \Delta_r$, called gates, independently of the evaluation log (Jordan & Jacobs, 1994). After release, learn the action distributions U:

$$\pi_U(a \mid x) = \sum_{l=1}^{r} \phi_l(x) U_{la}, \qquad U \in (\Delta_A)^r. \tag{10}$$

The fixed gate map defines A^r deterministic heads. Every stochastic head has a value determined by the same rA moments $Q_{la} = \mathbb{E}[\phi_l(X)m(X,a)]$. For example, two context regions and two actions give four deterministic heads, all evaluated by four region–action moments. Acquiring this table supports head optimization after release. The gate map restricts the policy class but imposes no realizability assumption on $m(x,a)$.

Assume a known floor $\mu(a \mid x) \geq \kappa > 0$ for these actions. Accept each row with probability κ/μ_i, then draw $l \sim \phi(x_i)$ and assign (w_i, l, a_i). Rows not assigned to a signal key enter the same window's residual key, so every reward contributes once. Propensity thinning (Li et al., 2011) followed by categorical encoding uses at most $rA+1$ totals per window. Pooling signal totals gives

$$\widehat{Q}_{la} = \frac{\sum_w S_{w,l,a}}{N\kappa}, \qquad \widehat{V}(U) = \sum_{l,a} U_{la}\widehat{Q}_{la}. \tag{11}$$

Each row's expected coefficient is $\pi_U(a_i \mid x_i)/\mu_i$, as in the library construction ((13)). The table supports rowwise maximization in $O(rA)$ time, linear constraints by an LP, and entropy penalties by softmax; Algorithm 2 gives the procedure. The following uniform guarantee justifies choosing U from the released table.

Theorem 5.1 (A reusable release with a matched binary-action rate)**.** *Before minimum-size repairs, for every fixed metadata and response array, $\widehat{Q}_{la}$ is unbiased for the individual-feedback moment $N^{-1}\sum_i \phi_l(x_i)\mathbf{1}\{a_i = a\}y_i/\mu_i$. Under the i.i.d. population model, let $\widehat{U}$ maximize* (11)*. Then*

$$\mathbb{E}\,\mathrm{Reg}(\pi_{\widehat{U}}) \leq \min\left\{1,\ 2\sqrt{\frac{2r\log(2A)}{N\kappa}} + \frac{4r\log(2A)}{3N\kappa}\right\}. \tag{12}$$

For $A = 2$ and $\kappa \leq 1/2$, some gate maps and reward populations force regret at least $c\min\{1, \sqrt{r/(N\kappa)}\}$ even with individual rewards. Thus $2r+1$ totals in one window suffice for the optimal worst-case dependence on r, N, κ.

Uniform control makes post-release learning valid. Write $p_l = \mathbb{E}\phi_l(X)$ and $Q_{la} = \mathbb{E}[\phi_l(X)m(X,a)]$. A row contributes at most $1/\kappa$ to $\widehat{Q}_{la}$, with second moment at most p_l/κ. For *all* heads,

$$|\widehat{V}(U) - V(U)| \leq \sum_l \max_a |\widehat{Q}_{la} - Q_{la}|.$$

Bernstein's exponential-moment bound within each gate and $\sum_l \sqrt{p_l} \leq \sqrt{r}$ give (12). A binary hypercube of context-specific preferred actions gives the lower bound. Appendix E proves both results and a uniform high-probability version. □

The uniform event also applies conditional on gates learned from independent pilot data.

Reporting constraints and reuse. Windows and minimum group sizes are enforced before reward release by merging undersized groups. Appendix E.4 gives the resulting error charge; it is included in the experimental evaluation. A fixed gate map supports all heads without requiring a minimum gate mass. Selecting unrelated gate maps is a different acquisition problem (Corollary E.1). An entropy penalty of weight λ adds at most $\lambda \log A$ to unregularized optimization error.

6 Experiments

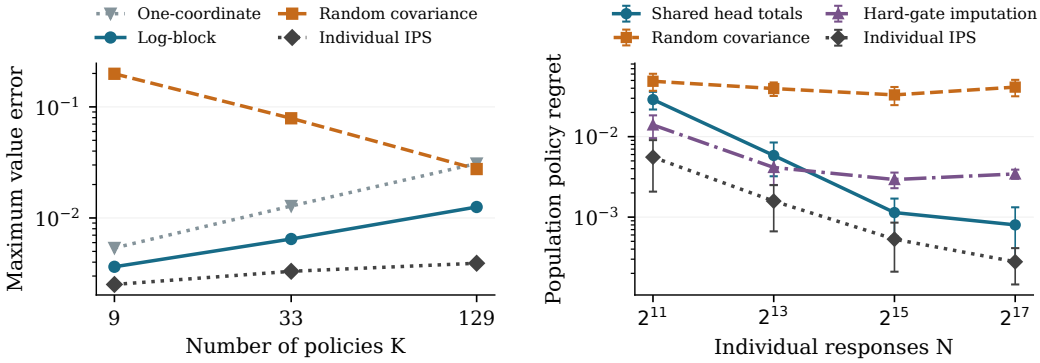
We test how grouping affects policy choice using population values in simulation, then separate same-log decision loss from later return on real clicks. Appendix G specifies every configuration; Appendix F adds two-policy, fixed-library, and release-stress comparisons.

6.1 More comparisons versus shared comparisons

Increasing the number of candidate policies. We use the K-policy family from the lower-bound proof, with $K = 9, 33, 129$, $N = 4096, 16384, 65536, 262144$, and 20 independent logs per setting.

The signal $\theta = \frac{1}{2}\sqrt{(K-1)/N}$ decreases with N to test local decision difficulty, rather than a learning curve at one fixed signal. The block sizes are one and $\lfloor \log_2(K-1) \rfloor$. Each random-bag covariance comparator receives the matching total budget. No method sees the realized rewards before forming groups.

Figure 1(a) reports maximum value error at $N = 262144$. For $K = 129$, logarithmic blocks have error 0.012555 using 2,432 totals, versus 0.027574 for random grouping at the same budget. One-coordinate coding uses 256 totals and has error 0.030901. The first comparison isolates grouping quality; the second trades budget for variance. Table 8 reports all settings. This linear-signal family differs from the nuisance family in Theorem 4.3. Appendix I tests batch-fitted trees and shows how their greater total comparison energy is distributed unevenly across directions.



(a) Library size at fixed $N = 2^{18}$.

(b) Learning an eight-gate stochastic head.

Figure 1: Maximum population-contrast error in (a) and population policy regret in (b). Random grouping in (a) matches the logarithmic-block budget; all aggregate methods in (b) have a 132-total cap. Individual IPS uses finer feedback. Points average 20 independent logs; bars in (b) are 1.96 Monte Carlo standard errors.

Fitting the head after release. Contexts have 64 binary feature vectors, four actions, and nonlinear, heterogeneous reward means. Public soft gates use independent random weights. We compare shared-head totals, an unbiased random-bag covariance estimator, hard-gate mean imputation, and individual-feedback IPS. Every method then optimizes the same head class; no candidate list or reward model is supplied. We vary gate counts $r = 4, 8, 16$, sample size, rare rewards, and weak overlap. Four windows and minimum group size four are enforced for all aggregate methods.

Figure 1(b) reports population regret for eight gates. At $N = 131072$, shared moments give 0.000801, compared with 0.041302 for random covariance and 0.003442 for hard-gate imputation. The release has at most 132 totals yet supports 4^8 deterministic heads and all their gatewise mixtures. At $N = 2048$, hard-gate imputation instead has lower population regret: 0.013995 versus 0.028924 for shared routing. Reconstruction and decision regret can therefore rank methods differently. Tables 9 and 10 retain every profile and repair cost.

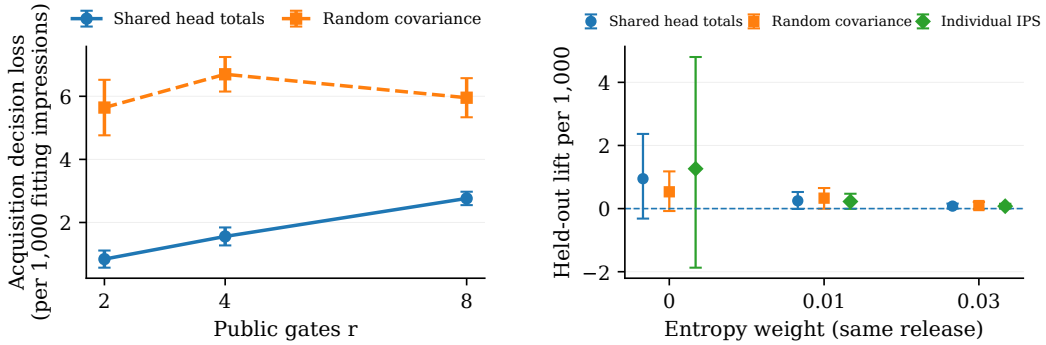
6.2 Real clicks: decision loss and later return

We retain all 150,936 position-1 impressions and 812 clicks from the Open Bandit Dataset (Saito et al., 2021). November 24–27, 2019 supplies 86,722 fitting responses with 459 clicks; November 28–30 supplies 64,214 evaluation responses with 353 clicks. The public gate map uses hashed context features, with two, four, or eight gates. Four day windows and minimum group size eight apply. The same totals fit an unconstrained head and entropy-regularized heads with predeclared weights 0.01 and 0.03; these alternatives do not trigger further releases or use held-out clicks for tuning.

We first ask how much the release changes the decision relative to individual feedback on the same log. For fixed entropy weight λ, let F_λ be the individual-feedback objective and U^*_λ its maximizer.

The *same-log acquisition loss* is $D_\lambda = F_\lambda(U^*_\lambda) - F_\lambda(\widehat{U}_\lambda)$. Both heads use the same responses, class, and λ, isolating acquisition loss from sampling and objective choice. For $\lambda > 0$, this measures the entropy-regularized objective. Later impressions separately assess transfer beyond the fitting log.

Figure 2(a) gives D_0. With eight gates, its mean is 2.763 for shared totals and 5.955 for random covariance, per 1,000 fitting impressions. The difference is 3.192 with a paired Monte Carlo bootstrap interval $[2.551, 3.809]$ across 40 alternative releases. All nine mean differences favor shared acquisition (Table 12); individual releases can reverse that ordering. The interval measures encoding variability conditional on this log.



(a) Decision loss caused by acquisition, $\lambda = 0$.

(b) Eight gates, three post-release objectives.

Figure 2: Same-log acquisition loss in (a) and later-period lift in (b), with cap $4(34r + 1)$ totals. Bars in (a) are 1.96 Monte Carlo standard errors across 40 releases, relative to the same individual-feedback optimum. Intervals in (b) resample held-out hours conditional on fitted heads; they do not cover calendar shift.

Figure 2(b) then evaluates later return. The unregularized shared head has lift $+0.947$ per 1,000 impressions with hour-block interval $[-0.317, 2.366]$; the individual-feedback reference also has an interval spanning zero. Better acquisition therefore does not establish return dominance. The full-log objective is itself estimated, and later impressions differ from fitting data. Table 11 retains all return results and class-wide reconstruction errors. Heads, regularization weights, and grouping procedures are specified or fitted without held-out clicks. All alternative releases use the same fitting log.

7 Conclusion

Grouping can retain the reward–weight association needed for policy choice. Randomized encoders estimate library contrasts, while shared moments support every head in a fixed representation. Lower bounds cover arbitrary outcome-blind batch partitions and match the encoders at polynomial budgets and, for linear budgets, sufficiently large sample sizes. The intermediate linear-budget transition remains open. Experiments separate the decision loss of acquiring these quantities from later-period return, which also reflects sampling and temporal change.

AI use statement

Generative AI tools were used extensively in checking proofs, implementing and debugging code, polishing the manuscript. The submitting authors are responsible for the final text, mathematical claims, code, and interpretation.

Reproducibility statement

All models, algorithms, constants, proofs, and experiment definitions are contained in the paper. Appendices A–E give the routing, constraint, and information-theoretic arguments; Appendix J proves

the large-sample linear-budget converse. Appendices F, G, and I specify every experiment and numerical summary. Encoders accept metadata only; collectors form unweighted totals, and decoders accept those totals. Fixed seeds determine every reported run. Unfavorable configurations and repair costs are retained. The real-log study evaluates post-release policy fitting on a single original log and does not treat alternative partitions as independent deployments.

Ethics statement

The work studies a restricted feedback interface, not a privacy definition. Grouping does not itself provide differential privacy, anonymity, or protection against auxiliary information. Practical use requires a collector authorized to access outcomes, a lawful metadata policy, and an audit of the proposed disclosure mechanism. The real-log study is a retrospective simulation of this interface, not a deployed intervention. It makes no claim about user-level causal identification or privacy guarantees.

References

Sushant Agarwal, Yukti Makhija, Rishi Saket, and Aravindan Raghuveer. Aggregating data for optimal and private learning. *arXiv preprint arXiv:2411.19045*, 2024. URL https://arxiv.org/abs/2411.19045.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/6c340f25839e6acdc73414517203f5f0-Abstract.html.

Amazon Web Services. Aggregation analysis rule. AWS Clean Rooms User Guide, n.d.a. URL https://docs.aws.amazon.com/clean-rooms/latest/userguide/analysis-rules-aggregation.html. Accessed September 23, 2026.

Amazon Web Services. Considerations and limitations. AWS Clean Rooms User Guide, n.d.b. URL https://docs.aws.amazon.com/clean-rooms/latest/userguide/custom-considerations.html. Accessed September 23, 2026.

Amazon Web Services. Minimum aggregation thresholds. AWS Clean Rooms User Guide, n.d.c. URL https://docs.aws.amazon.com/clean-rooms/latest/userguide/custom-min-agg-thresholds.html. Section: Allowing nested expressions in aggregate functions. Accessed September 23, 2026.

Amazon Web Services. Analysis rules in AWS Clean Rooms. AWS Clean Rooms User Guide, n.d.d. URL https://docs.aws.amazon.com/clean-rooms/latest/userguide/analysis-rules.html. Accessed September 23, 2026.

Lorne Applebaum, Travis Dick, Claudio Gentile, Haim Kaplan, and Tomer Koren. Optimal learning from label proportions with general loss functions. *arXiv preprint arXiv:2509.15145v2*, 2026. URL https://arxiv.org/abs/2509.15145v2.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021. doi: 10.3982/ECTA15732.

Leighton Pate Barnes, Yanjun Han, and Ayfer Özgür. Lower bounds for learning distributions under communication constraints via fisher information. *Journal of Machine Learning Research*, 21 (236):1–30, 2020. URL https://www.jmlr.org/papers/v21/19-737.html.

Andrew Bennett and Nathan Kallus. Efficient policy learning from surrogate-loss classification reductions. In *Proceedings of the International Conference on Machine Learning*, volume 119 of *PMLR*, pp. 788–798, 2020. URL https://proceedings.mlr.press/v119/bennett20a.html.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Anand Brahmbhatt, Mohith Pokala, Rishi Saket, and Aravindan Raghuveer. LLP-Bench: A large scale tabular benchmark for learning from label proportions. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2024. URL https://research.google/pubs/llp-bench-a-large-scale-tabular-benchmark-for-learning-from-label-proportions/.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy L. Nguyen, and David P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the ACM Symposium on Theory of Computing*, 2016. URL https://arxiv.org/abs/1506.07216.

Róbert Busa-Fekete, Heejin Choi, Travis Dick, Claudio Gentile, and Andres Munoz Medina. Easy learning from label proportions. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3085fd61063840fdb2e6eafac58589f8-Abstract-Conference.html.

Robert Istvan Busa-Fekete, Travis Dick, Claudio Gentile, Haim Kaplan, Tomer Koren, and Uri Stemmer. Nearly optimal sample complexity for learning with label proportions. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 5943–5976, 2025. URL https://proceedings.mlr.press/v267/busa-fekete25a.html.

Lin Chen, Gang Fu, Amin Karbasi, and Vahab Mirrokni. Learning from aggregated data: Curated bags versus random bags. *arXiv preprint arXiv:2305.09557*, 2023. URL https://arxiv.org/abs/2305.09557.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. doi: 10.1002/047174882X.

Persi Diaconis and Robert Griffiths. Exchangeable pairs of Bernoulli random variables, Krawtchouck polynomials, and Ehrenfest urns. *Australian and New Zealand Journal of Statistics*, 54(1):81–101, 2012. doi: 10.1111/j.1467-842X.2012.00654.x.

Persi Diaconis and Robert C. Griffiths. Reproducing kernel orthogonal polynomials on the multinomial distribution. *Journal of Approximation Theory*, 242:1–30, 2019. doi: 10.1016/j.jat.2019.01.007.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the International Conference on Machine Learning*, pp. 1097–1104, 2011. URL https://arxiv.org/abs/1103.4601.

Werner Ehm. Binomial approximation to the Poisson binomial distribution. *Statistics & Probability Letters*, 11(1):7–16, 1991. doi: 10.1016/0167-7152(91)90170-V.

Qing Feng, Ben Letham, Hongzi Mao, and Eytan Bakshy. High-dimensional contextual policy search with unknown context rewards using Bayesian optimization. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://neurips.cc/virtual/2020/public/poster_faff959d885ec0ecf70741a846c34d1d.html.

Venkata Gandikota, Daniel Kane, Raj Kumar Maity, and Arya Mazumdar. vqSGD: Vector quantized stochastic gradient descent. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 130 of *PMLR*, pp. 2197–2205, 2021. URL https://proceedings.mlr.press/v130/gandikota21a.html.

Paul W. Goldberg and Mark R. Jerrum. Bounding the Vapnik–Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2–3):131–148, 1995. doi: 10.1007/BF00993408.

Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. In *Proceedings of the Conference on Learning Theory*, volume 75 of *PMLR*, pp. 3163–3188, 2018. URL https://proceedings.mlr.press/v75/han18a.html.

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

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Adel Javanmard, Matthew Fahrbach, and Vahab Mirrokni. PriorBoost: An adaptive algorithm for learning from aggregate responses. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 21410–21429, 2024. URL https://proceedings.mlr.press/v235/javanmard24a.html.

Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994. doi: 10.1162/neco.1994.6.2.181.

Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/6616758da438b02b8d360ad83a5b3d77-Abstract.html.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018. doi: 10.3982/ECTA13288.

Hendrik Kück and Nando de Freitas. Learning about individuals from group statistics. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 332–339, 2005. URL https://ora.ox.ac.uk/objects/uuid:2a9663db-d06c-491f-bbab-08d53815df75.

Gene Li, Lin Chen, Adel Javanmard, and Vahab Mirrokni. Optimistic rates for learning from label proportions. In *Proceedings of the Conference on Learning Theory*, volume 247 of *PMLR*, pp. 3437–3474, 2024. URL https://proceedings.mlr.press/v247/li24b.html.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2011. URL https://arxiv.org/abs/1003.5956.

Tianhao Ma, Wei Wang, Ximing Li, Gang Niu, and Masashi Sugiyama. Learning from label proportions via proportional value classification. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6fc46679a7ba2ec82183cf01b80e5133-Abstract-Conference.html.

Giorgio Patrini, Richard Nock, Paul Rivera, and Tiberio Caetano. (almost) no label no cry. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/d3313de3f431fd64513431c4326d237c-Abstract.html.

David Pollard. A note on insufficiency and the preservation of Fisher information. arXiv preprint arXiv:1107.3797v2, 2012. URL https://arxiv.org/abs/1107.3797v2.

Novi Quadrianto, Alex J. Smola, Tiberio S. Caetano, and Quoc V. Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10(82):2349–2374, 2009. URL https://jmlr.org/papers/v10/quadrianto09a.html.

Sundeep Ramachandran and Marco Masciola. Managing data confidentiality for Scope 3 emissions using AWS Clean Rooms. AWS Architecture Blog, May 2023. URL https://aws.amazon.com/blogs/architecture/managing-data-confidentiality-for-scope-3-emissions-using-aws-clean-rooms/. Published May 31, 2023.

Bero Roos. Binomial approximation to the Poisson binomial distribution: The Krawtchouk expansion. *Theory of Probability and Its Applications*, 45(2):258–272, 2001. doi: 10.1137/S0040585X9797821X.

Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/33e75ff09dd601bbe69f351039152189-Abstract-round2.html.

Clayton Scott and Jianxin Zhang. Learning from label proportions: A mutual contamination framework. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/fcde14913c766cf307c75059e0e89af5-Abstract.html.

Nir Shlezinger, Yonina C. Eldar, and Miguel R. D. Rodrigues. Hardware-limited task-based quantization. *IEEE Transactions on Signal Processing*, 67(20):5223–5238, 2019. doi: 10.1109/TSP.2019.2935864. URL https://discovery.ucl.ac.uk/id/eprint/10085350/.

Ananda Theertha Suresh, Felix X. Yu, Sanjiv Kumar, and H. Brendan McMahan. Distributed mean estimation with limited communication. In *Proceedings of the International Conference on Machine Learning*, volume 70 of *PMLR*, pp. 3329–3337, 2017. URL https://proceedings.mlr.press/v70/suresh17a.html.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the International Conference on Machine Learning*, volume 37 of *PMLR*, pp. 814–823, 2015. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009. doi: 10.1007/b13794.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the International Conference on Machine Learning*, volume 70 of *PMLR*, pp. 3589–3597, 2017. URL https://proceedings.mlr.press/v70/wang17a.html.

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. TernGrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/89fcd07f20b6785b92134bd6c1d0fa42-Abstract.html.

Felix X. Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. On learning from label proportions. *arXiv preprint arXiv:1402.5902*, 2014. URL https://arxiv.org/abs/1402.5902.

Yuchen Zhang, John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2328–2336, 2013. URL https://papers.neurips.cc/paper_files/paper/2013/hash/d6ef5f7fa914c19931a55bb262ec879c-Abstract.html.

Pan Zhao, Antoine Chambaz, Julie Josse, and Shu Yang. Positivity-free policy learning with observational data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 238 of *PMLR*, pp. 1918–1926, 2024. URL https://proceedings.mlr.press/v238/zhao24a.html.

Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023. doi: 10.1287/opre.2022.2271.

APPENDIX CONTENTS

A Complete reconstruction and policy-selection proofs **15**

A.1 The routing identity and its conditional probability space 15

A.2 Proof of Theorem 3.1 16

A.3 Data-dependent choice of reward center 16

A.4 Proof of Corollary 3.2 17

A.5 From value contrasts to more general finite comparisons 17

B Release constraints and local score ranges **17**

B.1 Release constraints and observed heterogeneity 17

B.2 Proof of Theorem B.1 18

B.3 A computable local-range refinement 19

B.4 A random-partition covariance comparator 20

C Exact count divergence and minimax lower bounds **20**

C.1 Likelihood of a single count 20

C.2 Proof of the data-independent partition lower bound 21

C.3 Two policy-aware totals and the full-information lower bound 21

D Policy-count lower bounds for outcome-blind grouping **22**

D.1 Two-sided count information for an arbitrary group composition 22

D.2 A uniform bound over all metadata-fitted partitions 23

D.3 The local information matrix and what its trace omits 24

D.4 Proof of Theorem 4.2 24

D.5 Proof of Corollary 4.4 25

D.6 Near-linear budgets: a code bottleneck and the remaining gap 26

D.7 Sharper identities for predeclared kernels 26

E A reusable release for an infinite policy class **27**

E.1 Conditional correctness and optimization 28

E.2 Proof of Theorem 5.1 28

E.3 A matching binary-action lower bound 29

E.4 Minimum-size repairs and representation scope 29

E.5 Why unrestricted gate selection has an acquisition cost 29

F Experimental protocols and complete summaries **30**

F.1 Information access, metrics, and verification 30

F.2 Earlier finite-library comparisons 30

F.3 Two-policy fixed-budget experiment 31

F.4 Nonlinear finite-context population and frozen library 32

F.5 Reporting-window and minimum-size stress test 35
F.6 Real contextual recommendation-log replay 35

G Library growth and shared-comparison evaluations **36**
G.1 Growing-library testing family 37
G.2 Nonlinear public gates and policy fitting 38
G.3 Random-bag covariance estimator 41
G.4 Post-release heads on the Open Bandit Dataset 41
G.5 Decision loss attributable to the release 43

H Related work and the information interfaces **44**
H.1 Learning with group-level labels 45
H.2 Policy learning and release reuse 45
H.3 Randomized encoding and communication lower bounds 46

I What batch-adaptive grouping preserves **47**
I.1 Data, grouping algorithms, and a common decoder 47
I.2 Information concentration and decision error 47

J A direct-count converse at linear total budget **47**
J.1 A smaller fixed metadata distribution 48
J.2 A spectral constraint valid for every nonlinear channel 50
J.3 Gaussian prior shifts through a discrete observation kernel 51
J.4 Homogenizing each count while retaining the shared prior 52
J.5 Completion in the original batch experiment 53
J.6 The small-sample endpoint and the remaining transition 55

A Complete reconstruction and policy-selection proofs

A common decoding identity. The two-policy example extends by grouping according to several comparison weights at once. Both the library code below and the shared-head construction in Section 5 use the same identity. For a policy score $t_\pi(z)$, choose metadata-only probabilities $q_g(z)$ and decoding coefficients $a_g(\pi)$ satisfying

$$t_\pi(z) = \sum_g q_g(z) a_g(\pi). \tag{13}$$

Then $N^{-1}\sum_g a_g(\pi) S_g$ is unbiased for $N^{-1}\sum_i t_\pi(z_i) y_i$, simultaneously as an identity for all policies. A uniform error bound makes optimization after release valid. Below t_π is a library contrast; Section 5 uses the same identity for shared policy values. The lower bounds in Section 4 also allow decoders beyond this construction.

A.1 The routing identity and its conditional probability space

The metadata matrix $d \in \mathbb{R}^{N\times p}$, window labels, and reward vector $y \in [0,1]^N$ are fixed throughout this subsection. All expectations are over independent per-row routing coins. Within a row the sign coordinates may be independent, as in Algorithm 1; the coordinatewise guarantees only require their

specified marginals. The outcome collector acts after the assignment and returns each outcome in exactly one total.

Let $q(j)$ denote the block containing coordinate j. The random coefficient assigned to row i is

$$a_{ij} = JB\xi_{ij}\mathbf{1}\{U_i = q(j)\}.$$

Conditionally on $U_i = q(j)$, the mean sign is d_{ij}/B. Consequently

$$\mathbb{E}_U a_{ij} = d_{ij}, \qquad \mathbb{E}_U a_{ij}^2 = JB^2, \qquad \mathrm{Var}_U(a_{ij}) = JB^2 - d_{ij}^2. \tag{14}$$

These identities also hold for a padded last block; padded coordinates are not evaluated. Every nonzero decoder coefficient is attached to the public group key, so summation over rows and summation over released groups give the same estimator.

For a fixed $c \in \{0, 1/2, 1\}$,

$$N\widetilde{T}_j = c\sum_i d_{ij} + \sum_g A_{gj}(S_g - cn_g) \tag{15}$$

$$= c\sum_i d_{ij} + \sum_i a_{ij}(y_i - c). \tag{16}$$

Taking expectations and using (14) proves unbiasedness conditional on the fixed reward vector. The argument applies to every bounded vector, including dependent and heterogeneous responses.

A.2 Proof of Theorem 3.1

Let $Z_{ij} = (a_{ij} - d_{ij})(y_i - c)$. The Z_{ij} are independent over i, centered, and satisfy

$$|Z_{ij}| \le (J+1)BD_c =: L_c, \qquad \sum_i \mathbb{E}Z_{ij}^2 = \sum_i (JB^2 - d_{ij}^2)(y_i - c)^2 \le JB^2 \sum_i (y_i - c)^2.$$

For $y \in [0,1]$, $y^2 \le y$, $(1-y)^2 \le 1-y$, and $(y-1/2)^2 \le 1/4$. Hence $\sum_i (y_i - c)^2 \le v_c$, where the three v_c are those in the main text. Conditional on the fixed rewards, $S = \sum_i y_i$ is deterministic even though its group decomposition is random. Its appearance in the bound is therefore not a plug-in variance assumption.

The two-sided Bernstein inequality (Boucheron et al., 2013) implies

$$\mathbb{P}_U\left(\left|\sum_i Z_{ij}\right| > \sqrt{2JB^2 v_c t} + \frac{2L_c t}{3}\right) \le 2e^{-t}.$$

One way to check the safe inversion is to substitute $x = \sqrt{2vt} + 2Lt/3$ into $x^2 \ge 2t(v + Lx/3)$. With $t = \log(2p/\delta)$, a union bound over the p nonbaseline coordinates proves (5). The baseline has exactly zero contrast and error. Each of L windows has at most $J2^b$ keys, giving the claimed release budget. Empty keys require no outcome release.

For a random log, condition on its full metadata and reward vector and integrate this inequality. This establishes the same failure bound whenever the routing coins have their specified conditional law and are independent of the rewards. It does not allow an adversary to choose rewards after observing those coins. This distinction is important: unbiasedness holds for every fixed reward vector, not uniformly on a single event for all post-routing reward assignments.

A.3 Data-dependent choice of reward center

Let $\mathcal{C} = \{0, 1/2, 1\}$. Apply Theorem 3.1 separately at failure probability $\delta/3$ and take a union bound. All three simultaneous inequalities hold with probability at least $1 - \delta$. Thus a center chosen after release, for example by minimizing $r_c(\delta/3) + \max_j \beta_j(c)$, inherits its chosen bound. This requires no extra total because S and all group counts are already available. The estimator selected this way need not remain unbiased; the simultaneous high-probability bound, rather than an unqualified unbiasedness claim, justifies the selection. The earlier fixed-library experiments use $c = 0$; the two-policy and growing-library experiments use $c = 1/2$.

A.4 Proof of Corollary 3.2

Define the fine-feedback values

$$\widehat{V}_j = \frac{1}{N}\sum_i \frac{\pi_j(A_i \mid X_i)}{\mu(A_i \mid X_i)} Y_i.$$

Under the population model and overlap, these are averages of independent variables in $[0, W]$ with means $V(\pi_j)$. By Hoeffding's inequality (Hoeffding, 1963), with probability at least $1-\eta$,

$$\max_j |\widehat{V}_j - V(\pi_j)| \le c_N(\eta) := W\sqrt{\frac{\log(2K/\eta)}{2N}}. \tag{17}$$

This is an event for the fine-feedback variables in the analysis; the learner need not observe them. Also $T_j = \widehat{V}_j - \widehat{V}_0$. On the routing event, if $\widehat{j}$ maximizes $\widetilde{T}_j$, then for every j,

$$\widehat{V}_j - \widehat{V}_{\widehat{j}} = T_j - T_{\widehat{j}} \le 2r_c(\delta/2).$$

Combining with (17) at $\eta = \delta/2$ proves (6). The two events need not be independent. Conditional routing validity gives its unconditional failure probability, and a union bound suffices. A library trained on an independent sample is handled by conditioning on that library, provided the same overlap bound holds.

The deterministic choice $v_{1/2} = N/4$, together with $B \le W$, gives the displayed high-probability order. An explicit expectation bound follows without choosing a vanishing failure probability. For $c = 1/2$, let

$$a = W\sqrt{J/(2N)}, \quad b = (J+1)W/(3N), \quad t_0 = \log(2p).$$

The maximum coding error has tail at most e^{-u} above $a\sqrt{t_0+u}+b(t_0+u)$ for $u \ge 0$. It is therefore stochastically bounded by $a\sqrt{t_0+E}+b(t_0+E)$, where E is a unit exponential variable. Since $\sqrt{t_0+E} \le \sqrt{t_0}+\sqrt{E}$ and $\mathbb{E}\sqrt{E} < 1$, its expectation is at most $a(\sqrt{t_0}+1)+b(t_0+1)$. Similarly, the maximum IPS error has expectation at most $W(\sqrt{\log(2K)}+1)/\sqrt{2N}$. The empirical selection inequality holds on every realization with the actual maximum errors, so

$$\mathbb{E}\,\mathrm{Reg}(\widehat{\pi}) \le \frac{2W}{\sqrt{2N}}(\sqrt{\log(2K)}+1) + 2a(\sqrt{t_0}+1) + 2b(t_0+1).$$

This is $O(W\sqrt{J\log(2K)/N} + JW\log(2K)/N)$. It gives the sample-size dependence at fixed K; Section 4 compares it with lower bounds when the library grows.

A.5 From value contrasts to more general finite comparisons

The conditional theorem actually estimates any supplied bounded matrix of weights against an arbitrary bounded response vector. Policy selection uses that property with importance-weight differences. For example, if a binary-label loss has the form $\ell(f_j(x), y) = u_j(x) + v_j(x)y$, its empirical loss differences can be encoded by taking $d_{ij} = v_j(x_i) - v_0(x_i)$ and adding the known offset $N^{-1}\sum_i(u_j - u_0)$. This uses the established loss decomposition from learning with label proportions (Busa-Fekete et al., 2023; Applebaum et al., 2026). The encoded weight matrix remains declared before the release.

B Release constraints and local score ranges

B.1 Release constraints and observed heterogeneity

Minimum group sizes. Small groups may not be releasable. Suppose each reporting window contains at least h rows. After raw routing and before any reward release, merge every nonempty group smaller than h into the largest group in that window. If all groups are small, merge the whole window. All nonempty released groups then have at least h rows, and the number of totals does not increase.

Write A_{ij} for the original row coefficient and A'_{ij} for the coefficient after reassignment. The metadata-computable charge is

$$\beta_j(c) = \frac{D_c}{N} \sum_i |A'_{ij} - A_{ij}|, \qquad \beta_0(c) = 0. \tag{18}$$

Let $\widetilde{T}'_j$ use repaired totals. Select the policy maximizing $\widetilde{T}'_j - \beta_j(c)$, including the baseline. The common radius need not be subtracted because it is the same for every candidate.

Theorem B.1 (A repair guarantee for every realized design)**.** *The repaired partition obeys the windows, disjointness, and minimum size h. On the event of Theorem 3.1,*

$$|\widetilde{T}'_j - T_j| \le r_c(\delta) + \beta_j(c) \quad \textit{for all } j.$$

Under the population assumptions, with probability at least $1 - \delta$, its selected policy satisfies, simultaneously for each comparator j,

$$V(\pi_j) - V(\widehat{\pi}) \le 2W\sqrt{\frac{\log(4K/\delta)}{2N}} + 2r_c(\delta/2) + 2\beta_j(c). \tag{19}$$

If r rows move, $\max_j \beta_j(c) \le 2JBD_c r/N$ and $r \le LJ2^b(h-1)$.

The last estimate is conservative but explicit. At fixed K, L, h, repair is a lower-order $O(N^{-1})$ charge. For many short windows or large minimum sizes it may dominate, and the method may return the baseline. Appendix F tests a regime in which the charge becomes severe.

Adapting to visible score variation. A few large importance weights can determine the global bound B. Localizing the code makes its variance depend on visible within-cell variation. Partition each window using only d into metadata cells C. Choose known centers $m_C \in \mathbb{R}^p$ and compute

$$B_C = \max_{i \in C} \|d_i - m_C\|_\infty.$$

Apply the same code to $d_i - m_C$, and add m_C to the group's decoding coefficients. In the proof, $JB^2 v_c$ is replaced by

$$J \sum_C B_C^2 v_{C,c}, \qquad v_{C,0} = \sum_{i \in C} y_i, \quad v_{C,1/2} = |C|/4, \quad v_{C,1} = |C| - v_{C,0}. \tag{20}$$

All quantities are obtained from metadata and released totals. With R cells, the maximum number of totals is $RJ2^b$. The charge (18) still applies when repairs stay within cells. Proposition B.2 gives the full bound.

For a simple zero-center construction, place $r_i = \|d_i\|_\infty$ into T dyadic levels with widths $B, B/2, \ldots, B/2^{T-1}$. Before merging small cells,

$$\sum_C |C| B_C^2 \le 4 \sum_i r_i^2 + N(B/2^{T-1})^2. \tag{21}$$

This deterministic bound uses the observed scores. Tiny cells can be merged within their window and their widths recomputed; the valid radius then uses the resulting widths. All original scores and responses are retained. Additional cells help when the variance reduction offsets their extra totals and repair costs.

B.2 Proof of Theorem B.1

All repairs are computed from the raw bucket counts and known coefficient vectors, before rewards are examined. In a window, let the largest nonempty raw bucket be the anchor. If it already has at least h rows, every other nonempty bucket of size below h is merged into it and every remaining bucket is legal. If the anchor has fewer than h rows, all buckets are small; merging the entire window is legal because that window contains at least h rows. Disjointness and window containment are preserved. No new total is added, and none of the original totals needs to be released before repair.

Let a_{ij} and a'_{ij} be row coefficients before and after the merge. Both estimators use the same known offset $c\sum_i d_{ij}$. For every reward vector,

$$|\widetilde{T}'_j - \widetilde{T}_j| = \frac{1}{N}\left|\sum_i (a'_{ij} - a_{ij})(y_i - c)\right| \le \frac{D_c}{N}\sum_i |a'_{ij} - a_{ij}| = \beta_j(c).$$

This deterministic inequality holds even though the repair is data-dependent through metadata and the random partition. Intersect it with the event of Theorem 3.1. The repaired estimates are not generally unbiased; the charge explicitly accounts for their possible bias.

Let $\widehat{j} \in \arg\max_j(\widetilde{T}'_j - \beta_j)$. For any comparator j,

$$T_{\widehat{j}} \ge \widetilde{T}'_{\widehat{j}} - \beta_{\widehat{j}} - r \ge \widetilde{T}'_j - \beta_j - r \ge T_j - 2\beta_j - 2r.$$

The population concentration event (17) adds $2c_N(\delta/2)$, proving (19). The charge of the selected policy does not appear because it was already subtracted in selection.

Each row coefficient belongs to $\{-JB, 0, JB\}$, so a moved row contributes at most $2JB$ per coordinate. Only rows in raw groups of size at most $h-1$ move; at most $LJ2^b$ such groups exist. This proves $\max_j \beta_j \le 2JBD_c r/N$ and $r \le LJ2^b(h-1)$. These bounds remain true in the all-small case. Large minimum sizes can make the bound uninformative, as the constraint experiment demonstrates.

B.3 A COMPUTABLE LOCAL-RANGE REFINEMENT

Proposition B.2 (Metadata-cell coding)**.** *Fix a partition into R metadata cells, each contained in one window, with known centers m_C and $B_C \ge \max_{i\in C}\|d_i - m_C\|_\infty$. Code residuals using Algorithm 1; each decoding vector is m_C plus the residual code. For fixed $c \in \{0, 1/2, 1\}$ the unrepaired estimate is unbiased, uses at most $RJ2^b$ totals, and with probability at least $1-\delta$ has maximum error at most*

$$r_c^{\mathrm{local}}(\delta) = \frac{\sqrt{2Jt\sum_C B_C^2 v_{C,c}}}{N} + \frac{2(J+1)B_{\max}D_c t}{3N}, \qquad t = \log(2p/\delta), \quad B_{\max} = \max_C B_C. \tag{22}$$

Repair within cells satisfies the same deterministic charge (18)*, provided each cell has at least h rows.*

Proof. For $i \in C$, let $e_i = d_i - m_C$ and $a_i = m_C + JB_C\xi_i\mathbf{1}\{j \in U_i\}$. Then $\mathbb{E}a_{ij} = d_{ij}$ and

$$a_{ij} - d_{ij} = JB_C\xi_{ij}\mathbf{1}\{j \in U_i\} - e_{ij}.$$

The known cell center cancels in the coding error. Consequently its coordinate variance is at most JB_C^2 and its magnitude at most $(J+1)B_C$. The proof of Theorem 3.1 applies with variance sum (20). When $B_C = 0$, use coefficient m_C directly and release one cell total; the conditional error is exactly zero. Repair within a cell leaves that cell's total reward mass invariant and the triangle inequality proves the charge. Taking these constructed quantities as inputs assumes no favorable geometry of the unknown rewards. □

Dyadic implementation. Let $B = \max_i r_i > 0$. The first $T-1$ cells in each window consist of $r_i \in (B/2^{t+1}, B/2^t]$, for $t = 0, \dots, T-2$, and the last contains $r_i \le B/2^{T-1}$. Set every center to zero. Outside the last level, $B_C < 2r_i$, and inside it $B_C = B/2^{T-1}$. Summing squared widths proves (21). Sorting or bucket assignment takes $O(Np)$ time including score norms. If a cell has fewer than h rows, merge it with the largest occupied cell in the same window and recompute its valid width. The displayed bound (22) uses the recomputed widths; the factor-four premerge estimate is not asserted to survive arbitrary merges.

Other metadata partitions. A balanced tree can split a cell along its widest contrast coordinate at a median, stopping after a specified number of leaves or when a child would have fewer than h rows. Coordinatewise midranges supply m_C and B_C. This procedure produces a valid finite partition whose radii are computed from the resulting cells. The ablation reports how much it reduces the error bound and where it fails to improve decisions. Likewise, shared random numbers within a selected block preserve each sign marginal in (3); this changes joint errors but not the coordinatewise bound. The main method uses independent signs.

B.4 A RANDOM-PARTITION COVARIANCE COMPARATOR

Random aggregate-label estimators use products of feature sums and label sums, with centering to remove cross-example terms (Busa-Fekete et al., 2023; Applebaum et al., 2026). We instantiate a finite-population version with an exact correction for the matched partition sizes.

Fix $n \geq 2$ rows within a window, their contrast values d_i and rewards y_i, and nonempty group sizes $h_1, \ldots, h_m$ summing to n, with $m \geq 2$. Uniformly randomly assign rows to these fixed sizes. Set

$$Q = \frac{1}{n} \sum_g \Big(\sum_{i \in g} d_i \Big) S_g, \qquad \rho = \frac{\sum_g h_g(h_g - 1)}{n(n-1)}, \qquad \widehat{T}_{\mathrm{cov}} = \frac{Q - \rho \bar{d} \sum_i y_i}{1 - \rho}.$$

For distinct rows, their co-membership probability is ρ, so

$$\mathbb{E}Q = \frac{1}{n} \sum_i d_i y_i + \frac{\rho}{n} \sum_{i \neq j} d_i y_j = (1 - \rho) T + \rho \bar{d} \sum_i y_i.$$

Thus $\mathbb{E}\widehat{T}_{\mathrm{cov}} = T$ for any fixed arrays, including unequal group sizes. Window estimates are combined with weights n/N. This calculation does not apply to content-dependent groups: their pairwise inclusion probabilities are not the same constant. Our score-cluster baseline instead uses bag-mean imputation, whose bias is part of the comparison.

C EXACT COUNT DIVERGENCE AND MINIMAX LOWER BOUNDS

We first compute the likelihood of a group count together with its observed metadata. Averaging this information over data-independent groups gives the lower bound; grouping by the policy comparison gives the contrasting upper bound.

C.1 LIKELIHOOD OF A SINGLE COUNT

For one bag of h rows in (7), write $U_i = 2Y_i - 1 \in \{-1, 1\}$. Under the null $\theta = 0$, (Z_i, U_i) are independent uniform signs. The full-data likelihood ratio is

$$L_\theta(Z, U) = \prod_{i=1}^{h} (1 + 2\theta Z_i U_i).$$

Only the total $S = (h + \sum_i U_i)/2$ is observed along with Z. Define elementary sign polynomials

$$e_k(z) = \sum_{I \subseteq [h] : |I| = k} \prod_{i \in I} z_i, \qquad \kappa_k(s) = \sum_{\ell=0}^{k} (-1)^{k-\ell} \binom{s}{\ell} \binom{h-s}{k-\ell}.$$

These are orthogonal Bernoulli polynomials (Diaconis & Griffiths, 2012). Terms with an invalid binomial index are zero. For any realization with exactly s positive U_i, $e_k(U) = \kappa_k(s)$. Exchangeability under the null gives

$$\mathbb{E}_0 \Big[\prod_{i \in I} U_i \,\Big|\, S \Big] = \frac{\kappa_k(S)}{\binom{h}{k}}, \qquad |I| = k.$$

Therefore the observed-data likelihood ratio is

$$\overline{L}_\theta(Z, S) = \mathbb{E}_0[L_\theta \mid Z, S] = \sum_{k=0}^{h} (2\theta)^k \frac{e_k(Z) \kappa_k(S)}{\binom{h}{k}}. \tag{23}$$

For independent uniform signs, products indexed by distinct subsets are orthogonal. Hence

$$\mathbb{E}_0[e_k(Z) e_\ell(Z)] = \binom{h}{k} \mathbf{1}\{k = \ell\}, \qquad \mathbb{E}_0[\kappa_k(S) \kappa_\ell(S)] = \binom{h}{k} \mathbf{1}\{k = \ell\}.$$

The second identity follows by writing $\kappa_k(S) = e_k(U)$. Under the null Z and S are independent, so squaring (23) and applying these identities yields

$$\mathbb{E}_0 \overline{L}_\theta^2 = \sum_{k=0}^{h} (4\theta^2)^k.$$

Subtracting one gives the count identity used for testing:

$$1 + \chi^2(P_{\theta,h} \| P_{0,h}) = \sum_{k=0}^{h} (4\theta^2)^k \le (1 - 4\theta^2)^{-1}. \tag{24}$$

The extra observation of (X_i, A_i) instead of just Z_i changes nothing: conditional on Z_i, the remaining sign is ancillary and independent of Y_i, with a law not depending on θ.

C.2 Proof of the data-independent partition lower bound

Condition on any fixed data-independent partition into $m \le M$ nonempty groups of sizes h_g. Independence across rows makes the group experiments independent, so

$$1 + \chi^2(P_\theta \| P_0) = \prod_{g=1}^{m} \sum_{k=0}^{h_g} (4\theta^2)^k \le (1 - 4\theta^2)^{-M}. \tag{25}$$

For a randomized partition independent of the data, the partition is itself part of the observation. Its law is common under both parameters. Conditional divergence therefore averages over partitions, preserving the same bound. The reasoning also covers any extra randomness used by the selector via data processing.

Take $t = 1/(8\sqrt{M})$. Then $4t^2 = 1/(16M)$ and

$$(1 - 1/(16M))^{-M} \le 16/15.$$

For completeness, the logarithm is $\sum_{k \ge 1} 1/(k16^k M^{k-1})$, which decreases with $M \ge 1$ and equals $\log(16/15)$ at $M = 1$. Using $\mathrm{TV}(P, Q) \le \frac{1}{2}\sqrt{\chi^2(P\|Q)}$ and the triangle inequality,

$$\mathrm{TV}(P_t, P_{-t}) \le \mathrm{TV}(P_t, P_0) + \mathrm{TV}(P_{-t}, P_0) \le \sqrt{1/15} < 1/2.$$

Let $q(\mathcal{D}) \in [0, 1]$ be the probability that a possibly randomized selector chooses π_+. Its expected regrets at t and $-t$ are $2t\mathbb{E}_t(1 - q)$ and $2t\mathbb{E}_{-t}q$. Their average is at least

$$t\{1 - \mathrm{TV}(P_t, P_{-t})\} \ge t/2 = \frac{1}{16\sqrt{M}}.$$

The maximum is at least the average. This proves (8). The two parameters $\pm t$ define fixed i.i.d. populations before the sample and partition are drawn; no outcome is selected in response to a realized grouping.

C.3 Two policy-aware totals and the full-information lower bound

Let $G_+ = \{i : Z_i = 1\}$ and $G_- = \{i : Z_i = -1\}$. From at most two totals obtain

$$\widehat{\Delta} = \frac{2}{N}(S_+ - S_-) = \frac{1}{N}\sum_i 2Z_iY_i, \qquad \mathbb{E}_\theta \widehat{\Delta} = 2\theta.$$

Choose π_+ if $\widehat{\Delta} \ge 0$ and π_- otherwise. On a wrong choice, $2|\theta| \le |\widehat{\Delta} - 2\theta|$. Also $\mathbb{E}(2ZY)^2 = 2$, so

$$\mathbb{E}_\theta \operatorname{Reg}(\widehat{\pi}) \le \mathbb{E}_\theta|\widehat{\Delta} - 2\theta| \le \sqrt{2/N}.$$

This proves the policy-aware upper bound in Theorem 4.1. If a fixed minimum group size is imposed, Theorem B.1 applies; its additional worst-case charge is $O(h/N)$. The two-group statement by itself does not assume a minimum size.

For individual outcomes, orthogonality before aggregation gives

$$1 + \chi^2(P_\theta^{\mathrm{full}} \| P_0^{\mathrm{full}}) = (1 + 4\theta^2)^N.$$

At $t = 1/(8\sqrt{N})$, the right side is at most $e^{1/16}$, and the same triangle argument gives $\mathrm{TV}(P_t^{\mathrm{full}}, P_{-t}^{\mathrm{full}}) \le \sqrt{e^{1/16} - 1} < 1/2$. The preceding decision argument yields $1/(16\sqrt{N})$. Any reward-blind grouping followed by total release is a randomized function of full data, so data processing extends this lower bound to every such grouping method. This matches the sample-size exponent, not all constants or policy-class dependencies.

D POLICY-COUNT LOWER BOUNDS FOR OUTCOME-BLIND GROUPING

We fix the policy library and testing populations before sampling. The encoder then partitions the N rows using the complete metadata sample and reward-independent coins. The selector observes all contexts, actions, assignments, and totals. To include fitted clustering and trees, we first control a count for every fixed group composition, then bound the best partition of the realized metadata. This order keeps the effect of selecting a partition inside the information bound.

D.1 TWO-SIDED COUNT INFORMATION FOR AN ARBITRARY GROUP COMPOSITION

Lemma D.1 (Directional information in a reward count)**.** *Fix $z \in \{-1,1\}^h$, $h \geq 1$. Let $P_{\theta,z}$ be the law of $S = \sum_{i=1}^h Y_i$ for independent $Y_i \sim \mathrm{Bernoulli}(1/2 + \theta z_i)$, and set $x^2 = (\sum_i z_i)^2/h$. For $|\theta| \leq 1/8$,*

$$\frac{\theta^2 x^2 + \theta^4}{40} \leq D_{\mathrm{KL}}(P_{\theta,z} \| P_{0,z}) \leq 20\theta^2 x^2 + 96\theta^4. \tag{26}$$

In particular, the simpler bound used in Theorem 4.2 remains valid:

$$D_{\mathrm{KL}}(P_{\theta,z} \| P_{0,z}) \leq \log\{1 + \chi^2(P_{\theta,z} \| P_{0,z})\} \leq 20\theta^2(1 + x^2). \tag{27}$$

Both inequalities hold for every fixed group composition.

Proof. Upper bound. Write $q = 4\theta^2 \leq 1/16$ and $e_k(z) = \sum_{|I|=k} \prod_{i \in I} z_i$. The null Bernoulli expansion from Appendix C, now without averaging the signs, gives

$$1 + \chi^2(P_{\theta,z} \| P_{0,z}) = \sum_{k=0}^h q^k a_k^2, \qquad a_k = \frac{e_k(z)}{\sqrt{\binom{h}{k}}}.$$

For $h = 1$, this equals $1 + q$ and $x^2 = 1$, so the upper claims follow. Suppose $h \geq 2$. These normalized Krawtchouk coefficients (Diaconis & Griffiths, 2012; 2019) obey

$$a_{k+1} = \frac{s}{\sqrt{(k+1)(h-k)}} a_k - \sqrt{\frac{k(h-k+1)}{(k+1)(h-k)}} a_{k-1}, \qquad s = \sum_i z_i.$$

Put $x = |s|/\sqrt{h}$ and $H_k(x) = \mathbb{E}(x + G)^k/\sqrt{k!}$ for $G \sim N(0,1)$. These moments are nonnegative and satisfy $H_0 = 1$, $H_1 = x$, and $H_{k+1} = xH_k/\sqrt{k+1} + \sqrt{k/(k+1)} H_{k-1}$. For $k < h/2$, the two recurrence coefficients for $|a_{k+1}|$ are at most $\sqrt{2}x/\sqrt{k+1}$ and $\sqrt{2k/(k+1)}$. Induction gives $|a_k| \leq 2^{k/2} H_k(x)$ for $0 \leq k \leq \lfloor h/2 \rfloor$. Also $|a_{h-k}| = |a_k|$. Pairing complementary indices and retaining $q^h \leq q^2$ yields

$$1 + \chi^2 \leq 1 + q^2 + 2 \sum_{k \geq 1} (2q)^k H_k(x)^2.$$

For independent standard normals G, G' and $0 \leq t < 1$, absolute integrability justifies

$$F(t,x) := \sum_{k \geq 0} t^k H_k(x)^2 = \mathbb{E} e^{t(x+G)(x+G')} = \frac{\exp\{tx^2/(1-t)\}}{\sqrt{1-t^2}}.$$

Since $F \geq 1$, $1 + q^2 + 2(F-1) \leq (1+q^2)F^2$. Thus

$$\log(1 + \chi^2) \leq \log(1+q^2) - \log(1-4q^2) + \frac{4qx^2}{1-2q} \leq 6q^2 + 5qx^2.$$

Here $q \leq 1/16$, $-\log(1-4q^2) \leq 4q^2/(1-4q^2)$, and $4/(1-2q) < 5$. Substituting $q = 4\theta^2$ proves the upper bound in (26); it is at most $20\theta^2(1+x^2)$. The case $\theta = 0$ is immediate.

Lower bound. The variational formula for relative entropy and the null binomial moment-generating bound give

$$D_{\mathrm{KL}}(P_{\theta,z} \| P_{0,z}) \geq \sup_t \{t\theta s - ht^2/8\} = 2\theta^2 x^2.$$

This controls the mean shift. To detect information remaining when the mean cancels, set $T = (2S - h)/\sqrt{h}$. Direct calculation gives

$$\mathbb{E}_\theta T^2 - \mathbb{E}_0 T^2 = 4\theta^2(x^2 - 1).$$

If $x^2 \le 1/2$, the magnitude is at least $2\theta^2$. Write $U = T - \mathbb{E}_\theta T$ and $v = 1 - 4\theta^2$. Its summands are centered Bernoulli variables with equal variance before normalization. Expanding gives $\mathbb{E}_\theta U^4 = 3v^2 + (4v - 6v^2)/h \le 3 + 1/h \le 4$. Furthermore $|\mathbb{E}_\theta T| = 2|\theta|x \le 1/(4\sqrt{2})$, so $\mathbb{E}_\theta T^4 < 33$, while $\mathbb{E}_0 T^4 \le 3$. For $H^2(P, Q) = \sum_s (\sqrt{P(s)} - \sqrt{Q(s)})^2$, Cauchy–Schwarz implies

$$|\mathbb{E}_\theta T^2 - \mathbb{E}_0 T^2|^2 \le 2H^2(P_{\theta,z}, P_{0,z})(\mathbb{E}_\theta T^4 + \mathbb{E}_0 T^4) \le 72H^2(P_{\theta,z}, P_{0,z}).$$

Using $D_{\mathrm{KL}} \ge H^2$, we obtain $D_{\mathrm{KL}} \ge \theta^4/18$. Combining this with $2\theta^2 x^2$ gives $D_{\mathrm{KL}} \ge \theta^2 x^2 + \theta^4/36 \ge (\theta^2 x^2 + \theta^4)/40$. If $x^2 > 1/2$, the mean-shift bound alone gives the stated inequality because $\theta^2 \le 1/64$. □

For a fixed partition G with m nonempty groups, define P_G as the orthogonal projection onto groupwise-constant vectors. Independent group counts add their relative entropies, hence (26) gives

$$\frac{\theta^2 \|P_G z\|^2 + m\theta^4}{40} \le D_{\mathrm{KL}}(P_{\theta,z,G} \| P_{0,z,G}) \le 20\theta^2 \|P_G z\|^2 + 96m\theta^4. \tag{28}$$

This is a finite-signal characterization of the *count-to-null experiment* in this testing family, conditional on the entire metadata and partition. It is not a general formula for policy regret: a balanced group can distinguish signal magnitude from zero while losing its sign, as shown below.

D.2 A UNIFORM BOUND OVER ALL METADATA-FITTED PARTITIONS

For a nonempty partition G, let P_G be its $N \times N$ orthogonal projection onto groupwise-constant vectors: $(P_G)_{ii'} = \mathbf{1}\{i, i' \text{ share a group } g\}/|g|$. For an $N \times p$ sign matrix Z, define its preserved comparison energy

$$\mathcal{A}(G, Z) = \|P_G Z\|_F^2 = \sum_{j=1}^{p} \sum_g \frac{(\sum_{i \in g} Z_{ij})^2}{|g|}.$$

This is an observable statistic, not an assumed condition on the quality of the partition.

Lemma D.2 (Uniform partition-energy budget)**.** *If all entries of Z are independent uniform signs, then for $M \ge 2$,*

$$\mathbb{E} \sup_{G:\, |G| \le M} \mathcal{A}(G, Z) \le 4N \log M + 2pM \log 2. \tag{29}$$

The supremum includes every joint metadata-dependent partition.

Proof. For any fixed projection P of rank $m \le M$ and $0 < \lambda < 1/2$, Gaussian linearization and $\cosh u \le e^{u^2/2}$ give, for one sign column Z_j,

$$\mathbb{E}_{Z_j} e^{\lambda \|PZ_j\|^2} = \mathbb{E}_{Z_j, G} e^{\sqrt{2\lambda} G^\top P Z_j} \le \mathbb{E}_G e^{\lambda \|PG\|^2} = (1 - 2\lambda)^{-m/2}.$$

The p columns are independent for fixed P. At most M^N labeled assignments can induce the partition, so

$$\mathbb{E} \exp\{\lambda \sup_G \|P_G Z\|_F^2\} \le M^N (1 - 2\lambda)^{-pM/2}.$$

Jensen's inequality followed by $\lambda = 1/4$ proves (29). Taking a supremum before expectation is precisely what permits fitting the partition to the realized metadata. No independence across its assignments is used. □

D.3 THE LOCAL INFORMATION MATRIX AND WHAT ITS TRACE OMITS

For a fixed real metadata matrix $Z \in \mathbb{R}^{N\times d}$, consider independent Bernoulli means $1/2 + Z_i^\top \beta$ in a neighborhood of $\beta = 0$. Conditional on Z and a fixed outcome-blind partition, the full-data score is $4\sum_i Z_i(Y_i - 1/2)$. The score after taking a statistic is its conditional expectation (Pollard, 2012). Under the null, $\mathbb{E}[Y_i \mid S_g] = S_g/|g|$, so the count score and Fisher information are

$$\dot{\ell}_G(0) = 4\sum_g \frac{\sum_{i\in g} Z_i}{|g|}(S_g - |g|/2), \qquad \mathcal{I}_G(0) = 4Z^\top P_G Z.$$

Indeed, independent null counts have variances $|g|/4$. Thus

$$\mathcal{I}_{\text{individual}}(0) - \mathcal{I}_G(0) = 4Z^\top(I - P_G)Z \succeq 0.$$

For a direction u, all of its local Fisher information is retained exactly when $(I - P_G)Zu = 0$: its rowwise signal must be constant within every group. This specializes the classical score-projection identity to count feedback. Lemma D.1 adds two-sided finite-signal control for the sign directions used by the lower bound.

Equal total energy need not preserve the same comparisons. Take the eight rows

$$Z = \begin{pmatrix} -1 & -1 \\ -1 & 1 \\ 1 & -1 \\ 1 & 1 \\ -1 & -1 \\ -1 & 1 \\ 1 & -1 \\ 1 & 1 \end{pmatrix}$$

and partitions with assignment vectors $g^{(1)} = (0,0,0,0,0,0,1,2)$ and $g^{(2)} = (0,0,0,1,0,1,1,2)$. Both have three groups, and direct summation gives

$$Z^\top P_{g^{(1)}} Z = \begin{pmatrix} 8/3 & 0 \\ 0 & 2 \end{pmatrix}, \qquad Z^\top P_{g^{(2)}} Z = \frac{7}{3}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Their traces are equal to $14/3$, but the second loses direction $u = (1,-1)$ entirely at first order. The loss is stronger than a local degeneracy: for means $1/2 \pm t(Z_{i1} - Z_{i2})$, $0 < t \le 1/8$, each group of $g^{(2)}$ contains balanced positive and negative signals. Exchanging their signs preserves the complete group-count law. In $g^{(1)}$, singleton row 7 has different means, so the count laws differ. Thus the scalar trace $\mathcal{A}$ does not determine which policy comparison is distinguishable.

Even zero first-order energy need not mean no information against a null. For $h = 2$, $z = (1,-1)$, the count probabilities are

$$P_{\theta,z}(S = 0,1,2) = (1/4 - \theta^2,\ 1/2 + 2\theta^2,\ 1/4 - \theta^2).$$

They distinguish $|\theta|$ from zero at order θ^4, as in (26), but are identical for θ and $-\theta$. Decision difficulty therefore depends on the alternatives being compared, as well as on the energy and count-to-null divergence.

D.4 PROOF OF THEOREM 4.2

Sample X uniformly from $\{-1,1\}^p$, and log an independent uniform sign action A. Let $Z = AX$. The rows of Z have mutually independent uniform coordinates. The reference policy π_0 is uniform; policy π_j chooses action X_j. Under hypothesis $j \in [p]$,

$$m_j(x,a) = 1/2 + \theta a x_j, \qquad 0 < \theta \le 1/8. \tag{30}$$

The policies, logging law, and populations are fixed before sampling. Policy j has value $1/2 + \theta$; all others have value $1/2$. Importance ratios are at most two.

We may additionally reveal every reward-independent coin, since this only strengthens the learner. Let $\mathcal{M}$ contain these coins and the full metadata. The partition is determined by $\mathcal{M}$, whose law is

identical under every hypothesis and the null population P_0 of mean $1/2$. Conditional on $\mathcal{M}$, the reward counts in disjoint groups are independent. Apply Lemma D.1 to each fixed group and use the chain rule for relative entropy:

$$D_{\mathrm{KL}}(P_j \| P_0) \le 20\theta^2 \mathbb{E}\left[M + E_j(G, Z)\right], \qquad E_j(G, Z) = \sum_g \frac{(\sum_{i \in g} Z_{ij})^2}{|g|}. \tag{31}$$

This bounds the joint experiment that retains all metadata, not merely the marginal law of the counts. Averaging over j and invoking Lemma D.2,

$$\frac{1}{p}\sum_j D_{\mathrm{KL}}(P_j \| P_0) \le 20\theta^2 \left[(1 + 2\log 2)M + \frac{4N \log M}{p}\right] \le 80\theta^2 D, \qquad D = M + \frac{N \log M}{p}.$$

Set $\theta = \min\{1/8, \sqrt{\log p/(320D)}\}$. Fano's inequality (Cover & Thomas, 2006; Tsybakov, 2009) then gives average classification error at least $1 - (\frac{1}{4}\log p + \log 2)/\log p \ge 1/4$. A randomized mixture over the library induces a classifier by drawing an index, with expected regret exactly θ times its probability of not drawing the optimal index. If index zero is drawn, assigning it to any testing label can only reduce the classification error. Thus

$$\sup_j \mathbb{E}_j \operatorname{Reg}(\widehat{\pi}) \ge \frac{1}{4} \min\left\{\frac{1}{8}, \sqrt{\frac{\log p}{320(M + N \log M/p)}}\right\}. \tag{32}$$

This proves (9) with an absolute constant. The worst hypothesis is selected from fixed i.i.d. populations, not after observing data. Metadata-fitted splitting, merging, and randomization are already included in the supremum over final partitions. Reward-dependent assignments, overlapping releases, or uncounted pilot outcomes are not part of this experiment.

D.5 Proof of Corollary 4.4

Take one window and $b = \lfloor \log_2 p \rfloor \ge 2$. Algorithm 1 has $J = \lceil p/b \rceil \le 2p/b$ and uses $M_{\log} = J2^b \le p^2$ keys. Let the allowed cap satisfy $M_{\log} \le M \le p^2$. Set $c = 1/2$ and use a population-valid width $B \le W \le 2$, so the encoding kernel is specified before the log. Conditional on the log, the proof of Theorem 3.1 gives a sub-gamma bound for each signed error, with variance proxy $v \le JW^2/(4N)$ and scale $c \le (J+1)W/(6N)$. For $t = \log(2p)$, the log-sum-exp inequality yields

$$\mathbb{E}_{\mathrm{route}} \max_j |\widetilde{T}_j - T_j| \le \inf_{0<\lambda<1/c}\left\{\frac{t}{\lambda} + \frac{v\lambda}{2(1-c\lambda)}\right\} \le \sqrt{2vt} + ct.$$

Here $B \le W$ is deterministic. Averaging over the log gives

$$\mathbb{E} \max_j |\widetilde{T}_j - T_j| \le C_0\left(\sqrt{\frac{J \log(2p)}{N}} + \frac{(J+1)\log(2p)}{N}\right).$$

The ordinary IPS uniform deviation adds $O(\sqrt{\log K/N})$. Hence selection regret is at most $C_1(\sqrt{p/N} + p/N)$, uniformly over all K-policy libraries with overlap at most two. Since $2^b \ge 2b$, we have $M_{\log} \ge 2p$. Under $N \ge pM/\log p$ this implies $N \ge p$ and hence $p/N \le \sqrt{p/N}$. In (32),

$$D = M + \frac{N \log M}{p} \le \frac{N \log p}{p} + \frac{2N \log p}{p} = \frac{3N \log p}{p}.$$

The capped testing bound is therefore at least $\sqrt{p/N}/(4\sqrt{960})$; the cap does not weaken this inequality because $N \ge p$. This proves the claimed rate throughout $M_{\log} \le M \le p^2$. Finally, $p/2 < 2^b \le p$ and $b = \Theta(\log p)$ give $M_{\log} = \Theta(p^2/\log p)$. At its actual code budget, the sufficient sample threshold is thus of order $p^3/(\log p)^2$. The choice $M = p^2$ recovers the more conservative threshold $p^3/\log p$. The lower bound covers every outcome-blind batch partition; the attaining code remains rowwise.

D.6 NEAR-LINEAR BUDGETS: A CODE BOTTLENECK AND THE REMAINING GAP

We first give the general linear-budget bracket, then distinguish it from the stronger large-sample result in Table 1. Use one window, $p \geq 4$, $M = 2p$, overlap $W \leq 2$, and $N \geq p^2/\log p$. In (32),

$$D = 2p + \frac{N \log(2p)}{p} \leq \frac{4N \log p}{p},$$

since $2p \leq 2N \log p/p$ and $\log(2p) \leq \frac{3}{2} \log p$. The capped testing choice therefore yields a lower bound $c\sqrt{p/N}$ for an absolute $c > 0$. For the upper bound, set $b = 1$, so $J = p$ and the code uses $2p$ keys. The expected-maximum calculation above gives

$$\mathbb{E}\,\mathrm{Reg}(\widehat{\pi}) \leq C \left\{ \sqrt{\frac{p \log(2p)}{N}} + \frac{p \log(2p)}{N} \right\}.$$

The second term is bounded by a constant times the first on this sample range, giving $O(\sqrt{p \log p/N})$. These two bounds leave a factor $\sqrt{\log p}$ in regret, or $\log p$ on the squared-error scale. Theorem 4.3 closes it when $N \geq Cp^{5/2} \log(2p)$; the smaller-sample transition is not settled by this calculation. No positive minimum group size is imposed here; Section E.4 charges repairs separately.

Proposition D.3 (All-decoder lower bound for the block encoder)**.** *Use one window and Algorithm 1 with fixed block size $1 \leq b \leq p$, $J = \lceil p/b \rceil$, width $B = 1$, and $M_0 = J2^b$ keys. On the fixed p-comparison family* (30)*, for $p \geq 4$, every selector receiving the complete metadata, assignments, and totals satisfies*

$$\sup_j \mathbb{E}_j\,\mathrm{Reg}(\widehat{\pi}) \geq \frac{1}{4} \min \left\{ \frac{1}{8}, \sqrt{\frac{\log p}{160(M_0 + N/J)}} \right\}. \tag{33}$$

For $N \geq JM_0$, the worst-case rate for this encoder is $\Theta(\sqrt{J \log p/N})$. In particular, at $b = 1$, $M_0 = 2p$, and $N \geq 2p^2$, the logarithmic factor in its $\Theta(\sqrt{p \log p/N})$ rate cannot be removed by changing only the decoder.

Proof. In this family $d_{ij} = Z_{ij}$. Conditional on a row's chosen block and sign key, coordinate j is fixed to its key sign if its block was chosen; otherwise it remains an independent uniform sign. This property also holds with independent fair padding bits in a final short block. Thus rows assigned to a key containing j contribute their group size to E_j. The expected total size of such keys is N/J. Conditional on any nonempty other key of size h, the squared sum of its jth signs has expectation h, contributing one to E_j. Therefore

$$\mathbb{E}E_j \leq N/J + M_0.$$

(The exact expression for equal-size full blocks is $N/J + (M_0 - 2^b)[1 - (1 - 1/M_0)^N]$.) By (31),

$$D_{\mathrm{KL}}(P_j \| P_0) \leq 20\theta^2(2M_0 + N/J) \leq 40\theta^2(M_0 + N/J).$$

Set $\theta = \min\{1/8, \sqrt{\log p/[160(M_0 + N/J)]}\}$. Fano gives error at least $1/4$, and the same classifier reduction as in Theorem 4.2 proves (33), including arbitrary biased, nonlinear, or randomized selectors. For $N \geq JM_0$, the lower bound is $\Omega(\sqrt{J \log p/N})$. The expected-maximum calculation in the proof of Corollary 4.4 gives the matching upper bound; $M_0 \geq 2p$ keeps the linear remainder no larger than the square-root term on this range. □

This proposition is a restriction on a specified encoder, whereas Theorem 4.2 ranges over all outcome-blind batch partitions. The stronger all-encoder converse of Theorem 4.3 applies in its sample-rich range; this proposition has the sharper sample threshold for the specified code alone.

D.7 SHARPER IDENTITIES FOR PREDECLARED KERNELS

The next two identities give an alternative, sharper-constant argument when assignments are independently generated by one fixed kernel. They are not assumptions of the batch-adaptive theorem.

Biased conditional metadata.

Lemma D.4 (Count divergence under biased metadata). *Let $Z_1, \dots, Z_h$ be independent signs with $\mathbb{E}Z_i = m \in [-1, 1]$. Given Z, let Y_i be independent* $\mathrm{Bernoulli}(1/2 + \theta Z_i)$ *variables, and observe (Z, S) with $S = \sum_i Y_i$. For $q = 4\theta^2 < 1$, the observation laws satisfy*

$$1 + \chi^2(P_{\theta,h,m} \| P_{0,h,m}) = \sum_{k=0}^{h} q^k \sum_{u=0}^{\min(k,h-k)} \binom{k}{u}\binom{h-k}{u} m^{2u} \le \frac{\exp\{hqm^2/(1-q)\}}{1-q}. \quad (34)$$

The same formula holds if each row includes additional public metadata, provided its marginal Z has the stated law and the conditional response depends only on Z.

Proof. Under the null, write $V_i = 2Y_i - 1$. For $0 \le k \le h$, put $e_k(z) = \sum_{|I|=k} \prod_{i \in I} z_i$ and $F_k(S) = \mathbb{E}_0[\prod_{i=1}^{k} V_i \mid S]$. Exchangeability of the null signs gives $\mathbb{E}_0 F_k F_l = \mathbf{1}\{k = l\}/\binom{h}{k}$, by the orthogonality calculation in Appendix C. Conditional expectation of the full likelihood ratio therefore gives

$$L_\theta(Z, S) = \sum_{k=0}^{h} (2\theta)^k e_k(Z) F_k(S), \qquad 1 + \chi^2 = \sum_{k=0}^{h} q^k \frac{\mathbb{E}e_k(Z)^2}{\binom{h}{k}}.$$

For two k-subsets I, J, let $u = |I \setminus J| = |J \setminus I|$. The expectation of the product is m^{2u}. For each I, exactly $\binom{k}{u}\binom{h-k}{u}$ choices of J have this overlap. This proves the equality. For the bound, use $\binom{h-k}{u} \le h^u/u!$ and extend the nonnegative sum in k:

$$\sum_{k=0}^{h} q^k \sum_{u} \binom{k}{u}\binom{h-k}{u} m^{2u} \le \sum_{u \ge 0} \frac{(hm^2)^u}{u!} \sum_{k \ge u} \binom{k}{u} q^k = \frac{1}{1-q} \exp\left(\frac{hqm^2}{1-q}\right).$$

Additional metadata does not enter the response likelihood ratio; averaging over it yields the same expression. □

Assignment information.

Lemma D.5 (An entropy budget for policy directions). *If Z is uniform on $\{-1, 1\}^p$ and G has at most M values, define $a_j = \mathbb{E}[(\mathbb{E}[Z_j \mid G])^2]$. Then*

$$\sum_{j=1}^{p} a_j \le 2I(Z; G) \le 2 \log M. \quad (35)$$

Proof. For a sign of mean m, binary Pinsker gives (Cover & Thomas, 2006; Tsybakov, 2009) $D_{\mathrm{KL}}(P_Z \| \mathrm{Unif}\{-1, 1\}) \ge m^2/2$. Relative entropy to a product distribution is at least the sum of its marginal relative entropies. Apply this to the conditional law of Z given each $G = g$, then average over g. The product reference is the unconditional uniform law of Z, giving $I(Z; G)$. Finally $I(Z; G) \le H(G) \le \log M$. This is an information budget for the *metadata assignment*, not a bound on the bit precision of the released reward totals. □

E A REUSABLE RELEASE FOR AN INFINITE POLICY CLASS

Fix a measurable gate map $\phi : \mathcal{X} \to \Delta_r$ independently of the evaluation log. Its coordinates may be nonlinear and overlapping. The policy class is $\pi_U(a \mid x) = \sum_{l=1}^{r} \phi_l(x) U_{la}$, where every row of U is in Δ_A. This defines the policies being learned, not a model for $m(x, a)$. The output head U is chosen after the reward totals are available. Gates trained on an independent pilot are allowed by conditioning on that pilot.

Algorithm 2 One moment release, post-release head learning

Require: Public gates ϕ, logged actions and propensities, windows, floor κ.
1: **for** each row i, before observing its reward **do**
2: Accept with probability κ/μ_i using a fresh independent coin.
3: If accepted, sample $l \sim \phi(x_i)$ and assign (w_i, l, a_i); otherwise assign the residual key of w_i.
4: **end for**
5: Merge undersized keys within each window if required.
6: Collector releases each nonempty key's unweighted total once.
7: Form $\widehat{Q}$ by (11).
8: Fit any head objective over $(\Delta_A)^r$ using this same table.

E.1 Conditional correctness and optimization

Use a known $0 < \kappa \le \mu(a \mid x)$ on the target support. For row i, accept with probability κ/μ_i; if accepted, draw $L_i \sim \phi(x_i)$ and route to key (w_i, L_i, a_i). Otherwise use the window's residual key. Define

$$\widehat{Q}_{la} = \frac{1}{N\kappa} \sum_{w} S_{w,l,a}, \qquad \widehat{V}(U) = \sum_{l,a} U_{la} \widehat{Q}_{la}.$$

For the raw release before merging, every fixed log and response array satisfies

$$\mathbb{E}_{\text{route}} \widehat{Q}_{la} = \frac{1}{N} \sum_{i} \phi_l(x_i) \frac{\mathbf{1}\{a_i = a\}}{\mu_i} y_i.$$

Thus each fixed head has an unbiased estimate of its individual-feedback value. Choosing the head after release additionally requires the uniform bounds proved below.

An unconstrained head maximizes the linear objective by assigning each gate to an action attaining $\max_a \widehat{Q}_{la}$. This takes $O(rA)$ work. Linear constraints on U give an explicit linear program with rA variables. For known gate weights p_l and entropy penalty $\lambda > 0$, maximizing $\sum_{la} U_{la} \widehat{Q}_{la} + \lambda \sum_l p_l H(U_l)$ gives $U_{la} \propto \exp\{\widehat{Q}_{la}/(\lambda p_l)\}$ for $p_l > 0$. These optimizations work directly with rA variables and use the same table for every head objective.

For a fixed entropy weight $\lambda \ge 0$ and normalized nonnegative gate weights summing to one, let $\mathcal{H}(U)$ denote the weighted entropy, so $0 \le \mathcal{H}(U) \le \log A$. A maximizer of $\widehat{V}(U) + \lambda \mathcal{H}(U)$ obeys

$$V(U^\star) - V(\widehat{U}_\lambda) \le 2 \sup_U |\widehat{V}(U) - V(U)| + \lambda \log A.$$

Indeed, empirical optimality bounds $\widehat{V}(U^\star) - \widehat{V}(\widehat{U}_\lambda)$ by $\lambda\{\mathcal{H}(\widehat{U}_\lambda) - \mathcal{H}(U^\star)\} \le \lambda \log A$. The same event holds for every declared penalty, but selecting a penalty does not remove this optimization term.

E.2 Proof of Theorem 5.1

Under i.i.d. logging, put $p_l = \mathbb{E}\phi_l(X)$ and $Q_{la} = \mathbb{E}[\phi_l(X) m(X, a)]$. The per-row contribution to $\widehat{Q}_{la}$ is $Z_{i,la} = \kappa^{-1} \mathbf{1}\{G_i = (l, a)\} Y_i$, with the window coordinate summed out. Cancellation of the logging propensity yields

$$\mathbb{E} Z_{i,la} = Q_{la}, \qquad 0 \le Z_{i,la} \le 1/\kappa, \qquad \mathbb{E} Z_{i,la}^2 \le p_l/\kappa.$$

Rows remain independent regardless of the reward distribution within $[0, 1]$. Let $E_{la} = \widehat{Q}_{la} - Q_{la}$. For every head, including one chosen from the release,

$$|\widehat{V}(U) - V(U)| \le \sum_l \max_a |E_{la}|. \tag{36}$$

For each l, the centered bounded-sum moment bound gives

$$\mathbb{E} \max_a |E_{la}| \le \sqrt{\frac{2 p_l \log(2A)}{N\kappa}} + \frac{2 \log(2A)}{3N\kappa}.$$

For completeness, apply the logarithmic moment bound to the $2A$ signed errors and optimize the common moment parameter; the factor two in the linear term is conservative. Summing, using $\sum_l \sqrt{p_l} \le \sqrt{r}$, and applying the usual two-uniform-errors comparison proves

$$\mathbb{E}\{V(U^\star) - V(\widehat{U})\} \le 2\sqrt{\frac{2r\log(2A)}{N\kappa}} + \frac{4r\log(2A)}{3N\kappa}. \tag{37}$$

The bound uses the shared gate–action coordinates rather than a union bound over a head list. The same deterministic inequality holds for any constrained subset of heads, with its own comparator and any explicit optimization error added.

A high-probability version follows directly from Bernstein. For $t = \log(2rA/\delta)$, with probability at least $1 - \delta$,

$$\sup_{U \in (\Delta_A)^r} |\widehat{V}(U) - V(U)| \le \sqrt{\frac{2rt}{N\kappa}} + \frac{2rt}{3N\kappa}. \tag{38}$$

This event is uniform over the entire class, so optimization, regularization and repeated comparisons within that same class do not require a new release. Choosing a new gate map from these outcomes is not covered by this event.

E.3 A MATCHING BINARY-ACTION LOWER BOUND

The lower-bound experiment supplies only the N evaluation rows and fixed one-hot gates. Let X be uniform on $[r]$. Action 1 is logged with probability $\kappa \le 1/2$, and action 0 otherwise. Under $v \in \{-1, 1\}^r$, set

$$m_v(l, 1) = 1/2 + \varepsilon v_l, \qquad m_v(l, 0) = 1/2.$$

Every optimal policy is a deterministic head in the stated class. The hypercube testing argument (Tsybakov, 2009) compares adjacent coordinates. Flipping one coordinate of v changes only responses with that context and action 1. For $\varepsilon \le 1/4$, the single Bernoulli KL is at most $16\varepsilon^2$, so the N-sample full-data KL is at most $16N\kappa\varepsilon^2/r$. At $\varepsilon = \min\{1/4, \sqrt{r/(128N\kappa)}\}$, total variation is at most $1/4$. Pairing adjacent sign vectors, even a randomized head has average wrong-action probability at least $3/8$ in every context. Thus average regret is at least $3\varepsilon/8$, and some fixed population attains this lower bound. Group totals are a function of the full log, so the bound applies to every release mechanism as well. Together with (37) at $A = 2$, this matches the r, N, κ dependence up to absolute constants, with regret truncated at one.

E.4 MINIMUM-SIZE REPAIRS AND REPRESENTATION SCOPE

If each reporting window has at least h rows, merge its nonempty undersized raw keys into a largest key before releasing any reward, exactly as in Appendix B. Each head assigns decoder coefficients in $[0, 1/\kappa]$ to signal keys and zero to residual keys. If R_{move} rows change keys, then for every reward array and *all* heads,

$$|\widehat{V}_{\mathrm{repaired}}(U) - \widehat{V}(U)| \le \frac{R_{\mathrm{move}}}{N\kappa}, \qquad R_{\mathrm{move}} \le L(rA + 1)(h - 1).$$

A learned head adds at most twice this charge on the uniform-deviation event. For expected regret, the additive term is $2\mathbb{E}R_{\mathrm{move}}/(N\kappa)$, bounded deterministically by $2L(rA+1)(h-1)/(N\kappa)$. Windows do not alter raw moments when their signal keys are pooled, but more windows increase the total budget and repair burden.

One-hot gates give context-cell policies; smooth gates give mixtures that vary continuously across contexts. Both constructions allow arbitrary reward means. The comparator is the best head in the declared class. The binary lower bound uses a subclass to establish the necessity of the r, N, κ dependence, while the upper bound applies to the full fixed-gate class.

E.5 WHY UNRESTRICTED GATE SELECTION HAS AN ACQUISITION COST

Uniform control over heads of one representation does not automatically extend to a collection of unrelated representations. The following consequence of the count lower-bound experiment separates model-selection complexity from the information needed to collect those models' values.

Corollary E.1 (Two gates per map can still create a large release problem)**.** *For every $F \geq 4$, there is a declared family of F two-gate maps and uniform two-action logging such that every outcome-blind batch partition into at most $M \geq 2$ totals, followed by any learner over the union of their stochastic-head classes or mixtures of these classes, has, on some fixed i.i.d. reward population,*

$$\mathbb{E}\,\mathrm{Reg}(\widehat{\pi}) \geq c \min\left\{1, \sqrt{\frac{F \log F}{FM + N \log M}}\right\}. \tag{39}$$

The encoder may know all the maps and inspect the complete metadata batch before grouping.

Proof. Let X be uniform on $\{-1,1\}^F$ and log an independent uniform sign action A. Declare $\phi^{(l)}(x) = (\mathbf{1}\{x_l = 1\}, \mathbf{1}\{x_l = -1\})$ for $l \in [F]$. Under hypothesis j, take $m_j(x,a) = 1/2 + \theta a x_j$. For a stochastic head on map l, let $a_l(x_l) \in [-1,1]$ denote its conditional mean action. Its value is

$$\frac{1}{2} + \theta \mathbb{E}[X_j a_l(X_l)].$$

For $l \neq j$, this equals $1/2$ by independence and symmetry. For $l = j$, it is at most $1/2 + \theta$, attained by choosing action X_j. Thus the best policy in the union has value $1/2 + \theta$. For any output mixture with weights w_l on the maps, its regret is at least $\theta(1 - w_j)$. Drawing a map index according to those weights produces a classifier with error $1 - w_j$; a single output head is the degenerate case. If the policy has multiple representations, any fixed measurable decomposition into this finite-dimensional convex hull suffices. The lower-bound experiment and KL bound from Appendix D apply verbatim with $p = F$, and Fano proves (39). □

At fixed M, large N, and fixed two-action coverage, the bound is of order $\sqrt{F \log F / N}$ rather than $\sqrt{\log F / N}$. Thus an unchanged total budget cannot generally support arbitrary gate families at only a logarithmic cost in F. Representations sharing a small moment table remain compatible with reuse; that shared structure determines which quantities the release must acquire.

F Experimental protocols and complete summaries

F.1 Information access, metrics, and verification

Each run has an encoder, collector, selector, and evaluator. The encoder takes only the contrast matrix, window labels, a release configuration, and a random seed. The collector returns group counts and sums. The selector receives these aggregate outputs and the known decoder; it never observes individual outcomes. The evaluator separately computes individual-feedback contrasts and population values when they are known. Selection is deterministic with the smallest-index tie rule. No evaluation value, realized regret, or held-out outcome chooses a routing seed or configuration.

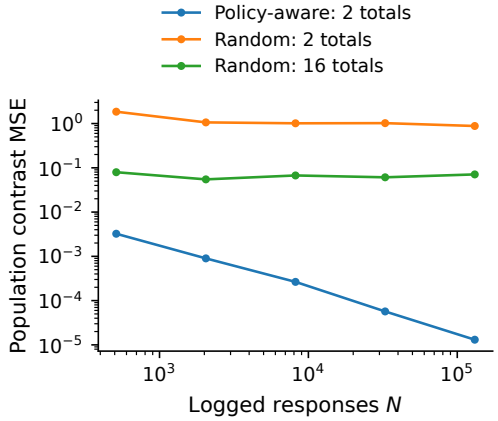
For synthetic data, population policy regret is $\max_j V_j - V_{\widehat{j}}$, with all V_j computed exactly over the finite context population. Reconstruction error is $\max_j |\widetilde{T}_j - T_j|$ relative to the same realized fine-feedback log, including the exactly zero baseline. This separates routing error from statistical sampling error. The empirical selection gap is $\max_j \widehat{V}_j - \widehat{V}_{\widehat{j}}$. A routing coverage check compares each realized contrast error to its common radius plus its policy-specific repair charge; an observed coverage frequency is not a proof of the bound.

The implementation uses NumPy double precision, SciPy, pandas, and scikit-learn; no proprietary optimizer is required. All runs use one BLAS/OpenMP thread. The following independent checks were executed: exhaustive single-bag likelihood enumeration for $h \leq 8$; exact code moments; all binary responses on an eight-row repair example; exact random-partition covariance averages; window and minimum-size validation; independent summary reconstruction; and equality of the collector's bucket output with its specified row sum. These are numerical and interface checks, not a proof-assistant certificate. All formal arguments appear in Appendices A–C.

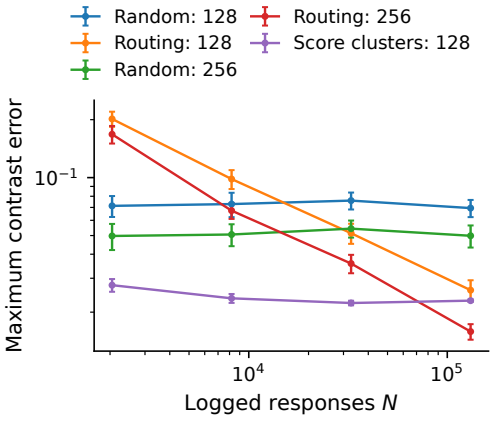
F.2 Earlier finite-library comparisons

Figure 3 retains the fixed-signal two-policy scaling experiment (panel a) and the nonlinear frozen-library reconstruction comparison (panel b). Figure 4 separates population regret (panel a) from the

measured cost of enforcing short reporting windows (panel b). These experiments use the finite-library encoder; Appendix G evaluates post-release head learning.

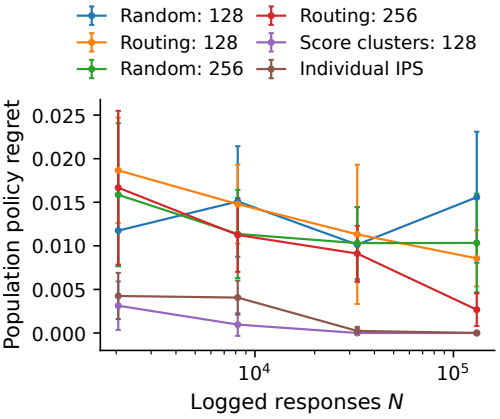


(a) More rows, fixed total budget.

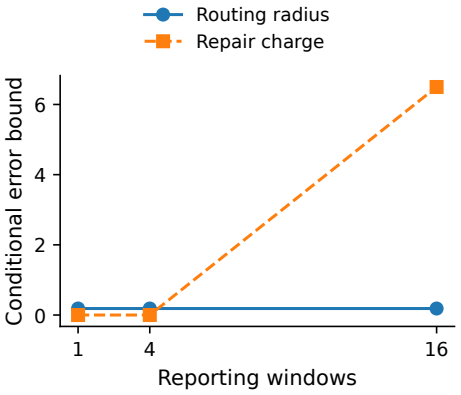


(b) Nonlinear contextual contrasts.

Figure 3: Population contrast error in (a) and fixed-log reconstruction error in (b). The two-policy model in (a) keeps each total budget fixed as N grows; (b) compares 17 policies in the dense nonlinear profile. Bars show 1.96 Monte Carlo standard errors of the displayed means.



(a) Population policy regret.



(b) The cost of short windows.

Figure 4: Population policy regret in (a) and constraint-induced error charges in (b). Panel (a) uses the dense nonlinear profile. At minimum group size 32, panel (b) separates the routing radius from the charge for repairing undersized groups. Short windows can make the repair charge dominate.

F.3 Two-policy fixed-budget experiment

The population is (7) at $\theta = 0.08$. We generate independent $X, A \in \{-1, 1\}$ and Bernoulli rewards. Sample sizes are $512, 2048, 8192, 32768, 131072$, with 100 independent logs at each size. The data seed is $380011 + 3N + s$ for $s \in \{0, \ldots, 99\}$. Policy-aware routing uses $b = 1$, $B = 2$, and center $1/2$; its two extreme-sign groups are deterministic. The minimum group size is four. Random covariance uses two or 16 equal groups, formed by an independent uniform permutation with seed $90000 + s$. The labels are unweighted sums for every aggregate method. Table 3 reports all means. Zero observed regret for policy-aware routing at this fixed signal does not mean zero worst-case regret over θ.

Table 3: Two-policy fixed-signal experiment, 100 independent logs per sample size. Regret uncertainty is one Monte Carlo standard error.

Responses	Method	Contrast MSE	Regret ± SE
512	Policy-aware: 2 totals	0.003241	0.0000 ± 0.0000
512	Random: 16 totals	0.079576	0.0464 ± 0.0073
512	Random: 2 totals	1.854186	0.0752 ± 0.0080
2,048	Policy-aware: 2 totals	0.000899	0.0000 ± 0.0000
2,048	Random: 16 totals	0.054818	0.0416 ± 0.0071
2,048	Random: 2 totals	1.062344	0.0560 ± 0.0077
8,192	Policy-aware: 2 totals	0.000265	0.0000 ± 0.0000
8,192	Random: 16 totals	0.067216	0.0400 ± 0.0070
8,192	Random: 2 totals	1.013845	0.0688 ± 0.0080
32,768	Policy-aware: 2 totals	0.000057	0.0000 ± 0.0000
32,768	Random: 16 totals	0.061041	0.0512 ± 0.0075
32,768	Random: 2 totals	1.021780	0.0864 ± 0.0080
131,072	Policy-aware: 2 totals	0.000013	0.0000 ± 0.0000
131,072	Random: 16 totals	0.070953	0.0400 ± 0.0070
131,072	Random: 2 totals	0.880760	0.0624 ± 0.0078

F.4 Nonlinear finite-context population and frozen library

Contexts are the 64 vectors $x \in \{-1, 1\}^6$, uniformly distributed, and actions are $a \in \{0, 1, 2, 3\}$. Use generator seed 73011. Draw $G \in \mathbb{R}^{6\times 5}$ with independent standard normal entries and define

$$\phi(x) = \big(x, \sin(x^\top G), x_1x_2, x_3x_4, x_5x_6\big) \in \mathbb{R}^{14}.$$

Draw independent $C \in \mathbb{R}^{14\times 4}$ with entries $N(0, 0.45^2)$ and $q \in \mathbb{R}^4$ with entries $N(0, 0.3^2)$. Put $f_a(x) = \phi(x)^\top C_{:a} + q_a$. Dense rewards have mean $0.1 + 0.75\,\mathrm{sigmoid}(f_a(x))$; rare rewards have mean $0.005+0.065\,\mathrm{sigmoid}(f_a(x))$. Weak-overlap rewards use the dense mean. Logging is uniform except in the weak-overlap profile, where probabilities are $(0.7, 0.1, 0.1, 0.1)$.

The baseline policy is the logger. For $k = 0, \ldots, 15$, draw a separate matrix C_k with entries $N(0, 0.45^2)$ and set

$$\pi_{k+1}(a \mid x) = 0.25\mu(a \mid x) + 0.75\,\mathrm{softmax}_a\big((0.25 + 1.1k/15)f(x) + \phi(x)^\top C_k\big).$$

The frozen candidates have different signal alignment and dense nonlinear perturbations. Their construction shares the synthetic signal f with the reward generator; they are not fitted from the realized evaluation labels or claimed to be independent of the synthetic population specification. The learner receives only policy probabilities. No knowledge of f or the conditional reward mean is used in routing or selection. Exact values are computed by summing over all 64 contexts and four actions.

For each profile and $N \in \{2048, 8192, 32768, 131072\}$, use 20 independent data seeds $811000 + s + 17N + 10^6 c$, where $c = 0, 1, 2$ is the profile index and $s = 0, \ldots, 19$. Four consecutive blocks of $N/4$ rows are reporting windows. All aggregate methods require at least four rows in every released group. Routing uses B equal to the maximum contrast over the finite metadata support, block sizes two or four, center zero, $\delta = 0.05$, and seeds $912000 + s + 100b$. The matching random partitions use seed $913000 + s + 100b$, with the same maximum number of totals. If a desired budget exceeds $\lfloor n_w/4 \rfloor$ groups in a window, it is capped at that feasible number; no small-group exception is given to the comparator.

Metadata-clustering comparator. Apply k-means with 32 centers to the 256 possible context/action contrast vectors, using five initializations and seed 42. This uses known finite metadata support, not rewards. Assign observed rows by the resulting cluster and window. Before releasing outcomes, merge nonempty groups below size four into the largest group in that window. The comparator assigns each row its group reward mean and evaluates the corresponding importance-weighted contrasts. It is a strong task-aware baseline, but generally biased. Tables 4 and 5 retain settings where it outperforms unbiased coding.

Additional routing configurations. The dyadic method uses six levels and block size two, so at most $4 \cdot 6 \cdot 8 \cdot 4 = 768$ totals; its row seed is $913371 + s$. Matching random partitions use seed $954001 + s$ and at most 192 legal groups per window. A residual-tree ablation uses four or eight cells per window and block size two, with the same width-based decoder and repair. A coupling ablation uses one uniform variate per row to generate all signs in its selected block instead of independent variates; marginal sign probabilities are unchanged. These configurations are fixed before evaluating each dataset. Table 5 reports every profile at the largest sample size, including settings where other methods outperform routing.

Table 4: Complete main nonlinear sweep. Each condition uses 20 independent logs, shared across methods. Regret is evaluated against the exact finite-population best policy; error is relative to individual IPS on the realized log. SE is one Monte Carlo standard error of regret. Minimum-size repair is applied to all aggregate baselines.

Profile	Responses	Method / budget	Regret	SE	Max. error
Dense	2,048	Individual IPS	0.00425	0.00135	0.00000
Dense	2,048	Random: 128	0.01175	0.00206	0.07132
Dense	2,048	Random: 256	0.01586	0.00419	0.04975
Dense	2,048	Routing: 128	0.01866	0.00309	0.20177
Dense	2,048	Routing: 256	0.01667	0.00450	0.16790
Dense	2,048	Clusters	0.00313	0.00142	0.02762
Dense	8,192	Individual IPS	0.00406	0.00100	0.00000
Dense	8,192	Random: 128	0.01509	0.00323	0.07287
Dense	8,192	Random: 256	0.01136	0.00258	0.05063
Dense	8,192	Routing: 128	0.01478	0.00231	0.09830
Dense	8,192	Routing: 256	0.01125	0.00216	0.06737
Dense	8,192	Clusters	0.00096	0.00066	0.02360
Dense	32,768	Individual IPS	0.00024	0.00024	0.00000
Dense	32,768	Random: 128	0.01014	0.00220	0.07596
Dense	32,768	Random: 256	0.01031	0.00211	0.05432
Dense	32,768	Routing: 128	0.01132	0.00408	0.05142
Dense	32,768	Routing: 256	0.00911	0.00162	0.03578
Dense	32,768	Clusters	0.00000	0.00000	0.02231
Dense	131,072	Individual IPS	0.00000	0.00000	0.00000
Dense	131,072	Random: 128	0.01558	0.00384	0.06945
Dense	131,072	Random: 256	0.01033	0.00289	0.04987
Dense	131,072	Routing: 128	0.00856	0.00164	0.02601
Dense	131,072	Routing: 256	0.00267	0.00096	0.01585
Dense	131,072	Clusters	0.00000	0.00000	0.02295
Rare	2,048	Individual IPS	0.00119	0.00029	0.00000
Rare	2,048	Random: 128	0.00262	0.00048	0.02758
Rare	2,048	Random: 256	0.00245	0.00051	0.02013
Rare	2,048	Routing: 128	0.00209	0.00044	0.05643
Rare	2,048	Routing: 256	0.00797	0.00051	0.04033
Rare	2,048	Clusters	0.00127	0.00028	0.00451
Rare	8,192	Individual IPS	0.00083	0.00019	0.00000
Rare	8,192	Random: 128	0.00333	0.00054	0.02909
Rare	8,192	Random: 256	0.00300	0.00052	0.02116
Rare	8,192	Routing: 128	0.00246	0.00051	0.03150
Rare	8,192	Routing: 256	0.00194	0.00040	0.01853
Rare	8,192	Clusters	0.00083	0.00025	0.00306
Rare	32,768	Individual IPS	0.00039	0.00010	0.00000
Rare	32,768	Random: 128	0.00328	0.00056	0.02835
Rare	32,768	Random: 256	0.00263	0.00047	0.02057
Rare	32,768	Routing: 128	0.00145	0.00043	0.01268
Rare	32,768	Routing: 256	0.00135	0.00034	0.00877
Rare	32,768	Clusters	0.00024	0.00010	0.00205

Continued on next page

Table 4 (continued)

Profile	Responses	Method / budget	Regret	SE	Max. error
Rare	131,072	Individual IPS	0.00008	0.00005	0.00000
Rare	131,072	Random: 128	0.00218	0.00045	0.02752
Rare	131,072	Random: 256	0.00328	0.00052	0.01885
Rare	131,072	Routing: 128	0.00127	0.00028	0.00711
Rare	131,072	Routing: 256	0.00109	0.00019	0.00484
Rare	131,072	Clusters	0.00002	0.00002	0.00210
Weak overlap	2,048	Individual IPS	0.00436	0.00115	0.00000
Weak overlap	2,048	Random: 128	0.01165	0.00276	0.12629
Weak overlap	2,048	Random: 256	0.01316	0.00377	0.07735
Weak overlap	2,048	Routing: 128	0.03711	0.00583	0.58103
Weak overlap	2,048	Routing: 256	0.02043	0.00457	0.50807
Weak overlap	2,048	Clusters	0.01259	0.00452	0.06661
Weak overlap	8,192	Individual IPS	0.00537	0.00154	0.00000
Weak overlap	8,192	Random: 128	0.01783	0.00508	0.11036
Weak overlap	8,192	Random: 256	0.01379	0.00288	0.07726
Weak overlap	8,192	Routing: 128	0.01801	0.00255	0.30924
Weak overlap	8,192	Routing: 256	0.01424	0.00266	0.21964
Weak overlap	8,192	Clusters	0.00644	0.00166	0.03773
Weak overlap	32,768	Individual IPS	0.00095	0.00044	0.00000
Weak overlap	32,768	Random: 128	0.01265	0.00270	0.11302
Weak overlap	32,768	Random: 256	0.01173	0.00346	0.08284
Weak overlap	32,768	Routing: 128	0.02238	0.00451	0.14722
Weak overlap	32,768	Routing: 256	0.01972	0.00430	0.10920
Weak overlap	32,768	Clusters	0.00205	0.00088	0.03447
Weak overlap	131,072	Individual IPS	0.00000	0.00000	0.00000
Weak overlap	131,072	Random: 128	0.01936	0.00287	0.11318
Weak overlap	131,072	Random: 256	0.01489	0.00394	0.07952
Weak overlap	131,072	Routing: 128	0.00975	0.00172	0.07034
Weak overlap	131,072	Routing: 256	0.00888	0.00158	0.05372
Weak overlap	131,072	Clusters	0.00000	0.00000	0.03588

Table 5: All additional routing configurations at $N = 131{,}072$, 20 logs per profile. Four/eight tree cells allow 512/1,024 totals; dyadic coding allows 768; coupled block sizes two/four allow 128/256. Each random comparator is matched to the maximum budget and minimum size. Configurations are fixed before their performance is evaluated.

Profile	Method	Regret	SE	Max. error
Dense	Coupled $b = 2$	0.00925	0.00170	0.02576
Dense	Coupled $b = 4$	0.00353	0.00131	0.01558
Dense	Dyadic routing	0.00693	0.00168	0.02324
Dense	Random: 1024	0.00665	0.00167	0.02655
Dense	Random: 512	0.00803	0.00136	0.03515
Dense	Random: 768	0.00556	0.00158	0.02915
Dense	Tree cells: 4	0.00401	0.00126	0.01631
Dense	Tree cells: 8	0.00650	0.00145	0.01708
Rare	Coupled $b = 2$	0.00112	0.00012	0.00680
Rare	Coupled $b = 4$	0.00127	0.00036	0.00469
Rare	Dyadic routing	0.00143	0.00024	0.00655
Rare	Random: 1024	0.00215	0.00044	0.00994
Rare	Random: 512	0.00179	0.00043	0.01357
Rare	Random: 768	0.00173	0.00037	0.01060
Rare	Tree cells: 4	0.00088	0.00014	0.00459
Rare	Tree cells: 8	0.00105	0.00019	0.00442

Continued on next page

Table 5 (continued)

Profile	Method	Regret	SE	Max. error
Weak overlap	Coupled b =2	0.01235	0.00220	0.07344
Weak overlap	Coupled b =4	0.01079	0.00323	0.04225
Weak overlap	Dyadic routing	0.01100	0.00213	0.03957
Weak overlap	Random: 1024	0.00741	0.00223	0.04047
Weak overlap	Random: 512	0.00930	0.00177	0.05806
Weak overlap	Random: 768	0.00922	0.00212	0.04666
Weak overlap	Tree cells: 4	0.00641	0.00116	0.03607
Weak overlap	Tree cells: 8	0.00641	0.00146	0.03555

F.5 Reporting-window and minimum-size stress test

We fix the dense population and $N = 8192$, using 50 independent seeds $922100 + s$. For each log we evaluate all nine combinations of $L \in \{1, 4, 16\}$ equal consecutive windows and $h \in \{1, 8, 32\}$. Routing has block size two, center zero, and seed $12300 + s$, shared across the constraint settings. Thus changing windows alone preserves the underlying row coefficients: without repairs, the reconstructed estimates are identical. More windows increase the number of potential totals and decrease raw group sizes. Table 6 reports moved rows, actual totals, regret, and certificate components. The largest charge is permitted to exceed one; it then gives no nontrivial population guarantee. The empirical policy regret remains in $[0, 1]$.

Table 6: Window and minimum-size stress at $N = 8192$ (50 logs). "Cap" is the maximum allowed number of totals, and "Actual" is the mean released count. Radius and repair are separate conditional error components. Every reported group meets the minimum size.

L	h	Cap	Actual	Moved	Regret	Radius	Repair
1	1	32	32.00	0.00	0.01661	0.185	0.000
1	8	32	32.00	0.00	0.01661	0.185	0.000
1	32	32	32.00	0.00	0.01661	0.185	0.000
4	1	128	128.00	0.00	0.01664	0.185	0.000
4	8	128	128.00	0.00	0.01664	0.185	0.000
4	32	128	128.00	0.00	0.01664	0.185	0.000
16	1	512	512.00	0.00	0.01699	0.185	0.000
16	8	512	506.78	33.52	0.01713	0.185	0.044
16	32	512	16.00	7789.68	0.10589	0.185	6.497

F.6 Real contextual recommendation-log replay

We use the supplied men's-campaign random-policy log from the Open Bandit Dataset (Saito et al., 2021). It contains 452,949 original impressions and 2,321 binary clicks. Position 1 has 150,936 impressions and 812 clicks; we retain all of them. The reported propensity is $1/34$. We sort by timestamp, use November 24–27, 2019 for aggregate fitting, and November 28–30 for evaluation. Fitting has 86,722 impressions, 459 clicks, and four day windows. Evaluation has 64,214 impressions, 353 clicks, and 72 nonempty hourly blocks.

Four public hashed user-feature strings are transformed, independently of outcomes, to numerical coordinates in $[-1, 1)$ by taking the first eight hexadecimal digits of SHA256, dividing the resulting integer by 2^{32}, and applying $2u - 1$. The 34 user–item affinities are past-click summaries documented by the data provider; no current or held-out label is added to them. Let these inputs be $x \in [-1, 1)^4$ and $a \in [0, \infty)^{34}$. With seed 460019, draw biases $q_{kj} \sim N(0, 0.7^2)$ and interactions $G_{k\ell j} \sim N(0, 0.6^2)$ for 16 candidates. Set τ_k to the 16 equally spaced values from 0.7 to 1.6 and

$$q_k(x, a) = q_k + \tanh(x^\top G_k) + (0.2 + 0.05k)\log(1 + a),$$
$$\pi_{k+1}(j \mid x, a) = 0.7/34 + 0.3\,\mathrm{softmax}_j(q_k(x, a)/\tau_k), \qquad k = 0, \ldots, 15.$$

The baseline is uniform. These fixed contextual score functions exercise nonconstant policy comparisons on the original heterogeneous responses. This experiment evaluates selection from that declared library; Appendix G separately evaluates head training after release. All choices, including the 0.3 mixing weight, are fixed before inspecting fit or held-out performance. The population importance-weight bound from this construction is $W = 0.7 + 0.3 \cdot 34 = 10.9$; the visible fitting contrast maximum used for coding is $B = 3.8926943414$.

Each method receives one disjoint release per routing replication. Basic routing uses block sizes two/four, at most 128/256 totals, minimum size eight, and seed $720001 + s$ for $s = 0, \ldots, 39$. Random covariance uses the same budgets with seeds $840001 + s$. Dyadic routing uses six levels, block size two, minimum size eight, and seed $93201 + s$; its matched 768-total covariance comparator uses seed $844001 + s$. Dyadic coding releases a mean 630.55 totals, including all responses; its upper budget is 768. No raw outcome is disclosed twice within a run. Forty repetitions evaluate alternative releases of one stored log. Each repetition is a separate single-release protocol; their variability is conditional on the same responses.

The main conditional metric is the maximum contrast error relative to individual-feedback IPS on the identical fitting responses. The individual-feedback selector is a richer-information benchmark, not an aggregate algorithm. The later-date paired lift of a selected policy π is

$$\widehat{\Delta}_{\text{late}}(\pi) = \frac{1}{N_{\text{late}}} \sum_{i \in \text{late}} \left(\frac{\pi(A_i \mid X_i)}{1/34} - 1 \right) Y_i.$$

For a method, average across its 40 selected policies before resampling 72 evaluation-hour blocks with replacement 2,000 times, using seed 315791. Each resample carries both the policy and baseline terms from the same impressions. Percentile endpoints at 2.5% and 97.5% describe held-out hour variation conditional on the fitted choices. They are not joint confidence intervals for selection, repeated-user dependence, calendar shifts, or counterfactual potential outcomes. No result treats a noisy held-out lift as population ground truth for the routing bound.

The conditional coding guarantee applies to this fixed log, including its observed dependence. Population transfer is assessed separately by the later-date estimates in Table 7. The retrospective grouping evaluates the proposed protocol on real responses; it does not document deployment of this interface by the original platform.

Table 7: Real-log replay. Errors and lift are per 1,000 impressions. Error uncertainty is one Monte Carlo standard error over 40 alternative releases of the same fitting log; lift intervals resample held-out hours. Budgets count unweighted totals. Dyadic routing uses at most 768, with mean 630.55 actually released. Individual IPS has richer labels.

Method	Budget	Max. error ± SE	Later lift [interval]
Random covariance	128	4.349 ± 0.232	-0.066 [-0.119, -0.014]
Routing ($b = 2$)	128	5.889 ± 0.221	+0.002 [-0.050, +0.056]
Random covariance	256	3.267 ± 0.212	-0.104 [-0.172, -0.038]
Routing ($b = 4$)	256	4.005 ± 0.194	-0.033 [-0.093, +0.022]
Random covariance	768	1.709 ± 0.100	-0.090 [-0.161, -0.022]
Dyadic routing	768	1.914 ± 0.092	+0.019 [-0.030, +0.066]
Individual IPS	86,722	0.000 ± 0.000	-0.039 [-0.227, +0.183]

G Library growth and shared-comparison evaluations

The growing-library experiment uses 240 independent logs. The head-learning experiment uses 240 further independent logs, with different gate counts evaluated on shared logs: 400 metadata designs and 1,600 method outputs. Growing-library comparisons contribute 1,200 method outputs. The real-data experiment uses one original log, 120 shared releases and 120 matched random releases, each reused for three head objectives. With three individual-feedback designs, this gives 729 evaluated heads derived from one real log. Additional fixed-library simulations appear in Appendix F.

G.1 GROWING-LIBRARY TESTING FAMILY

The metadata statistic is $Z \in \{-1, 1\}^p$, with independent uniform coordinates, $p \in \{8, 32, 128\}$. This is the statistic AX in the population used in Theorem 4.2. The baseline score is zero, and contrast j has coefficient Z_j. For each $N \in \{4096, 16384, 65536, 262144\}$, set

$$\theta = \min\{0.20, \tfrac{1}{2}\sqrt{p/N}\}, \qquad Y \mid Z \sim \text{Bernoulli}(1/2 + \theta Z_{j^\star}).$$

The cap is inactive at all reported settings. Repetition $s \in \{0, \ldots, 19\}$ uses $j^\star = 1 + (s \bmod p)$ and data seed $490003 + 13N + 1009p + s$. The $j^\star$ value is fixed before drawing that repetition's log. The policy-value contrasts are exactly θ for $j^\star$ and zero for all other candidates. A selection error, including choosing the baseline, loses θ.

All rows belong to one window. Block sizes are $b = 1$ and $b = \lfloor \log_2 p \rfloor$, with deterministic $B = 1$, center $1/2$, no minimum-size repair, and routing seed $550003 + s + 100b$. Random covariance uses an independent uniform fixed-size partition with seed $660003 + s + 100b$ and the same $J2^b$ total cap. For $p = 8, 32, 128$, the one-coordinate budgets are 16,64,256 and logarithmic-block budgets are 24,224,2432. These logarithmic-block budgets equal $M_{\log}$. This sweep evaluates the linear-signal family, not the separate nuisance family used in Theorem 4.3. Corollary 4.4 guarantees a matched rate when $N \geq pM_{\log}/\log p$; other points evaluate the finite-sample construction without asserting minimax matching.

The score maximizer selects a policy after release. We report the actual population regret, the error to individual-feedback empirical contrasts, and the error to the known population contrasts. Figure 1a uses the last quantity, averaged after taking the maximum across coordinates in each dataset. It does not average at each coordinate before taking that maximum. Figure 1b instead shows population regret for the head experiment below. Table 8 gives all growing-library values and Monte Carlo errors.

Table 8: Growing-library experiment. Each condition averages 20 independent logs. Regret is against the best policy in the fixed population for that condition; the signal $\theta = \frac{1}{2}\sqrt{(K-1)/N}$ changes with N. The error column averages $\max_j |\widehat{T}_j - (V_j - V_0)|$. The half-width is 1.96 Monte Carlo standard errors of regret, not a uniform-risk confidence interval.

K	N	Method	Totals	Regret	MC half-width	Value error
9	4,096	IPS	4,096	0.005524	0.004302	0.019666
9	4,096	Random	16	0.018783	0.003548	0.215356
9	4,096	Random	24	0.016573	0.004302	0.173846
9	4,096	Route $b = 1$	16	0.012153	0.004943	0.038659
9	4,096	Route $b = 3$	24	0.008839	0.004868	0.028412
9	16,384	IPS	16,384	0.002210	0.001987	0.010132
9	16,384	Random	16	0.009391	0.001774	0.219735
9	16,384	Random	24	0.011049	0.000000	0.188767
9	16,384	Route $b = 1$	16	0.004972	0.002472	0.019928
9	16,384	Route $b = 3$	24	0.004419	0.002434	0.013901
9	65,536	IPS	65,536	0.001657	0.001138	0.004891
9	65,536	Random	16	0.005248	0.000541	0.232576
9	65,536	Random	24	0.004419	0.000994	0.178402
9	65,536	Route $b = 1$	16	0.003315	0.001217	0.010642
9	65,536	Route $b = 3$	24	0.003038	0.001236	0.007512
9	262,144	IPS	262,144	0.000691	0.000538	0.002520
9	262,144	Random	16	0.002348	0.000443	0.255849
9	262,144	Random	24	0.002486	0.000373	0.198648
9	262,144	Route $b = 1$	16	0.001933	0.000569	0.005370
9	262,144	Route $b = 3$	24	0.001243	0.000618	0.003635
33	4,096	IPS	4,096	0.002210	0.004331	0.025928
33	4,096	Random	224	0.030936	0.009107	0.076227
33	4,096	Random	64	0.039775	0.005962	0.147134

Continued on next page

Table 8 (continued)

K	N	Method	Totals	Regret	MC half-width	Value error
33	4,096	Route $b = 1$	64	0.041984	0.004331	0.108752
33	4,096	Route $b = 5$	224	0.022097	0.009936	0.050165
33	16,384	IPS	16,384	0.002210	0.002981	0.013479
33	16,384	Random	224	0.020992	0.002166	0.083412
33	16,384	Random	64	0.018783	0.003548	0.141981
33	16,384	Route $b = 1$	64	0.019887	0.002981	0.050629
33	16,384	Route $b = 5$	224	0.007734	0.004739	0.026985
33	65,536	IPS	65,536	0.000000	0.000000	0.006048
33	65,536	Random	224	0.010496	0.001083	0.077383
33	65,536	Random	64	0.011049	0.000000	0.152126
33	65,536	Route $b = 1$	64	0.007734	0.002277	0.026392
33	65,536	Route $b = 5$	224	0.005524	0.002484	0.012626
33	262,144	IPS	262,144	0.000276	0.000541	0.003310
33	262,144	Random	224	0.005524	0.000000	0.078965
33	262,144	Random	64	0.005524	0.000000	0.147086
33	262,144	Route $b = 1$	64	0.004419	0.000994	0.012814
33	262,144	Route $b = 5$	224	0.002762	0.001242	0.006466
129	4,096	IPS	4,096	0.000000	0.000000	0.031677
129	4,096	Random	2,432	0.000000	0.000000	0.035925
129	4,096	Random	256	0.053033	0.019471	0.088121
129	4,096	Route $b = 1$	256	0.083969	0.008662	0.245691
129	4,096	Route $b = 7$	2,432	0.079550	0.011923	0.100544
129	16,384	IPS	16,384	0.000000	0.000000	0.015723
129	16,384	Random	2,432	0.002210	0.004331	0.030090
129	16,384	Random	256	0.030936	0.009107	0.087996
129	16,384	Route $b = 1$	256	0.041984	0.004331	0.128867
129	16,384	Route $b = 7$	2,432	0.022097	0.009936	0.051541
129	65,536	IPS	65,536	0.000000	0.000000	0.007629
129	65,536	Random	2,432	0.017678	0.003974	0.029176
129	65,536	Random	256	0.020992	0.002166	0.086529
129	65,536	Route $b = 1$	256	0.020992	0.002166	0.061330
129	65,536	Route $b = 7$	2,432	0.015468	0.004553	0.025135
129	262,144	IPS	262,144	0.000000	0.000000	0.003912
129	262,144	Random	2,432	0.008286	0.002151	0.027574
129	262,144	Random	256	0.011049	0.000000	0.087464
129	262,144	Route $b = 1$	256	0.010496	0.001083	0.030901
129	262,144	Route $b = 7$	2,432	0.006629	0.002434	0.012555

G.2 Nonlinear public gates and policy fitting

We use the 64 contexts, action probabilities, and nonlinear dense/rare reward functions of Appendix F, without its frozen policy library. For gate count r, seed $772003 + r$ draws a 6×3 standard Gaussian matrix G and then a $11 \times r$ standard Gaussian matrix C. The public gate map is

$$f(x) = (x, x_1x_2, x_3x_4, \sin(x^\top G)), \qquad \phi(x) = \mathrm{softmax}\Big(1.2 f(x)^\top C/\sqrt{11}\Big).$$

These parameters are independent of rewards, candidates, and evaluation data. Population moment entries $Q_{la} = 64^{-1} \sum_x \phi_l(x) m(x, a)$ and the best head value $\sum_l \max_a Q_{la}$ are computed exactly on the finite context space. The mean reward need not lie in the span of the gates.

Dense rewards use $r = 4, 8, 16$; rare rewards and weak overlap use $r = 8$. Each design has $N = 2048, 8192, 32768, 131072$ and 20 repetitions. Data seed $890003 + 17N + 10^6 c + s$ uses profile index $c = 0, 1, 2$. For fixed profile, N and s, all gate counts reuse the same contexts, logged actions and rewards. Four consecutive row blocks are windows. Minimum nonempty group size is four, and the logging floor is $1/4$ in the dense and rare profiles, $1/10$ in weak overlap.

Shared routing uses seed $940003 + s + r$ and at most $4(4r+1)$ totals. Random covariance uses seed $990003+s+r$, the same window and size restrictions, and the same maximum total budget. Its vector coefficients are the $4r$ known functions $\phi_l(x)\mathbf{1}\{a = a'\}/\mu(a \mid x)$. A proposed random budget above $\lfloor n_w/h \rfloor$ is capped at that feasible count before partitioning, so the covariance comparator obeys the same minimum size.

Imputation and individual-feedback comparisons. Hard-gate mean imputation forms keys $(w, \arg\max_l \phi_l(x), a)$, merges undersized keys within the window, and substitutes each group's mean reward for its rows. It uses at most the same budget and is generally biased for overlapping gates. Group means and fitted heads are computed from observed feedback; the population oracle is used only for evaluation. Individual IPS uses the finer feedback of all original rewards. Each method then independently maximizes its moment objective over all heads, assigning one action per gate. Population evaluation uses the resulting policy $\phi(x)U$, not the hard routing assignment. Table 9 gives all methods, while Table 10 shows actual repair counts and their pathwise charges.

Uniform reconstruction error. For $E = \widehat{Q} - Q^{\mathrm{IPS}}$, the exact fixed-log class deviation is

$$\sup_{U \in (\Delta_A)^r} |\langle U, E \rangle| = \max \left\{ \left| \sum_l \max_a E_{la} \right|, \left| \sum_l \min_a E_{la} \right| \right\}.$$

This identity evaluates the entire stochastic-head class exactly. The plots' Monte Carlo bars use 1.96 times the sample standard error across 20 independent logs at that setting. They are not simultaneous population-confidence bands. Larger r increases class expressiveness, sampling error and the total budget simultaneously; comparisons across r are reported as this tradeoff rather than as equal-budget dominance.

Table 9: Post-release policy learning over all heads. D: dense rewards, R: rare rewards, W: weak overlap. The population comparator is the best head in the same gate class. Error is the exact uniform deviation from individual-feedback empirical values over that class. Each row averages 20 logs; the MC column is 1.96 standard errors of the regret mean. Different gate counts reuse the same underlying logs.

Case	r	N	Method	Totals	Regret	MC half-width	Error
D	4	2,048	Hard-gate mean	68	0.011187	0.006937	0.020945
D	4	2,048	Individual IPS	2,048	0.007417	0.004099	0.000000
D	4	2,048	Random covariance	68	0.050547	0.011627	0.145317
D	4	2,048	Shared moments	68	0.027600	0.011778	0.056898
D	4	8,192	Hard-gate mean	68	0.003056	0.001306	0.020038
D	4	8,192	Individual IPS	8,192	0.001951	0.001355	0.000000
D	4	8,192	Random covariance	68	0.051576	0.014042	0.168512
D	4	8,192	Shared moments	68	0.004834	0.003635	0.028602
D	4	32,768	Hard-gate mean	68	0.001548	0.000949	0.020507
D	4	32,768	Individual IPS	32,768	0.000299	0.000403	0.000000
D	4	32,768	Random covariance	68	0.036895	0.010267	0.168706
D	4	32,768	Shared moments	68	0.000592	0.000644	0.012708
D	4	131,072	Hard-gate mean	68	0.001548	0.000949	0.020570
D	4	131,072	Individual IPS	131,072	0.000000	0.000000	0.000000
D	4	131,072	Random covariance	68	0.040243	0.011839	0.163118
D	4	131,072	Shared moments	68	0.000819	0.000650	0.006677
D	8	2,048	Hard-gate mean	132	0.013995	0.004412	0.029359
D	8	2,048	Individual IPS	2,048	0.005564	0.003489	0.000000
D	8	2,048	Random covariance	132	0.048833	0.011666	0.136456
D	8	2,048	Shared moments	132	0.028924	0.007164	0.079246
D	8	8,192	Hard-gate mean	132	0.004143	0.001633	0.026684
D	8	8,192	Individual IPS	8,192	0.001582	0.000916	0.000000
D	8	8,192	Random covariance	132	0.039648	0.007581	0.143964
D	8	8,192	Shared moments	132	0.005841	0.002622	0.039159

Continued on next page

Table 9 (continued)

Case	r	N	Method	Totals	Regret	MC half-width	Error
D	8	32,768	Hard-gate mean	132	0.002932	0.000645	0.027405
D	8	32,768	Individual IPS	32,768	0.000532	0.000322	0.000000
D	8	32,768	Random covariance	132	0.033055	0.008394	0.143043
D	8	32,768	Shared moments	132	0.001139	0.000560	0.019566
D	8	131,072	Hard-gate mean	132	0.003442	0.000447	0.026999
D	8	131,072	Individual IPS	131,072	0.000279	0.000133	0.000000
D	8	131,072	Random covariance	132	0.041302	0.009543	0.148466
D	8	131,072	Shared moments	132	0.000801	0.000524	0.009871
D	16	2,048	Hard-gate mean	260	0.024234	0.005265	0.033177
D	16	2,048	Individual IPS	2,048	0.012339	0.003215	0.000000
D	16	2,048	Random covariance	260	0.037034	0.005763	0.096069
D	16	2,048	Shared moments	260	0.044195	0.007723	0.135862
D	16	8,192	Hard-gate mean	260	0.008971	0.001540	0.026803
D	16	8,192	Individual IPS	8,192	0.003415	0.001931	0.000000
D	16	8,192	Random covariance	260	0.033582	0.008240	0.100982
D	16	8,192	Shared moments	260	0.019950	0.004125	0.061911
D	16	32,768	Hard-gate mean	260	0.006493	0.000910	0.026129
D	16	32,768	Individual IPS	32,768	0.000557	0.000400	0.000000
D	16	32,768	Random covariance	260	0.031857	0.008840	0.104993
D	16	32,768	Shared moments	260	0.006367	0.001852	0.029756
D	16	131,072	Hard-gate mean	260	0.006808	0.000794	0.026002
D	16	131,072	Individual IPS	131,072	0.000041	0.000032	0.000000
D	16	131,072	Random covariance	260	0.028168	0.004925	0.097270
D	16	131,072	Shared moments	260	0.001557	0.000427	0.014945
R	8	2,048	Hard-gate mean	132	0.003762	0.000794	0.006076
R	8	2,048	Individual IPS	2,048	0.003156	0.000796	0.000000
R	8	2,048	Random covariance	132	0.005414	0.000990	0.053035
R	8	2,048	Shared moments	132	0.004833	0.000912	0.023209
R	8	8,192	Hard-gate mean	132	0.002052	0.000651	0.003521
R	8	8,192	Individual IPS	8,192	0.001859	0.000809	0.000000
R	8	8,192	Random covariance	132	0.006389	0.000941	0.058837
R	8	8,192	Shared moments	132	0.003696	0.000650	0.011401
R	8	32,768	Hard-gate mean	132	0.000842	0.000276	0.002740
R	8	32,768	Individual IPS	32,768	0.000521	0.000312	0.000000
R	8	32,768	Random covariance	132	0.006093	0.000927	0.051374
R	8	32,768	Shared moments	132	0.002000	0.000432	0.005530
R	8	131,072	Hard-gate mean	132	0.000265	0.000079	0.002391
R	8	131,072	Individual IPS	131,072	0.000091	0.000051	0.000000
R	8	131,072	Random covariance	132	0.006386	0.000978	0.053084
R	8	131,072	Shared moments	132	0.000299	0.000164	0.002814
W	8	2,048	Hard-gate mean	132	0.020613	0.006957	0.036563
W	8	2,048	Individual IPS	2,048	0.012133	0.005405	0.000000
W	8	2,048	Random covariance	132	0.052571	0.012148	0.242177
W	8	2,048	Shared moments	132	0.042831	0.006906	0.164318
W	8	8,192	Hard-gate mean	132	0.005979	0.002577	0.028253
W	8	8,192	Individual IPS	8,192	0.001878	0.001101	0.000000
W	8	8,192	Random covariance	132	0.046442	0.012344	0.213596
W	8	8,192	Shared moments	132	0.020959	0.005856	0.071725
W	8	32,768	Hard-gate mean	132	0.002808	0.000779	0.027514
W	8	32,768	Individual IPS	32,768	0.000422	0.000159	0.000000
W	8	32,768	Random covariance	132	0.056206	0.010023	0.234850
W	8	32,768	Shared moments	132	0.005106	0.001989	0.035514
W	8	131,072	Hard-gate mean	132	0.003254	0.000530	0.027034
W	8	131,072	Individual IPS	131,072	0.000467	0.000339	0.000000
W	8	131,072	Random covariance	132	0.049802	0.009285	0.211980

Continued on next page

Table 9 (continued)

Case	r	N	Method	Totals	Regret	MC half-width	Error
W	8	131,072	Shared moments	132	0.001251	0.000696	0.016183

Table 10: Actual release costs for shared-head simulations. All runs use four windows and minimum nonempty group size four. The final column is the value-perturbation charge $R_{\mathrm{move}}/(N\kappa)$, before multiplying by two for policy regret. Charges are computed from the realized repairs and are retained even when large.

Profile	r	N	Used totals	Moved rows	Charge
dense	4	2,048	64.0	0.0	0.000000
dense	4	8,192	64.0	0.0	0.000000
dense	4	32,768	64.0	0.0	0.000000
dense	4	131,072	64.0	0.0	0.000000
dense	8	2,048	128.0	0.1	0.000293
dense	8	8,192	128.0	0.0	0.000000
dense	8	32,768	128.0	0.0	0.000000
dense	8	131,072	128.0	0.0	0.000000
dense	16	2,048	235.9	50.3	0.098242
dense	16	8,192	256.0	0.0	0.000000
dense	16	32,768	256.0	0.0	0.000000
dense	16	131,072	256.0	0.0	0.000000
rare	8	2,048	128.0	0.0	0.000000
rare	8	8,192	128.0	0.0	0.000000
rare	8	32,768	128.0	0.0	0.000000
rare	8	131,072	128.0	0.0	0.000000
weak overlap	8	2,048	115.2	40.3	0.196777
weak overlap	8	8,192	132.0	0.0	0.000000
weak overlap	8	32,768	132.0	0.0	0.000000
weak overlap	8	131,072	132.0	0.0	0.000000

G.3 Random-bag covariance estimator

For one window containing $n \geq 2$ rows, let a uniform random permutation produce predetermined nonempty group sizes $h_1, \ldots, h_m$. Put

$$\rho = \frac{\sum_g h_g(h_g - 1)}{n(n-1)}, \quad D_g = \sum_{i:g_i=g} d_i, \quad S_g = \sum_{i:g_i=g} y_i.$$

For a fixed center c, the estimator is

$$\widehat{T} = c\bar{d} + \frac{\sum_g D_g(S_g - ch_g) - \rho(\sum_i d_i)(\sum_i (y_i - c))}{n(1-\rho)}.$$

Every pair of distinct rows has common grouping probability ρ; separating diagonal and off-diagonal terms proves unbiasedness for the fixed empirical contrast. The decoder uses only totals and metadata. Across windows, average the estimates with weights n_w/N. This permits unequal group sizes and provides a covariance-debiasing baseline based on LLP (Busa-Fekete et al., 2023; Applebaum et al., 2026).

G.4 Post-release heads on the Open Bandit Dataset

The original record projection, timestamps, actions, click counts, and date split are those of Appendix F. We retain every position-1 record: 150,936 responses and 812 clicks. There are 86,722 fitting responses with 459 clicks and 64,214 later responses with 353 clicks. All propensities are $1/34$. Four fitting dates define the windows; each nonempty group must contain at least eight rows.

We use the four hashed metadata coordinates described in Appendix F; the 34 affinity fields are not used to form these gates. Center and scale each coordinate using the fitting metadata only, with standard deviation floored at 10^{-6}. For each $r \in \{2, 4, 8\}$, seed $317003 + r$ draws $G \in \mathbb{R}^{4\times 3}$ and $C \in \mathbb{R}^{8\times r}$ with standard Gaussian entries. Set

$$f(x) = (x, \sin(x^\top G), x_1 x_2), \qquad \phi(x) = \mathrm{softmax}(f(x)^\top C/\sqrt{8}).$$

The gate map is frozen before any outcome is used. Because centering is fitted to the public metadata of this log, the analysis here is conditional on that metadata. The independent-fixed-gate population guarantee of Theorem 5.1 has a different sampling requirement. Held-out contexts use the same fitting-metadata centering, scale, and weights.

Forty alternative partitions use routing seed $619003 + s + r$ and random-covariance seed $719003 + s + r$, $s = 0, \ldots, 39$. Maximum aggregate budgets are $4(34r + 1) = 276, 548, 1092$. Since logging is uniform, propensity acceptance is one and residual keys are empty. The shared releases therefore use exactly 272,544,1088 nonempty signal totals, with no minimum-size moves in these runs. All responses still contribute. The covariance comparator uses its full allowed budget. Both methods respect the same cap and legal group sizes.

One release, several head objectives. For each moment table and predeclared $\lambda \in \{0, 0.01, 0.03\}$, fit

$$\widehat{U}_\lambda \in \arg\max_U \left\{ \langle U, \widehat{Q} \rangle + \lambda \sum_l \widehat{p}_l H(U_l) \right\}, \qquad \widehat{p}_l = N^{-1} \sum_i \phi_l(x_i).$$

For $\lambda = 0$ use exact gatewise argmax; otherwise use the softmax solution stated in Appendix E. These are different optimization objectives on the same release, not candidates selected by held-out performance. For a fixed nonnegative penalty weight and gate weights summing to one, its worst effect on unregularized empirical optimization is at most $\lambda \log A$. The uniform event controls estimation for every fitted head, but the regret bound for the unpenalized optimizer should not be assigned unchanged to a regularized head.

Individual IPS fits the same heads with original outcomes, once per gate map, and is marked as finer feedback. Each method's fit gap compares its selected head with the *unregularized* individual-IPS optimum. The zero gap of individual IPS at $\lambda = 0$ is definitional; it says nothing about later population performance.

Later-period evaluation. For each fitted policy, compute paired importance-weighted lift over uniform logging on the held-out impressions. For method-level intervals, first average the fitted heads over its 40 partitions, then resample the 72 held-out hourly blocks with replacement 2,000 times using seed 191703. Each resample retains the pairing of policy and logger terms. The intervals condition on the fitted heads and do not cover date shift, repeat-user dependence, or selecting a regularizer from the plot. Display position and its co-display environment remain fixed. Table 11 reports every head objective, including outcomes unfavorable to shared routing.

Table 11: Post-release head fitting on the original OBD position-1 log. Error, empirical gap to the unconstrained individual-IPS head, lift, and intervals are per 1,000 impressions. The same aggregate release supports all three entropy weights. Each aggregate method averages 40 alternative partitions of one fitting log; the interval resamples held-out hours and does not include model-selection or calendar-shift uncertainty. The individual-IPS comparator has finer feedback.

r	Method	λ	Totals	Error	Fit gap	Lift	Hour interval
2	Individual IPS	0.00	86,722	0.000	0.000	+0.928	[-2.117, +4.245]
2	Individual IPS	0.01	86,722	0.000	5.844	+0.211	[-0.012, +0.444]
2	Individual IPS	0.03	86,722	0.000	6.278	+0.070	[-0.000, +0.143]
2	Random covariance	0.00	276	61.363	5.643	+0.191	[-0.650, +1.064]
2	Random covariance	0.01	276	61.363	5.986	+0.076	[-0.394, +0.594]
2	Random covariance	0.03	276	61.363	6.260	+0.037	[-0.105, +0.185]

Continued on next page

Table 11 (continued)

r	Method	λ	Totals	Error	Fit gap	Lift	Hour interval
2	Shared moments	0.00	276	3.098	0.839	+0.756	[-1.684, +3.306]
2	Shared moments	0.01	276	3.098	5.834	+0.212	[-0.015, +0.449]
2	Shared moments	0.03	276	3.098	6.277	+0.070	[-0.001, +0.144]
4	Individual IPS	0.00	86,722	0.000	0.000	+0.011	[-2.569, +2.804]
4	Individual IPS	0.01	86,722	0.000	6.609	+0.203	[-0.028, +0.445]
4	Individual IPS	0.03	86,722	0.000	7.194	+0.069	[-0.003, +0.143]
4	Random covariance	0.00	548	46.519	6.698	+0.767	[+0.081, +1.432]
4	Random covariance	0.01	548	46.519	6.894	+0.327	[-0.005, +0.649]
4	Random covariance	0.03	548	46.519	7.236	+0.102	[+0.002, +0.196]
4	Shared moments	0.00	548	5.480	1.558	+0.460	[-1.199, +2.130]
4	Shared moments	0.01	548	5.480	6.575	+0.213	[-0.031, +0.473]
4	Shared moments	0.03	548	5.480	7.190	+0.071	[-0.003, +0.148]
8	Individual IPS	0.00	86,722	0.000	0.000	+1.262	[-1.872, +4.803]
8	Individual IPS	0.01	86,722	0.000	6.331	+0.224	[-0.008, +0.470]
8	Individual IPS	0.03	86,722	0.000	6.936	+0.076	[+0.003, +0.153]
8	Random covariance	0.00	1,092	35.116	5.955	+0.532	[-0.078, +1.179]
8	Random covariance	0.01	1,092	35.116	6.358	+0.331	[-0.004, +0.651]
8	Random covariance	0.03	1,092	35.116	6.902	+0.097	[-0.015, +0.209]
8	Shared moments	0.00	1,092	9.034	2.763	+0.947	[-0.317, +2.366]
8	Shared moments	0.01	1,092	9.034	6.236	+0.247	[-0.012, +0.526]
8	Shared moments	0.03	1,092	9.034	6.925	+0.078	[+0.002, +0.158]

G.5 Decision loss attributable to the release

The estimates of future value and the conditional quality of a release answer different questions. To measure the latter without changing the data or fitted policies, let Q^{ind} be the empirical moment matrix from the individual fitting outcomes and set

$$F_\lambda(U) = \langle U, Q^{\mathrm{ind}} \rangle + \lambda \sum_l \widehat{p}_l H(U_l), \qquad U^*_\lambda \in \operatorname*{arg\,max}_{U \in (\Delta_A)^r} F_\lambda(U).$$

For each already-fitted aggregate head $\widehat{U}_\lambda$, define

$$D_\lambda = F_\lambda(U^*_\lambda) - F_\lambda(\widehat{U}_\lambda) \geq 0.$$

The reference maximizes the individual-feedback objective at the same λ: rowwise maximization applies at $\lambda = 0$, and the stated softmax formula applies otherwise. Individual outcomes enter only this evaluation and the reconstruction-error calculation; the aggregate head uses its original release.

This diagnostic is directly controlled by the acquisition theory. The metadata-only entropy term is identical in F_λ and the released objective $\widehat{F}_\lambda$. Since $\widehat{U}_\lambda$ maximizes the latter,

$$\begin{aligned} D_\lambda &\leq (F_\lambda - \widehat{F}_\lambda)(U^*_\lambda) + (\widehat{F}_\lambda - F_\lambda)(\widehat{U}_\lambda) \\ &\leq 2 \sup_{U \in (\Delta_A)^r} |\langle U, \widehat{Q} - Q^{\mathrm{ind}} \rangle|. \end{aligned}$$

No condition on later outcomes is used. Linear constraints can be handled in the same way when both optimizers use the same feasible set. By contrast, changing the penalty only for one reference would confound acquisition with objective choice.

Table 12 evaluates all 729 stored heads. The 40 shared and 40 random releases at each gate count are paired by their replication index on the same fitting log. Their mean differences and percentile Monte Carlo intervals use 10,000 bootstrap resamples of those 40 differences, with fixed seed 2026092109. Each interval describes repeated-encoding uncertainty conditional on the observed log, without simultaneous adjustment across configurations; it is not a population or calendar-shift confidence statement. All nine displayed means favor shared acquisition, while individual releases can reverse this ordering. At $\lambda = 0$, shared has lower loss in 37, 39, and 34 of the 40 comparisons for $r = 2, 4, 8$, respectively.

The later-period difference has a separate exact decomposition. Write $\widehat{V}_{\mathrm{fit}}$ and $\widehat{V}_{\mathrm{later}}$ for the paired IPS value functionals on the two samples. For the same reference and returned head,

$$\begin{aligned}&\widehat{V}_{\mathrm{later}}(U_\lambda^*) - \widehat{V}_{\mathrm{later}}(\widehat{U}_\lambda)\\ &\quad= \widehat{V}_{\mathrm{fit}}(U_\lambda^*) - \widehat{V}_{\mathrm{fit}}(\widehat{U}_\lambda) + (\widehat{V}_{\mathrm{later}} - \widehat{V}_{\mathrm{fit}})(U_\lambda^*) - (\widehat{V}_{\mathrm{later}} - \widehat{V}_{\mathrm{fit}})(\widehat{U}_\lambda).\end{aligned}$$

The final two terms combine sampling variability and distribution change, without separating their causal contributions. For eight gates and $\lambda = 0$, the mean shared acquisition loss is 2.763 per 1,000 impressions, whereas its held-out shortfall relative to the individual-feedback head is only about 0.315. Thus neither training-objective loss nor a single positive lift estimate alone describes the full empirical behavior. All original holdout intervals, including those crossing zero, are retained.

Table 12: Loss from aggregate acquisition on the fixed OBD fitting log, relative to the individual-feedback optimum for the *same* entropy weight. Units are objective value per 1,000 impressions. The last column is random minus shared with a 95% Monte Carlo bootstrap interval across 40 paired alternative releases; it does not quantify uncertainty over new users or dates.

Gates	λ	Shared	Random	Difference [MC interval]
2	0.00	0.839	5.643	4.804 [3.863, 5.699]
2	0.01	0.109	18.292	18.183 [16.557, 20.038]
2	0.03	0.034	11.280	11.246 [10.135, 12.454]
4	0.00	1.558	6.698	5.140 [4.538, 5.720]
4	0.01	0.305	14.267	13.962 [13.006, 14.942]
4	0.03	0.093	6.876	6.783 [6.274, 7.334]
8	0.00	2.763	5.955	3.192 [2.551, 3.809]
8	0.01	0.869	9.693	8.824 [8.239, 9.419]
8	0.03	0.243	3.882	3.639 [3.475, 3.798]

H Related work and the information interfaces

A documented collector configuration. AWS Clean Rooms separates a data owner's authorization of queries from a collaborator's ability to analyze its own metadata. Each table carries an owner-defined analysis rule, and the service enforces the participating tables' restrictions (Amazon Web Services, n.d.d). Under the aggregation rule, approved dimension columns support grouping, while aggregate functions take column arguments (Amazon Web Services, n.d.a). An AWS walkthrough demonstrates this separation: a reporting party groups suppliers' sums by a product dimension in its own table (Ramachandran & Masciola, 2023). A metadata-derived assignment column can play that role in a policy-learning application. This last mapping is our proposed use of the documented interface.

The restriction has an explicit rationale. In its custom-rule documentation, AWS recommends the `COLUMNS_ONLY` setting because aggregate expressions can select small subgroups within otherwise qualifying results; enabling `ANY_EXPRESSION` changes that permission (Amazon Web Services, n.d.c). Thus an approved grouped sum and an exact query summing row-specific reward–weight products are different permissions. These controls motivate the acquisition problem without making a privacy guarantee for our encoders.

The formal experiment fixes one complete cohort and one disjoint release with budget M. A joined implementation must preserve that cohort, including zero outcomes, and match each reward to exactly one assignment. These are application requirements: repeated queries or joins that duplicate reward rows would change the experiment. AWS enforces minimum distinct-identity thresholds and suppresses outputs that fail them (Amazon Web Services, n.d.a). Our row-count repair implements such a threshold when rows represent distinct governed identities; repeated identities require identity-aware handling. AWS also documents differencing risks across repeated queries (Amazon Web Services, n.d.b). The service documentation establishes the supported access restriction and its rationale; the Open Bandit Dataset study remains a retrospective replay, separate from a deployment of this application.

H.1 Learning with group-level labels

Learning individual labels or predictors from group statistics has a substantial history (Kück & de Freitas, 2005; Quadrianto et al., 2009; Yu et al., 2014). Proportion matching and loss-level debiasing address different aspects of this problem. Patrini et al. (2014) study learning with almost no instance labels, while Scott & Zhang (2020) develop a mutual-contamination framework for label proportions. Busa-Fekete et al. (2023) give EasyLLP estimators and empirical-risk algorithms from random bags. Optimistic classification-risk guarantees from debiased square-loss and EasyLLP learning rules (Li et al., 2024), nearly optimal sample complexity (Busa-Fekete et al., 2025), and general-loss results (Applebaum et al., 2026) show that unbiased estimation and learning guarantees are established tools in this literature. Recent proportional-value classification changes the learning objective to address the oversmoothing of proportion matching (Ma et al., 2026). Our objective is policy value: the known weights come from candidate policies and logging probabilities, and the totals must support comparisons of their values.

Bag formation is itself studied. Chen et al. (2023) compare curated and random bags, Javanmard et al. (2024) give adaptive aggregation for learning aggregate responses, and Agarwal et al. (2024) optimize aggregation under learning and privacy criteria. LLP-Bench makes the difference between synthetic random bags and feature-defined tabular bags empirically concrete (Brahmbhatt et al., 2024). These approaches establish task-aware bag formation as a learning tool. PriorBoost partitions data into stages, updates its response predictor using earlier aggregate rewards, and curates bags in fresh stages. Its adaptivity comes from previous rewards even when rows are used once. Theorem 4.2 covers any grouping fitted to the current metadata batch; the reward-driven updates in PriorBoost belong to a multistage experiment. The distinction concerns available information, not whether the partition is fitted or randomized.

H.2 Policy learning and release reuse

Importance-weighted policy values follow the sampling principle of Horvitz & Thompson (1952). Counterfactual risk minimization (Swaminathan & Joachims, 2015), empirical welfare maximization (Kitagawa & Tetenov, 2018), and observational policy learning (Athey & Wager, 2021) analyze selection and optimization over policy classes. Multi-action optimization and surrogate reductions extend the algorithmic scope (Zhou et al., 2023; Bennett & Kallus, 2020). Our post-release optimization uses this empirical policy-learning principle. Mean-operator methods already identify statistics supporting many downstream losses (Patrini et al., 2014). Our task is to acquire the policy-value moments through one disjoint unweighted-total release, without assuming reward homogeneity, and to quantify its uniform error and matching binary-action rate. Exact weighted-moment queries would supply the inputs required by this established downstream optimization step.

Propensity thinning is related to replay evaluation (Li et al., 2011). Doubly robust evaluation and adaptive estimators exploit reward models and variance tradeoffs (Dudík et al., 2011; Wang et al., 2017); balancing directly targets finite-sample distribution matching (Kallus, 2018). These methods operate on individual outcomes or residuals. A pre-fitted outcome model could similarly be used as a control variate in a richer release protocol, but our basic release assumes only bounded outcomes and known propensities. Positivity-free policy learning changes the policy target to avoid general overlap requirements (Zhao et al., 2024); our bounds retain explicit coverage for the declared policy class instead.

The fixed gate representation in (10) includes soft context partitions and overlaps with mixture-of-experts constructions (Jordan & Jacobs, 1994). The gates define the policy class while the reward mean remains arbitrary. Theorem 5.1 controls every head of the fixed map on one event, allowing head optimization after release. Selecting a new gate map requires information for additional value moments, as quantified by Corollary E.1. Contextual policy search with aggregate rewards has also been studied through Bayesian optimization (Feng et al., 2020); its surrogate-model assumptions and sequential queries differ from our single disjoint release. The real-log evaluation uses the Open Bandit Dataset (Saito et al., 2021); it is a retrospective test of the interface, not a field deployment.

Why a fixed gate map does not make arbitrary totals sufficient. For a fixed partition, let C be its group-incidence matrix: the release is Cy, and consider a known weighted sum $v^\top y$. It is determined by Cy for every $y \in [0,1]^N$ if and only if v is constant within each group. Sufficiency is immediate

by combining totals with the group constants. For necessity, if $v_i \neq v_j$ in one group, the two binary arrays $y = e_i$ and $y' = e_j$ have the same metadata and release but different weighted sums. This is the usual row-space criterion for recovering a linear functional. Applied to $v_i = \phi_l(x_i)\mathbf{1}\{a_i = a\}/\mu_i$, it locates the missing quantity: fixed gates specify the weights but do not supply their reward-weighted sums. Our encoding changes the partition before outcomes are observed and estimates these targets with explicitly bounded randomization error. The guarantee concerns values in the declared policy class; it does not require sufficiency for the entire response law. With one-hot gates and uniform logging over A actions, take $\kappa = 1/A$: every row is accepted, its gate is deterministic, and the gate–action totals give the empirical moments exactly. In this familiar special case, ordinary stratum–action totals already provide the moments. Soft gates and varying propensities are handled by stochastic encoding, with arbitrary reward variation within gates.

Two meanings of independent pilot data. Upper guarantees may condition on a gate map or valid width learned from data independent of the evaluation sample. This justifies evaluation and head fitting relative to the resulting fixed class. For a lower bound, pilot rewards from the same unknown environment carry information about the testing hypothesis even if sampled independently. A pilot-assisted lower bound must therefore account for those rewards, even when the fitted encoder is subsequently held fixed. The experiments in Theorems 4.1 and 4.2 supply only the N counted outcome-bearing rows; a pilot-assisted or multistage lower bound would require a different information account. Parameter-independent public randomness remains allowed.

H.3 Randomized encoding and communication lower bounds

Unbiased stochastic encoding is standard in communication-efficient estimation and optimization. Distributed mean estimation (Suresh et al., 2017), QSGD (Alistarh et al., 2017), TernGrad (Wen et al., 2017), and vector quantization (Gandikota et al., 2021) analyze related variance–communication tradeoffs. Task-based quantization explicitly optimizes for a downstream statistical task rather than signal reconstruction (Shlezinger et al., 2019). Our sign code applies this encoding principle to known policy coefficients: it selects a group before seeing the reward, and the collector subsequently emits an unweighted total for that group.

Information-theoretic limits for distributed estimation use data processing, geometric constraints, and information contractions (Zhang et al., 2013; Braverman et al., 2016; Han et al., 2018; Barnes et al., 2020). Here no communication limit is imposed on metadata or on the numerical precision of a total. A learner observes all row contexts, actions, and assignments but not which outcome belongs to which row. Lemma D.1 controls a count for every fixed group composition. Lemma D.2 then bounds the comparison energy preserved by the best partition of the full realized metadata. This combination accommodates data-dependent clustering without conditioning away its selection effect. The biased-count and assignment-entropy identities in Appendix D provide sharper refinements for independently assigned keys.

Directional information versus a scalar capacity. The score of a statistic is the conditional expectation of the full-data score, and equality of Fisher information is not in general global sufficiency (Pollard, 2012). We specialize this identity to obtain $4Z^\top P_G Z$ for grouped Bernoulli rewards. The new finite-signal step is the two-sided count bound in Lemma D.1; the trace then supports the uniform partition converse. Appendix D.3 gives equal-trace partitions with different exact distinguishability. The example shows why the relevant alternative populations, as well as the local matrix, matter for decision risk.

Our proofs use standard entropy, testing, and concentration tools (Cover & Thomas, 2006; Tsybakov, 2009; Hoeffding, 1963; Boucheron et al., 2013). Symmetric Bernoulli expansions and their kernel identities use Krawtchouk orthogonality (Diaconis & Griffiths, 2012; 2019). We give the full count calculations because the complete metadata remain observed, and treating only the released count as the statistical experiment would give an incorrect lower bound.

I What batch-adaptive grouping preserves

The converse permits fitting a partition jointly to the evaluation metadata. We compare a fitted tree with two other grouping procedures at the same sampling model, total budget, and decoder, measuring the comparison energy in Lemma D.2 and the resulting decision error.

I.1 Data, grouping algorithms, and a common decoder

Use the fixed family in (30), with $p \in \{8, 32\}$, $N \in \{2048, 8192, 32768\}$, one window, and cap $M = 2p$. Each setting has 32 independent logs. For repetition $s = 0, \ldots, 31$, let $j^\star = 1+(s \bmod p)$ and $\theta = \frac{1}{2}\sqrt{p/N}$ before sampling. Seed $2026092000 + 100000p + 10N + s$ generates independent uniform signs Z_{ij} and independent Bernoulli responses of mean $1/2 + \theta Z_{ij^\star}$. The context/action representation is equivalently $Z = AX$ from Section 4. There are 192 independent datasets and 576 method outputs; the three partitions of a dataset reuse exactly the same responses.

The first baseline is a uniformly permuted, balanced partition into $2p$ groups. The second independently chooses one coordinate per row and groups by the coordinate and its sign, as in the one-coordinate code. Their common random stream uses the data seed plus 10^9, independently of rewards. The third method fits a binary tree to the full metadata matrix. At each step, among every current leaf and every nonconstant sign coordinate, it chooses the split maximizing the increase in $\mathcal{A}(G, Z)$; ties use the smallest node and coordinate indices. Splitting continues to $2p$ leaves. Every grouping routine receives only Z and random coins. No minimum-size threshold is imposed in this diagnostic, consistently with the outcome-blind protocol.

For a group g, let $h_g = |g|$, $s_{gj} = \sum_{i \in g} Z_{ij}$ and S_g be its reward total. All methods use the same Gaussian count-score decoder,

$$\widehat{j} \in \arg\max_j \left\{ \sum_g \frac{s_{gj}}{h_g}(S_g - h_g/2) - \frac{\theta}{2} \sum_g \frac{s_{gj}^2}{h_g} \right\}.$$

This is the Gaussian likelihood score obtained by approximating a count's mean by $h_g/2 + \theta s_{gj}$ and its variance by $h_g(1/4 - \theta^2)$. The positive signal magnitude is supplied equally to all decoders; only the planted coordinate is unknown. This common approximate score keeps the decoder fixed so that the comparison isolates grouping. Population regret in this family is exactly $\theta \mathbf{1}\{\widehat{j} \neq j^\star\}$. The decoder uses group sizes, known metadata sums, and released counts, never individual outcomes.

I.2 Information concentration and decision error

Table 13 reports all settings. Figure 5(a) shows total energy; panel (b) shows the least-covered coordinate. These are one quarter of the local Fisher trace and minimum coordinate diagonal, not its minimum eigenvalue (Appendix D.3). At $p = 32$, $N = 32768$, $M = 64$, the tree's mean $\mathcal{A}/N$ is 6.078, versus 1.061 for coordinate routing and 0.062 for random bags; regret means are 0.008301, 0.012695, and 0.015625.

The tree distributes this gain unevenly: its mean $\min_j E_j/N$ is 0.003575, versus 0.030929 for coordinate routing. Early splits favor some coordinates, leaving others in small leaves. Equal traces can even conceal complete loss of a comparison (Appendix D.3). The worst-population converse and the tree's average-case advantage therefore describe different aspects of the release.

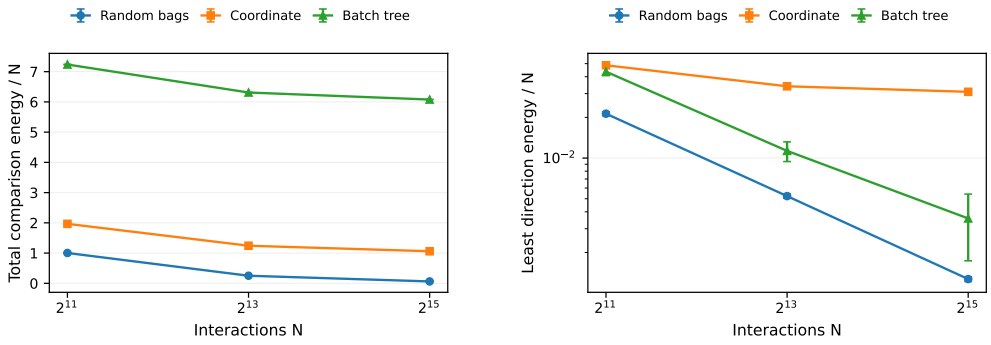
Signals shrink with N, so this is not a fixed-gap learning curve. Classification error can rise while regret falls because an error costs θ. The 1.96 standard errors pool independent logs and the stated cycling of $j^\star$, rather than give confidence intervals for a fixed or worst-case population.

J A direct-count converse at linear total budget

We prove Theorem 4.3 by constructing reward means from a Gaussian latent prior and transferring a testing bound to the original Bernoulli-count experiment. The metadata law, library, and prior are fixed before sampling and encoder randomness; truncation makes every prior draw a legal bounded reward mean.

Table 13: Outcome-blind batch grouping with $M = 2p$ totals. Each row averages 32 independent datasets; the regret half-width is 1.96 Monte Carlo standard errors. The full-data tree, coordinate route, and random bags use identical responses at each repetition and the same count-score decoder.

p	N	Grouping	Regret	$\mathcal{A}/N$	$\min_j E_j/N$
8	2,048	Random bags	0.026367 ± 0.003994	0.061	0.004335
8	2,048	Coordinate route	0.019531 ± 0.005326	1.055	0.120206
8	2,048	Batch-fitted tree	0.010742 ± 0.005225	4.046	0.110478
8	8,192	Random bags	0.013184 ± 0.001997	0.016	0.001081
8	8,192	Coordinate route	0.010254 ± 0.002612	1.014	0.121837
8	8,192	Batch-fitted tree	0.004883 ± 0.002550	4.012	0.107088
8	32,768	Random bags	0.006592 ± 0.000999	0.004	0.000277
8	32,768	Coordinate route	0.005127 ± 0.001306	1.003	0.122722
8	32,768	Batch-fitted tree	0.002197 ± 0.001237	4.003	0.078625
32	2,048	Random bags	0.054688 ± 0.007276	1.003	0.021305
32	2,048	Coordinate route	0.050781 ± 0.008588	1.967	0.048621
32	2,048	Batch-fitted tree	0.023438 ± 0.010652	7.237	0.043587
32	8,192	Random bags	0.027344 ± 0.003638	0.251	0.005231
32	8,192	Coordinate route	0.026367 ± 0.003994	1.243	0.033983
32	8,192	Batch-fitted tree	0.014648 ± 0.005490	6.311	0.011303
32	32,768	Random bags	0.015625 ± 0.000000	0.062	0.001272
32	32,768	Coordinate route	0.012695 ± 0.002147	1.061	0.030929
32	32,768	Batch-fitted tree	0.008301 ± 0.002745	6.078	0.003575



(a) Total preserved comparison energy.

(b) Least-covered comparison coordinate.

Figure 5: Total and minimum coordinate comparison energy at $p = 32$, using at most 64 totals. Curves average 32 independent logs per sample size; bars are 1.96 Monte Carlo standard errors. The fitted tree increases total energy while allocating it unevenly across coordinates.

Sections J.1–J.2 turn a property of the fixed metadata law into a spectral restriction for all nonlinear grouping channels. Section J.3 bounds information in the resulting shared-prior count mixtures. Section J.4 compares heterogeneous Bernoulli counts with mean-matched binomial counts, and J.5 applies these bounds to each realized batch partition. The sufficient range $N \geq Cp^{5/2}\log(2p)$ controls truncation, empirical metadata frequencies, and this comparison error. Constants are universal once the geometric parameter below is fixed.

J.1 A SMALLER FIXED METADATA DISTRIBUTION

We give the geometric input in full. General VC approximation bounds (Goldberg & Jerrum, 1995; Vapnik & Chervonenkis, 1971) provide one route to a finite support; the exponentially small mass

estimate below yields a smaller support here. The radius of a polytope is the largest norm of a stated vertex, relative to the origin.

Lemma J.1 (Projected sign vectors and few vertices)**.** *Fix $0 < \eta < 1/4$. For sufficiently large p and $d = \lfloor \eta p \rfloor$, there is a symmetric measure ν, uniform on $D \le C_\eta p \log(2p)$ indexed sign vectors, such that*

$$\mathbb{E}_\nu Z = 0, \qquad \tfrac{1}{2} I \preceq G := \mathbb{E}_\nu Z Z^\top \preceq 2I.$$

For every orthonormal $U \in \mathbb{R}^{p \times d}$ and every polytope P with at most $4p$ vertices,

$$\nu\{U^\top Z \in P\} \ge 7/8 \quad \Longrightarrow \quad \max_{v \in \mathrm{vert}(P)} \|v\| \ge a_\eta p, \tag{40}$$

where $a_\eta > 0$ depends only on η.

An exponential small-mass claim. First, for each fixed $L_0 < \infty$, there are $a, \gamma > 0$ such that every projected polytope with at most $L_0 p$ vertices and radius at most ap satisfies

$$\Pr\{U^\top Z \in P\} < e^{-\gamma p}, \qquad Z \sim \mathrm{Unif}\{-1,1\}^p. \tag{41}$$

Proof of the small-mass claim. Let X be uniform on $[-1,1]^p$ and, conditional on $X = x$, generate independent signs of means x_i. The marginal Z is uniform on the sign cube. Put $A = \{U^\top Z \in P\}$, $q = \mathbb{P}(A)$, and $a(x) = \mathbb{P}(A \mid X = x)$. If $q = 0$, the claim is immediate; assume $q > 0$. The set $B = \{x : a(x) \ge q/2\}$ has uniform-cube probability at least $q/2$, because

$$q = \mathbb{E} a(X) \le \mathbb{P}(B) + (q/2)\mathbb{P}(B^c).$$

For $x \in B$, conditioning the product sign law on A costs relative entropy $\log(1/a(x)) \le \log(2/q)$. Also

$$\log \mathbb{E}[e^{t^\top (Z - x)} \mid X = x] \le \|t\|^2/2.$$

The variational entropy inequality and optimization over t give

$$\|\mathbb{E}[Z \mid X = x, A] - x\| \le \sqrt{2\log(2/q)} =: r.$$

Since P is convex, $\mathbb{E}[U^\top Z \mid X = x, A] \in P$. Thus

$$\mathbb{P}\{U^\top X \in P + r B_d\} \ge q/2. \tag{42}$$

The density of $U^\top X$ is a $(p-d)$-dimensional orthogonal-section volume divided by 2^p. Each section lies in a ball of radius $\sqrt{p}$. For $d \le p/2$, the ball-volume formula therefore bounds this density by C_0^p, with an absolute $C_0 \ge 1$. Consequently

$$\mathrm{vol}_d(P + r B_d)^{1/d} \ge \left(\frac{q}{2 C_0^p} \right)^{1/d}. \tag{43}$$

We next bound the same volume from above. Since $r B_d \subseteq \mathrm{conv}\{\pm r\sqrt{d}\, e_k : k \le d\}$, the Minkowski sum lies in twice the convex hull of the original vertices, those $2d$ points, and zero. A polytope with L nonzero vertices of norms at most R' has volume at most $\binom{L}{d}(R')^d/d!$: triangulate its boundary, cone from zero, and bound each determinant by $(R')^d$. With the factor two and $L \le L_0 p + 2d + 1$, this gives

$$\mathrm{vol}_d(P + r B_d)^{1/d} \le C_{\eta, L_0} \left(a + \frac{r}{\sqrt{p}} \right). \tag{44}$$

Suppose now that $q \ge e^{-\gamma p}$. For sufficiently large p, $\eta p/2 \le d \le \eta p$, so the right side of (43) is at least

$$\tfrac{1}{2} e^{-2\gamma/\eta} C_0^{-2/\eta}.$$

The right side of (44) is at most $C_{\eta, L_0}\{a + \sqrt{2\gamma + 2\log 2/p}\}$. First choose $\gamma > 0$ sufficiently small (and $\gamma \le \eta$), then $a > 0$ sufficiently small, and finally p_0 sufficiently large. These choices make the latter expression strictly smaller than the former uniformly for $p \ge p_0$. This contradiction proves (41). □

Proof of Lemma J.1. Use the preceding small-mass claim with $L_0 = 6$, obtaining a_0, γ, p_0. Choose constants $a_\eta \le a_0/4$ and $\varepsilon \le a_0/(8\sqrt{\eta})$. Cover the Stiefel manifold by a Frobenius-ε net. A volume-packing argument in its ambient ball of radius $\sqrt{d}$ gives at most

$$(1 + 2\sqrt{d}/\varepsilon)^{pd}$$

net points, themselves on the Stiefel manifold. Cover the radius-$a_\eta p$ ball in $\mathbb{R}^d$ by a Euclidean-$\varepsilon\sqrt{p}$ net with centers in that ball, of size at most $(1 + 2a_\eta\sqrt{p}/\varepsilon)^d$. List $4p$ vertices, duplicating vertices if necessary. The joint net over U and these vertices has size at most

$$\exp\{C_\eta p^2 \log(2p)\}. \tag{45}$$

If $U^\top z \in \mathrm{conv}(v_1, \dots, v_{4p})$, nearest net points satisfy

$$U'^\top z \in P' + 2\varepsilon\sqrt{p}\, B_d, \qquad P' = \mathrm{conv}(v'_1, \dots, v'_{4p}),$$

because $\|z\| = \sqrt{p}$ and operator norm is bounded by Frobenius norm. The enlarged set is contained in a polytope with at most $4p + 2d \le 6p$ vertices and radius at most

$$\max\{2a_\eta p,\ 4\varepsilon\sqrt{pd}\} \le a_0 p.$$

The preceding small-mass claim bounds each fixed net event's cube probability by $e^{-\gamma p}$.

Draw $T = \lceil C'_\eta p \log(2p) \rceil$ independent uniform sign vectors. For any one net event, the probability that it contains at least $3T/4$ sampled vectors is at most

$$2^T e^{-3\gamma p T/4}.$$

Union-bound using (45). For C'_η and p_0 large enough, the probability that any such event contains at least $3T/4$ samples is less than $1/4$. Thus, with the complementary probability, every prohibited original polytope has empirical mass below $3/4$.

The sample covariance is simultaneously between $I/2$ and $2I$ with probability greater than $3/4$. Here is one standard elementary justification. For fixed unit u, $u^\top Z$ is sub-Gaussian with variance proxy one, so $((u^\top Z)^2 - 1)$ has moments bounded by $k! C^k$, and Bernstein gives

$$\mathbb{P}\left(\left|T^{-1}\sum_{i=1}^{T}(u^\top Z_i)^2 - 1\right| > 1/4\right) \le 2e^{-cT}.$$

A $1/4$ net of the unit sphere has at most 9^p points; the usual quadratic-form net inequality then gives operator error at most $1/2$ except with probability $2 \cdot 9^p e^{-cT} < 1/4$. The moment bound follows directly by integrating $\mathbb{P}(|u^\top Z| > t) \le 2e^{-t^2/2}$. Hence both events have positive joint probability.

Fix a successful sample and add its negatives. The resulting uniform indexed measure has $D = 2T$, zero mean, and the same covariance. For every prohibited event A, both A and its reflected event belong to the class controlled above, so the symmetrized mass is below $3/4$, hence below $7/8$. Repeated vectors are distinct public types, which gives the encoder additional metadata rather than reducing its permitted observations. □

J.2 A SPECTRAL CONSTRAINT VALID FOR EVERY NONLINEAR CHANNEL

Fix η for the moment and the measure from Lemma J.1. Complete the constant and the coordinates of Z to an orthogonal decomposition of $L^2(\nu)$. Write $\psi_1, \dots, \psi_s$, $s = D - p - 1$, for an orthonormal basis of the subspace orthogonal to both. Thus $\mathbb{E}\psi = 0$ and $\mathbb{E}Z\psi^\top = 0$. For any probabilities $q_g(t)$ on these types, with at most $M \le 2p$ outputs, put

$$P_g = \mathbb{E}_\nu q_g, \qquad B_g = \frac{\mathbb{E}_\nu[q_g Z^\top]}{\sqrt{P_g}}, \qquad H_g = \frac{\mathbb{E}_\nu[q_g \psi^\top]}{\sqrt{P_g}}.$$

Zero-mass rows are omitted. The probabilities may form any nonlinear channel; Section J.5 constructs such probabilities from a realized batch partition.

Lemma J.2 (Most contrast directions have bounded precision)**.** *Let ρ be any probability measure on the same indexed types with $1/2 \le \rho(t)/\nu(t) \le 2$. Define P_g, B_g, H_g using ρ in place of ν, but retain G and ψ from ν. There is a constant A_η such that*

$$J = 4G^{-1}B^\top (I + 4pHH^\top)^{-1} BG^{-1} \tag{J.2}$$

has fewer than $d = \lfloor \eta p \rfloor$ eigenvalues exceeding A_η/p, for every channel with at most $2p$ outputs.

Proof. Suppose an orthonormal U spans d eigenvectors with eigenvalues collected in $\Lambda \succeq \lambda I$. Put $R = I + 4pHH^\top$ and

$$C = 4R^{-1}BG^{-1}U\Lambda^{-1}.$$

Then $B^\top C = GU$ and $C^\top RC = 4\Lambda^{-1}$. For the vector-valued outputs $a_g = C_g/\sqrt{P_g}$, write $r(t) = \sum_g q_g(t) a_g$, $\ell(t) = \rho(t)/\nu(t)$, and $\bar{a} = \mathbb{E}_\rho r$. Expand the weighted conditional mean in the complete $L^2(\nu)$ basis:

$$\ell(t) r(t) = \bar{a} + U^\top Z_t + e(t), \qquad \mathbb{E}_\nu \|e\|^2 = \|H^\top C\|_F^2 \le \frac{d}{p\lambda} =: h^2.$$

Indeed $\mathbb{E}_\nu[Z(\ell r)^\top] = B^\top C = GU$, and e is its component in the ψ-subspace. Also $T := \sum_g P_g \|a_g\|^2 = \|C\|_F^2 \le 4d/\lambda$ and $\|\bar{a}\| \le \sqrt{T}$. Since $\ell \ge 1/2$, $\mathbb{E}_\nu \sum_g q_g(t)\|a_g\|^2 \le 2T$. Markov's inequality therefore gives a set of ν-mass at least $7/8$ on which

$$\sum_g q_g(t)\|a_g\|^2 \le 32T, \qquad \|e(t)\| \le 4h.$$

For nonnegative weights w_g summing to one and satisfying $\sum_g w_g \|a_g\|^2 \le t_0$, their mean lies in

$$2\,\mathrm{conv}\{0, b_1, \dots, b_M\}, \qquad b_g = a_g \min\{1, \sqrt{t_0}/\|a_g\|\}. \tag{J.3}$$

The rescaled coefficients sum to at most two: for the longer vectors, $\|a_g\|/\sqrt{t_0} \le \|a_g\|^2/t_0$. With $t_0 = 32T$, $\ell \le 2$, and the preceding expansion, the successful $U^\top Z_t$ belong to

$$4\,\mathrm{conv}\{0, b_1, \dots, b_M\} - \bar{a} + 4hB_d.$$

Replace the error ball by the convex hull of $\{\pm 4h\sqrt{d}\, e_l\}_{l=1}^d$. For any two convex sets K_1, K_2, $K_1 + K_2 \subseteq \mathrm{conv}(2K_1 \cup 2K_2)$. Thus this set is contained in a polytope with at most $M + 1 + 2d \le 4p$ vertices and radius at most

$$48\sqrt{T} + 8h\sqrt{d} \le 96\sqrt{d/\lambda} + 8d/\sqrt{p\lambda}.$$

At $\lambda = A_\eta/p$, the last expression is at most $104p\sqrt{\eta/A_\eta}$. Taking A_η large enough makes this smaller than $a_\eta p$, contradicting Lemma J.1. The argument treats all channels and all permitted ρ at once; no operator perturbation or vanishing relative-frequency tolerance is needed. □

J.3 Gaussian prior shifts through a discrete observation kernel

The spectral restriction becomes a testing bound by integrating the shared latent prior jointly. A shift of that latent variable and the relative-entropy chain rule control the observation divergence (Cover & Thomas, 2006), preserving the dependence created by the common prior.

Lemma J.3 (A variational bound for a shared Gaussian latent variable)**.** *Let $(K_t)_{t \in \mathbb{R}^m}$ be an observation kernel satisfying $D_{\mathrm{KL}}(K_t \| K_s) \le \kappa \|t - s\|^2$ for all s, t, with $\kappa > 0$. For $V \sim N(0, I_q)$, a fixed $m \times q$ matrix A, and $P_a = \mathbb{E}_V K_{a+AV}$, we have*

$$D_{\mathrm{KL}}(P_a \| P_0) \le \inf_u \left\{ \tfrac{1}{2}\|u\|^2 + \kappa \|a - Au\|^2 \right\} = \kappa a^\top (I + 2\kappa AA^\top)^{-1} a. \tag{46}$$

No product structure is assumed for the marginal laws P_a, P_0.

Proof. For any $u \in \mathbb{R}^q$, represent the source by $U \sim N(u, I_q)$ and the conditional law $K_{AU+a-Au}$, and the reference by $U \sim N(0, I_q)$ and K_{AU}. Their observation marginals are P_a, P_0. The relative entropy of the two joint laws is

$$\tfrac{1}{2}\|u\|^2 + \mathbb{E}_{U \sim N(u, I_q)} D_{\mathrm{KL}}(K_{AU+a-Au} \| K_{AU}) \le \tfrac{1}{2}\|u\|^2 + \kappa \|a - Au\|^2.$$

Discarding U can only decrease divergence. The shift u is a reparameterization of the two existing laws, not a new data-dependent prior. Minimizing this positive quadratic, with $u = (I + 2\kappa A^\top A)^{-1} 2\kappa A^\top a$, proves the identity. □

An exact count kernel. For any positive group sizes $h_1, \dots, h_m$, define

$$\ell_h(t) = \mathrm{clip}_{[1/4,3/4]}\left(\tfrac{1}{2} + \frac{t}{2\sqrt{h}}\right), \qquad K_t = \bigotimes_{g=1}^{m} \mathrm{Bin}(h_g, \ell_{h_g}(t_g)).$$

Bernoulli divergence is at most $(a-b)^2/[b(1-b)]$. Clipping is one-Lipschitz, so

$$D_{\mathrm{KL}}(K_t \| K_s) \le \frac{4}{3}\|t - s\|^2 \tag{47}$$

for every s, t, including singleton groups. Thus Lemma J.3 applies with $\kappa = 4/3$ to a discrete experiment, without a normal approximation.

Fix $\eta = 1/8192$ in Lemma J.2. Enlarge its constant to $A_* = (8/3)A_\eta$, and draw, before any metadata,

$$\begin{aligned} &J_0 \sim \mathrm{Unif}[p], \qquad \sigma^2 = \eta p / A_*, \qquad \delta = 8\sigma\sqrt{\log(32p)}, \\ &\beta = \delta e_{J_0} + \sigma\xi, \qquad \xi \sim N(0, I_p), \qquad v \sim N(0, pI_s). \end{aligned} \tag{48}$$

with independent ξ, v, J_0. Set

$$f(t) = Z_t^\top G^{-1}\beta + \psi(t)^\top v. \tag{49}$$

For every realization, $\mathbb{E}_\nu f = 0$ and $\mathbb{E}_\nu[Zf] = \beta$. The nuisance changes attribution but not these population policy contrasts.

Lemma J.4 (Testing shared-prior binomial mixtures)**.** *For every channel and measure ρ in Lemma J.2, and any positive group sizes, consider the mixture of the count kernel* (47) *at $t = 2BG^{-1}\beta + 2Hv$, with prior* (48)*. Every classifier of J_0 has error at least $3/8$. Under the prior,*

$$\Pr\{\beta_{J_0} \textit{ exceeds zero and all other coordinates by } \delta/2\} \ge 1 - 1/(512p).$$

Proof. Let P_j be the observation law conditional on $J_0 = j$, and let P_0 use the same centered Gaussian prior without the signal shift. Apply (46) with

$$a_j = 2\delta BG^{-1}e_j, \qquad A = \begin{bmatrix} 2\sigma BG^{-1} & 2\sqrt{p}\,H \end{bmatrix}.$$

The matrix identity for a rank update gives

$$D_{\mathrm{KL}}(P_j \| P_0) \le \frac{\delta^2}{2}\, e_j^\top J_*(I + \sigma^2 J_*)^{-1} e_j, \qquad J_* = \frac{32}{3} G^{-1} B^\top \left(I + \frac{32}{3} pHH^\top\right)^{-1} BG^{-1}. \tag{50}$$

Since $J_* \preceq (8/3)J$ for J in (J.2), fewer than d eigenvalues of J_* exceed A_*/p. Bound those eigenvalues of $J_*(I + \sigma^2 J_*)^{-1}$ by $1/\sigma^2$, and the remaining ones by A_*/p. Consequently

$$I(J_0; S) \le \frac{1}{p}\sum_j D_{\mathrm{KL}}(P_j \| P_0) \le \frac{\delta^2}{2p}\left(\frac{d}{\sigma^2} + A_*\right) \le 64\eta \log(32p) \le \tfrac{1}{8}\log p.$$

The last inequality holds for $p \ge 4$. Fano yields error at least $1 - 1/8 - \log 2/\log p \ge 3/8$. Finally,

$$\Pr\{\|\sigma\xi\|_\infty > \delta/4\} \le 2p\exp\{-\delta^2/(32\sigma^2)\} = 1/(512p).$$

On its complement the shifted coordinate is at least $3\delta/4$, whereas every other coordinate is at most $\delta/4$. □

J.4 Homogenizing each count while retaining the shared prior

The actual group contains heterogeneous Bernoulli means. We compare its count with a binomial count having the same average mean, controlling the error through the variation of those means within the group. Binomial approximation for these counts is classical (Ehm, 1991; Roos, 2001); the orthogonal argument below gives the quadratic relative-entropy bound used here.

Lemma J.5 (A quadratic within-group heterogeneity bound)**.** *Let P be the law of $\sum_{i=1}^h Y_i$ for independent $Y_i \sim \mathrm{Bernoulli}(a_i)$, and put*

$$a = h^{-1}\sum_i a_i \in (0,1), \qquad \tau = \frac{\sum_i (a_i - a)^2}{h\, a(1-a)}, \qquad Q = \mathrm{Bin}(h, a).$$

If $e\tau < 1$, then

$$D_{\mathrm{KL}}(P\|Q) \le \chi^2(P\|Q) \le \frac{(e\tau)^2}{1-e\tau}. \tag{51}$$

For $h \ge 2$, also $\chi^2(P\|Q) \ge h\tau^2/[2(h-1)]$. Thus the chi-square order is quadratic in τ. The upper bound is $2e^2\tau^2$ when $e\tau \le 1/2$; for a singleton both divergences vanish.

Proof. Under independent reference Bernoulli variables of mean a, write $U_i = (Y_i - a)/\sqrt{a(1-a)}$ and $t_i = (a_i - a)/\sqrt{a(1-a)}$. For $S = \sum_i Y_i$, the conditional average of $\prod_{i\in A} U_i$ depends only on $|A| = k$; denote it by $\kappa_k(S)$. Independence, centering, and unit variance give

$$\mathbb{E}_Q[\kappa_k(S)\kappa_l(S)] = \frac{\mathbf{1}\{k=l\}}{\binom{h}{k}}.$$

Conditioning the full likelihood $\prod_i(1+t_iU_i)$ on S and expanding it therefore gives

$$1 + \chi^2(P\|Q) = \sum_{k=0}^{h} \frac{e_k(t)^2}{\binom{h}{k}},$$

where e_k is the kth elementary symmetric polynomial. Here $e_0 = 1$ and $e_1 = 0$. For complex z with $|z| = r$,

$$\left|\prod_i (1+t_iz)\right|^2 \le \exp\left(2\,\mathrm{Re}(z)\sum_i t_i + r^2\sum_i t_i^2\right) = \exp(h\tau r^2).$$

Cauchy's coefficient bound with $r^2 = k/(h\tau)$ gives $|e_k(t)|^2 \le (eh\tau/k)^k$; the case $\tau = 0$ is immediate. Since $\binom{h}{k} \ge (h/k)^k$, summing over $k \ge 2$ proves (51). Retaining just $e_2 = -h\tau/2$ gives the stated lower bound. Finally, $D_{\mathrm{KL}} \le \log(1+\chi^2) \le \chi^2$. □

Integrating the shared prior jointly. For any fixed metadata batch, any partition, and a prior supported on $\|f\|_\infty \le R$, let P be the mixture of its actual Bernoulli group counts with means $1/2 + f_i/\sqrt{N}$. Let Q use the same single prior draw but replace each group's count by $\mathrm{Bin}(h_g, 1/2 + \bar{f}_g/\sqrt{N})$. If $N \ge C_0R^2$ and $\sup_t \mathbb{E}f(t)^4 \le L_4$, Lemma J.5 implies

$$D_{\mathrm{KL}}(P\|Q) \le \frac{CML_4}{N^2}, \qquad \mathrm{TV}(P,Q) \le \frac{C\sqrt{ML_4}}{N}. \tag{52}$$

Indeed $a_g \in [1/4, 3/4]$ and

$$\tau_g^2 \le \frac{C}{N^2}\left(\frac{1}{h_g}\sum_{i\in g} f_i^2\right)^2 \le \frac{C}{N^2 h_g}\sum_{i\in g} f_i^4.$$

Conditional on f, KL adds across groups. Integrate this conditional KL in the joint law of (f, S), then discard f by data processing and apply Pinsker once. The group counts need not be independent after this integration. In particular, no restriction on group sizes, balancing, or assignment independence is used.

J.5 Completion in the original batch experiment

Proof of Theorem 4.3. Use the antipodal pairs underlying ν as contexts, retaining their indices, and let the action be an independent uniform sign. If x represents a pair, policy π_j chooses action x_j and π_0 chooses uniformly. Then $Z = Ax$ has law ν, the importance-weight contrast is Z_j, and all importance ratios are at most two.

Draw the entire prior (48) before the metadata and encoder coins. Write

$$T = pD + p\delta^2 \le Cp^2\log(2p).$$

Uniformly in indexed type t and hypothesis j, the Gaussian variable $f(t)$ has squared mean at most $4p\delta^2$ and variance at most CpD. The complete $L^2(\nu)$ basis has squared row norm D, while $\|G^{-1}\| \le 2$. Hence

$$\mathbb{E}[f(t)^4 \mid J_0 = j] \le CT^2. \tag{53}$$

A Gaussian tail and union bound give a deterministic

$$R^2 \le CT\log(2000D) \le Cp^2\log^2(2p), \qquad \Pr(\|f\|_\infty > R \mid J_0 = j) \le 1/1000. \tag{54}$$

Condition each hypothesis's prior on this event's complement, retaining uniform J_0. This is a choice of legal reward environments before sampling, not an event imposed after observing a partition. Fourth moments increase by at most $1000/999$, and changing the prior changes any resulting observation law by at most $1/1000$ in TV.

For $N \ge Cp^{5/2}\log(2p)$, increasing the universal constant ensures $N \ge C_0R^2$, since $\log(2p)/\sqrt{p}$ is bounded. Every prior draw therefore defines a fixed bounded-response population

$$Y \mid t \sim \text{Bernoulli}\left(\tfrac{1}{2} + \frac{f(t)}{\sqrt{N}}\right), \qquad V(\pi_j) - V(\pi_0) = \frac{\beta_j}{\sqrt{N}}. \tag{55}$$

The Gaussian variables generate least-favorable means; the observations themselves are Bernoulli.

Let ρ be the empirical indexed-type distribution. Independently of the reward prior,

$$\Pr\Big\{\max_t |\rho(t)/\nu(t) - 1| > 1/2\Big\} \le 2De^{-N/(12D)} \le 1/1000. \tag{56}$$

On the complementary event, condition on the complete metadata array and encoder coins. Any resulting batch partition defines empirical probabilities

$$q_g(t) = \frac{\#\{i : \text{type}(i) = t,\ G_i = g\}}{\#\{i : \text{type}(i) = t\}}.$$

These are arbitrary nonnegative probabilities summing to one; the robust spectral lemma applies with this ρ and q. For actual group size $h_g = NP_g$, the normalized conditional mean is

$$\frac{2\sum_{i\in g} f_i}{\sqrt{Nh_g}} = 2B_gG^{-1}\beta + 2H_gv.$$

Thus the mean-matched binomial mixture is exactly the clipped count kernel in Lemma J.4; on the truncated prior the clipping is inactive.

Equation (52) and (53) give, uniformly over the realized metadata and its partition,

$$\text{TV}(P_j^{\text{actual}}, P_j^{\text{binomial,truncated}}) \le \frac{C\sqrt{M}\,T}{N} \le \frac{Cp^{5/2}\log(2p)}{N} \le 1/100.$$

Restore the untruncated prior in the valid clipped-binomial kernel at cost at most $1/1000$. Its testing error is at least $3/8$ by Lemma J.4. Averaging over metadata and subtracting the failure probability in (56) leaves classification error at least $1/3$ in the original experiment.

A mixture selector induces a classifier by sampling its chosen policy index; choosing policy zero is an error. Except on prior probability less than $1/100$ after truncation, a wrong index loses at least $\delta/(2\sqrt{N})$. The prior-average regret is consequently at least

$$\frac{\delta}{2\sqrt{N}}\left(\frac{1}{3} - \frac{1}{100}\right) \ge c\sqrt{\frac{p\log p}{N}}.$$

At least one fixed environment in the prior support attains this bound. The metadata law, library and prior are chosen before data and do not depend on the encoder. This proves the claim for all sufficiently large p. For the finite number of smaller $p \ge 4$, Theorem 4.2 gives $c\sqrt{p/N}$; bounded $\log p$ is absorbed into the universal constant. The unchanged $b = 1$ encoder gives the matching upper bound for all bounded-response populations with $W \le 2$. □

J.6 THE SMALL-SAMPLE ENDPOINT AND THE REMAINING TRANSITION

The direct-count proof applies to every group size and retains the joint shared-prior law. Its sample requirements are

$$N \geq C \max\{D \log(2000D),\ T \log(2000D),\ \sqrt{M}\, T\}, \qquad D = O(p \log p), \quad T = O(p^2 \log p).$$

They give the stated $p^{5/2} \log(2p)$ sufficient range. The $\sqrt{M}\, T$ term pays for replacing heterogeneous counts by mean-matched binomial mixtures; it can include information about nuisance-induced count variation unrelated to the best-policy index. Thus this comparison cost is not a proved policy-learning lower bound on the sample threshold itself.

A matched small-sample endpoint. Singleton groups are allowed in the present experiment. Use the first $n_s = \min\{N, 2p-1\}$ rows as singletons and place all remaining rows, if any, into one additional group. Every response is released once, using at most $2p$ totals. Maximizing the IPS value on the predetermined singleton rows gives, by the bounded-weight Hoeffding bound and a finite maximum,

$$\mathfrak{R}^{\mathrm{batch}}_{N,2p} \leq C \min\{1, \sqrt{\log(2p)/n_s}\}.$$

The singleton rows remain i.i.d.; they are not selected using their rewards. Even individual feedback has the finite-library lower bound $c \min\{1, \sqrt{\log p/N}\}$: use the fixed family in the proof of Theorem 4.2, bound full-data KL by $CN\theta^2$, and apply Fano with $\theta = c_0 \min\{1, \sqrt{\log p/N}\}$. For $N \leq p$ this matches the singleton upper bound. For $p < N \leq p^2/\log(2p)$, the denominator in (9) at $M = 2p$ is at most $3p^2$, giving the matching lower bound $c\sqrt{\log p/p}$. Hence

$$N \leq \frac{p^2}{\log(2p)} \quad \Longrightarrow \quad \mathfrak{R}^{\mathrm{batch}}_{N,2p} \asymp \min\left\{1, \sqrt{\frac{\log(2p)}{\min\{N,p\}}}\right\}. \tag{57}$$

This endpoint uses the absence of a minimum-group-size constraint, not an additional collector capability.

Between (57) and Theorem 4.3, the intermediate transition is still unresolved. In particular $N = p^2$ retains a logarithmic gap. The matching statement in Theorem 4.3 is restricted to its sufficient sample range; it does not assert the whole finite-sample transition.