The Decision Value of Imperfect Provenance: A Minimax Characterization for Shadow Evaluation

Anonymous authors
Paper under double-blind review

Abstract

Online systems use shadow evaluations to compare candidate models or agents with the production choice. Same-request comparisons can cancel shared variation, but imperfect pairing leaves their quality uncertain. We study how provenance determines the value of a fixed shadow budget under bounded i.i.d. losses and marginal-preserving independent resampling. Two channels can have the same erased-score law, matching accuracy, and mutual information about request matching, yet yield constant versus square-root minimax regret at full budget. The relevant quantity is the posterior-contamination spectrum $\mathsf{H}(s) = \mathbb{E}[1/(s+Z)]$, where Z is the probability of independent resampling conditional on the observed provenance. This spectrum governs achievable comparison precision and the hard-family information bound for complete adaptive transcripts, giving minimax regret bounds that match up to logarithmic factors. It leads to calibrated inverse-variance weighting and an order-only method that retains the polynomial regret rates without calibration. In controlled experiments at $T = 2^{18}$, full budget, and a near-clean tail exponent of $1/4$, the maximum tested-gap mean regret is 13.06 for calibrated weighting, 66.28 for unweighted pairing, and 27.28 for order-only service.

1 Introduction

Online systems must choose among candidate models while limiting the cost of evaluating alternatives. Model routing and cascades make this tradeoff explicit (Ding et al., 2024; Ong et al., 2025); similar choices arise among agent configurations (Wu et al., 2024; Hong et al., 2024). A *shadow* evaluation scores an alternative alongside the action used in production. Scoring both on the same request can cancel shared request-level variation, making their difference more informative than independently collected scores. Pairwise model evaluation provides a natural use for such comparisons (Chiang et al., 2024), and mirrored execution provides an interface for the additional call (Envoy Project, n.d.).

Existing learning methods exploit both additional observations and their statistical structure. Query-efficient bandit methods study the benefit of extra evaluations (Lu et al., 2024); difference-adaptive prediction (Thune & Seldin, 2018) and covariance-adaptive identification (Saad et al., 2023) exploit informative joint scores. In fixed-action linear heteroscedastic bandits, Zhao et al. (2026) obtain simple-regret bounds with nearly harmonic-mean dependence on noise variances and a nearly matching lower bound. These results establish the value of unequal observation quality. The question here is how that quality is determined by incomplete evidence about request pairing.

A shadow worker may evaluate a fresh request instead of the production request, while the evaluator retains only a partial record of which occurred. Such independent resampling can preserve both actions' marginal score distributions while changing the noise in their difference. Even matching accuracy or mutual information about the resampling event can miss the distinction that matters: how frequently the record identifies a nearly clean comparison. Thus equal evaluation counts and equal aggregate tag quality need not imply equal learning resources. We ask:

What property of provenance determines the decision value of a fixed shadow budget?

Our insight is that provenance should be evaluated through the comparison precision it supplies at each statistical scale. Let $F = 1$ indicate independent resampling, let C be the observed provenance,

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.

and set $Z = \mathbb{P}(F = 1 \mid C)$. At an intrinsic paired-noise scale $s > 0$, define

$$\mathsf{H}(s) = \mathbb{E}\left[\frac{1}{s+Z}\right]. \tag{1}$$

The spectrum accounts jointly for the frequency and residual contamination of useful comparisons. Under the marginal-preserving channel studied here, conditioning on provenance changes comparison quality while retaining the same mean-difference target. This permits weighting or selecting comparisons without changing which action is best. Appendix A.2 gives theorem-level comparisons.

1.1 Overview of Our Results

A minimax characterization. The same spectrum controls achievable comparison precision and the information bound for complete adaptive transcripts, yielding minimax regret bounds that match up to logarithmic factors. Equal erased-score laws, 75% matching accuracy, and equal mutual information can nevertheless give constant versus square-root regret at full budget. For fixed $N, v, \gamma = 1, \nu = 0$, a linear budget, and $\mathbb{P}(Z \le z) \asymp z^{\alpha}$ with $0 < \alpha < 1$, the rate is $T^{\alpha/(1+\alpha)+o(1)}$ without an exactly clean stratum.

Learning from calibrated or ordered provenance. Inverse-variance weighting uses calibrated contamination probabilities; a bank of ordered thresholds retains the polynomial rates using only a clean-to-dirty ordering. The spectrum also characterizes best-action certification resources up to logarithmic factors. Preparation-time evidence and request-lineage audits provide concrete target-preserving channels for these learning rules.

Controlled evidence. Matched erased-score paths isolate provenance use in two-action service. At $T = 2^{18}$, full budget, and $\alpha = 1/4$, maximum mean regret over twelve tested gaps is 13.06 for calibrated weighting, 66.28 for unweighted pairing, 27.28 for order-only service, and 203.86 for production-only learning. Multi-action experiments distinguish certification and shadow savings from service regret, and identify post-budget continuation as a separate determinant of performance.

Related work. We give a more detailed discussion of related work in A.2.

2 Settings and Preliminaries

We study one production decision per i.i.d. request, with a budget for evaluating an alternative. A paired Bernstein condition describes when same-request differences become easier to estimate as actions approach optimality.

Let $N \ge 2$ and $T \ge 1$ be integers, and write $[N] = \{1, \dots, N\}$. We use the loss convention, so lower values are better. On request t, the bounded loss vector $Y_t = (Y_{1,t}, \dots, Y_{N,t}) \in [0,1]^N$ is drawn independently from a fixed law P. Let $Y = (Y_1, \dots, Y_N) \sim P$ denote a generic draw. Write P_a for the marginal law of Y_a and P_{ab} for the joint marginal law of (Y_a, Y_b). Let $\mu_a = \mathbb{E}Y_a$, $\mu_* = \min_a \mu_a$, $\Delta_a = \mu_a - \mu_*$, and $\mathcal{O}(P) = \arg\min_a \mu_a$. Define

$$d_{ab} = \mu_a - \mu_b, \qquad q_{ab} = \mathbb{E}(Y_a - Y_b)^2 = \mathrm{Var}(Y_a - Y_b) + d_{ab}^2. \tag{2}$$

Let $\mathcal{H}_{t-1}$ denote the completed observable history before request t, including past production and shadow choices, returned scores, and any provenance metadata. A policy $\mathcal{A}$, using $\mathcal{H}_{t-1}$ and internal randomization, chooses a production action $A_t \in [N]$ and an optional distinct shadow action $J_t \in ([N] \setminus \{A_t\}) \cup \{\varnothing\}$ before the current scores are observed. Write $\widetilde{Y}_{a,t}$ for a returned score of action a. Under same-request feedback, it equals $Y_{a,t}$; Section 4 also allows independent-request resampling of the shadow score with the same marginal law. The production score is observed on every round.

An environment θ specifies the base score law and observation channel. Let $\mathcal{F}$ denote a class of such environments. Define the number of shadow evaluations by

$$M_T = \sum_{t=1}^{T} \mathbf{1}\{J_t \neq \varnothing\}.$$

For budget $0 \le B \le T$, let

$$\mathfrak{A}_B(\mathcal{F}) = \left\{ \mathcal{A} : \sup_{\theta \in \mathcal{F}} \mathbb{E}_\theta M_T \le B \right\}. \tag{3}$$

The realized pseudo-regret and its expectation are

$$\mathcal{R}_T = \sum_{t=1}^{T} \Delta_{A_t}, \qquad R_T(\mathcal{A}; \theta) = \mathbb{E}_\theta \mathcal{R}_T.$$

We write $\overline{B} = B \vee 1$ and define the minimax regret

$$\mathfrak{R}(T, B, N; \mathcal{F}) = \inf_{\mathcal{A} \in \mathfrak{A}_B(\mathcal{F})} \sup_{\theta \in \mathcal{F}} R_T(\mathcal{A}; \theta). \tag{4}$$

The implemented policies satisfy the stronger pathwise cap $M_T \le \lfloor B \rfloor$. Same-request paired evaluation is available, for example, when both model predictions can be evaluated against the same revealed label or fixed reference. We assume all scores needed for round t arrive before the next decision. The budget B counts only additional shadow evaluations. All regret statements compare with the best fixed action. The following condition controls the second moment of a same-request comparison relative to its optimality gap.

Definition 1. For $\gamma \in [1, 2]$, $v \ge 1$, and $\nu \in [0, 1]$, $P \in \mathcal{V}_\gamma(v, \nu)$ if some $a_* \in \mathcal{O}(P)$ satisfies

$$q_{aa_*} \le v \Delta_a^{2-\gamma} + \nu, \qquad \text{for all } a \in [N]. \tag{5}$$

We use the convention $0^0 = 1$ when $\gamma = 2$.

The second moment includes both variance and squared mean difference. For a binary-loss pair (a, a_*), put $p_+ = \mathbb{P}(Y_a - Y_{a_*} = 1)$ and $p_- = \mathbb{P}(Y_a - Y_{a_*} = -1)$. Then $q_{aa_*} = p_+ + p_-$ and $\Delta_a = p_+ - p_-$, so $\gamma = 1$ permits noisy wins in both directions when $v > 1$. A finite classifier set containing the Bayes classifier under a positive label margin m_0 satisfies $(\gamma, v, \nu) = (1, 1/m_0, 0)$ (Tsybakov, 2004; Massart & Nédélec, 2006); Appendix B.3 gives the derivation. At zero floor and $\gamma < 2$, tied optima must have identical losses almost surely.

For $0 < g \le 1/8$, let $E \sim \text{Bernoulli}(1/4)$ and $W \sim \text{Bernoulli}(2g)$ be independent, and let $\oplus$ denote XOR. The two action orderings of $(E, E \oplus W)$ have gap g and difference second moment $2g$, hence belong to $\mathcal{V}_1(2, 0)$. Both orderings give positive mass to all four binary score pairs. We use this family to instantiate the provenance spectrum while retaining noisy wins in both directions. Appendix B.2 gives its clean-versus-independent comparison.

3 Learning with Same-Request Paired Feedback

Perfectly matched shadow feedback provides the baseline: both scores refer to the production request, so learning can exploit their difference. We first separate the comparison schedule from the production decision; imperfect provenance will change the comparison statistics while retaining this separation.

Following confidence-based action elimination (Even-Dar et al., 2006), paired-confidence elimination (PCE) maintains an active set. Each phase freezes an active anchor, cycles through active targets, and intersects paired-difference intervals with marginal intervals. It eliminates actions only at complete-cycle boundaries, retaining an optimum on the joint confidence event. Either endpoint of a scheduled pair may serve in production: empirical orientation (EO) chooses the endpoint favored by past paired differences, with fair ties. After shadow exhaustion, marginal sampling and elimination continue. The budget-handoff variant instead freezes the active set and runs MOSS initialized from previous marginal scores. Appendix C.4 specifies PCE, and Appendix C.3 separates orientation, anchor refresh, and handoff.

Let R_m be the expected regret of a fixed-rate SODA learner over m paired rounds. Equation (9) and Lemma 1 of Thune & Seldin (2018), combined with Definition 1, give

$$R_m \le \frac{\log N}{\eta} + \eta(N-1)\{v m^{\gamma-1} R_m^{2-\gamma} + \nu m\}, \qquad 0 < \eta \le [2(N-1)]^{-1}. \tag{6}$$

Here η is the learning rate. Running this learner for $m = \lfloor B \rfloor \geq 1$ paired requests and then sampling uniformly from its past production actions gives $R_T = (T/m)R_m$. Optimizing η yields the paired scale

$$T\left[\left(\frac{vN}{B}\right)^{1/\gamma} + \sqrt{\frac{\nu N}{B}}\right]$$

up to logarithmic factors. Production-only MOSS supplies $\min\{T, \sqrt{NT}\}$ up to constants (Audibert & Bubeck, 2009; Lattimore & Szepesvári, 2020). Appendix C.9 proves the continuation identity; Appendix C.2 gives adaptive PCE guarantees and a matching two-resource lower bound. The comparison cost $v\Delta^{-\gamma} + \nu\Delta^{-2}$ follows from covariance-adaptive identification guarantees (Saad et al., 2023; Saad & Blanchard, 2023). This clean-pair benefit is the starting point. We allow coupling quality to vary across calls and ask how much of the benefit is recoverable from their provenance.

4 Learning with Imperfect Provenance

When pairing is imperfect, provenance indicates how much of the clean comparison remains useful. We model this information after endpoint selection, construct calibrated and order-only comparison banks, and characterize their attainable regret.

4.1 Provenance Model and Spectrum

A provenance channel may attach a discrete label, a confidence score, or a richer record to each shadow evaluation. Let $\mathcal{C}$ be a standard Borel tag space, let m_C be a probability law on $\mathcal{C}$, draw $C \sim m_C$, and let $F \in \{0, 1\}$ indicate whether the shadow score is independently resampled: $F = 0$ for a same-request pair and $F = 1$ for independent resampling. Define the posterior contamination

$$\zeta(c) = \mathbb{P}(F = 1 \mid C = c).$$

Fresh channel draws are independent across requests and of the completed history. The channel is fixed across actions and base environments; conditional on F, provenance is independent of score randomness. Both endpoints are chosen before the current provenance and scores arrive. For a base law P, the pair observed under provenance value c therefore has kernel

$$Q^c_{ab} = (1 - \zeta(c))P_{ab} + \zeta(c)(P_a \otimes P_b). \tag{7}$$

For the returned contrast $D_{ab} = \widetilde{Y}_a - \widetilde{Y}_b$, these marginals imply

$$\mathbb{E}[D_{ab} \mid C = c] = d_{ab} \quad \text{for } m_C\text{-almost every } c.$$

Thus conditioning on provenance preserves the production mean-difference target. The base moment q_{ab} continues to refer to a same-request pair. Every raw shadow call consumes budget. Write

$$Z = \zeta(C), \qquad G(z) = \mathbb{P}(Z \leq z).$$

For a known channel, Z retains all decision-relevant information in C; Appendix D.1 proves the equivalence of these observation interfaces. Averaging over provenance erases this information and leaves a pair kernel with contamination $\mathbb{E}Z$. At intrinsic comparison-noise scale $s > 0$, define

$$\mathsf{H}(s) = \mathbb{E}\left[\frac{1}{s + Z}\right] = \sup_{\substack{0 \leq w(C) \leq 1 \\ \mathbb{E}w(C) > 0}} \frac{(\mathbb{E}w(C))^2}{\mathbb{E}[w(C)^2(s + Z)]}. \tag{8}$$

If the clean comparison has second moment at most s, the conditional moment is at most $s + Z$. For this noise envelope, the variational optimum is attained by inverse-variance weights

$$w(C) = \frac{s}{s + Z}.$$

Thus $\mathsf{H}(s)$ is the available precision per raw shadow call relative to the envelope. The lower bound below shows that this scale is necessary in the worst case. For $P \in \mathcal{V}_\gamma(v, \nu)$, define the structural comparison scale

$$S(\varepsilon) = \min\{1, v\varepsilon^{2-\gamma} + \nu\}, \qquad \mathsf{J}(\varepsilon) = \varepsilon^2\mathsf{H}(S(\varepsilon)), \qquad e(u) = \sup\{\varepsilon \in [0, 1] : \mathsf{J}(\varepsilon) \leq u\}. \tag{9}$$

We set $\mathsf{J}(0) = 0$ and $e(u) = 1$ for $u \geq \mathsf{J}(1)$. Here $S(\varepsilon)$ is the structural moment scale, $\mathsf{J}(\varepsilon)$ is the corresponding decision-information scale, and $e(u)$ maps a per-comparison information requirement to a resolvable gap.

4.2 Provenance-Aware Learning

PCE keeps its phase-fixed anchors, complete-cycle elimination, and marginal observations. Provenance changes only the paired confidence bank. All variants know N, T, B and a confidence level $0 < \delta < 1$; none knows the gaps or an optimal action. Set

$$K = \lceil \log_2(2T) \rceil, \qquad L_{\mathrm{sp}} = \log \frac{4096N^2(K+1)^3(T+1)^3}{\delta}, \qquad \Lambda_{\mathrm{sp}} = KL_{\mathrm{sp}}.$$

Given $Z = \zeta(C)$, use the weights

$$w_j(C) = \frac{2^{-j}}{2^{-j} + Z}, \qquad 0 \le j \le K.$$

For each phase-local pair stream (C_i, D_i) and each weight, maintain $W_n = \sum_i w(C_i)$, $U_n = \sum_i w(C_i)D_i$, and $V_n = \sum_i w(C_i)^2 D_i^2$, with sums over $i \le n$. The interval has center U_n/W_n and radius $(3\sqrt{L_{\mathrm{sp}} V_n} + 4L_{\mathrm{sp}})/W_n$ when $W_n > 0$, and is $[-1, 1]$ otherwise. Intersect the clipped intervals with the marginal interval. An empty intersection invokes PCE's fallback. The bank adapts to the unknown comparison scale without G or (γ, v, ν).

Suppose C is real-valued and its contamination is nondecreasing. Use indicator weights for lower-tag prefixes instead. At doubling pair counts, sort by C, check dyadic prefix sizes, include whole tied groups, and include the whole record. Cache the intersection between checks. This growing bank uses tag order; Appendix D.10 gives the equivalent full-prefix construction and its computational cost.

All bank intervals estimate the same mean difference. PCE therefore retains its elimination rule and continues learning from production scores after the shadow budget is exhausted. Binary Tagged-PCE stratifies by the two labels; the fixed-cut experiment uses a finite numerical bank. Section 6 specifies these implementations separately from the growing rank bank.

4.3 Minimax and Adaptive Guarantees

Theorem 1. *For a fixed channel (7), let $\mathfrak{R}_{\mathrm{ch}}$ be the minimax regret over $\mathcal{V}_\gamma(v, \nu)$, with $\gamma \in [1, 2]$, $v \ge 1$, and $\nu \in [0, 1]$. Universal constants $c_1, c_2 > 0$ satisfy*

$$\mathfrak{R}_{\mathrm{ch}} \ge c_1 T \min\{1, \sqrt{N/T}, e(N/B)\}, \tag{10}$$

$$\mathfrak{R}_{\mathrm{ch}} \le c_2 T \min\{1, \sqrt{N/T}, e(N\log(2N)/B)\}. \tag{11}$$

The upper policy knows the full channel, including G, and (γ, v, ν), and satisfies a pathwise shadow cap. The lower bound grants the same knowledge under the expected-budget constraint.

The bounds differ only by $\log(2N)$ inside the spectral inverse. The oracle uses a fixed-rate SODA update with weights $\lambda/(\lambda+Z)$ normalized by their expectation, followed by uniform-past continuation. Appendix D.2 specifies its scale, update, and proof.

The calibrated PCE design above removes channel-distribution and structural-parameter inputs. For a universal $c_3 > 0$, it satisfies

$$R_T \le c_3 \min\left\{T, \sqrt{NTL_{\mathrm{sp}}}, Te(N\Lambda_{\mathrm{sp}}/B)\right\} + \delta T. \tag{12}$$

Its raw-comparison quota at precision ε is $O(L_{\mathrm{sp}}/\mathsf{J}(\varepsilon))$ per active pair. A threshold with contamination at most z retains mass $G(z)$ and has moment envelope $s + z$. Define $\mathsf{A}(s) = \sup_{0 \le z \le 1} G(z)/(s+z)$. Then

$$\mathsf{A}(s) \le \mathsf{H}(s) \le 2\{1 + \lceil \log_2(1/s) \rceil\}\mathsf{A}(s), \qquad 0 < s \le 1. \tag{13}$$

Ordered PCE satisfies (12) with $e(N(K+1)\Lambda_{\mathrm{sp}}/B)$ in the paired term. The extra factor is logarithmic. Appendices D.3–D.4 prove simultaneous coverage, phase completion, and adaptive bounds.

Let $\mathbb{P}_i^{\mathcal{A}}$ denote the full observable transcript in hard-family environment i. For $v \ge 8$ and $0 < \Delta \le 1/64$, the in-class construction with $d = N - 1$ and $s = \min\{1/4, S(\Delta)\}$ satisfies

$$\frac{1}{d}\sum_{j=2}^{N} D_{\mathrm{KL}}(\mathbb{P}_0^{\mathcal{A}} \| \mathbb{P}_j^{\mathcal{A}}) \le \frac{16T\Delta^2}{d} + \frac{512B\Delta^2}{d}\mathsf{H}(s). \tag{14}$$

Singleton observations and complete tagged pairs contribute separately, with counts bounded by T and B. The same spectral precision used by the comparison banks limits every adaptive decision rule on this family. Appendix D.5 proves the lower bound and extends it to $1 \le v < 8$ by rescaling.

4.4 Consequences of the Spectrum

For fixed N, v, $\gamma = 1$, $\nu = 0$, and $B = T$, Theorem 1 gives three regimes. A clean atom $\mathbb{P}(Z = 0) > 0$ yields $\Theta(1)$ minimax regret. If $\mathbb{P}(Z = 0) = 0$ and $\mathbb{E}[1/Z] < \infty$, the minimax regret is $\Theta(\sqrt{T})$. If there is no clean atom but $\mathbb{E}[1/Z] = \infty$, the regret lies strictly between these orders:

$$\omega(1) \qquad \text{and} \qquad o(\sqrt{T}).$$

These known-channel regimes follow from Appendix D.2. A power-law near-clean tail gives an explicit budget-dependent exponent.

Corollary 1. *Fix N, v, $\nu = 0$, and $\gamma \in [1, 2]$. Suppose $G(z) \asymp z^\alpha$ as $z \downarrow 0$ for $0 < \alpha < 1$, and put $p_\alpha = \gamma + (2 - \gamma)\alpha$. If $B = T^{\beta + o(1)}$, $0 \le \beta \le 1$, then*

$$\mathfrak{R}_{\mathrm{ch}} = T^{\rho + o(1)}, \qquad \rho = \min\{1/2, 1 - \beta/p_\alpha\}. \tag{15}$$

With $\delta = (T + 1)^{-2}$, ordered PCE attains the same exponent without knowing α or the numerical calibration. Polynomial improvement over $\sqrt{T}$ occurs when $\beta > p_\alpha/2$.

At $\gamma = 1$ and $B = T$, the exponent becomes $\alpha/(1 + \alpha)$, filling the interval between constant and square-root order without a pure clean stratum. At $\alpha = 1$, the known-channel minimax rate is $\Theta(\sqrt{T/\log T})$ for fixed N, v, $\nu = 0$, and $\gamma < 2$; for $\alpha > 1$, the spectrum remains bounded. Appendix D.6 treats the boundary and slowly varying cases.

To isolate provenance from the score law itself, let $C \sim \mathrm{Unif}(0, 1)$ and define

$$\zeta_\alpha(c) = \begin{cases} \frac{1}{4}(2c)^{1/\alpha}, & 0 < c \le 1/2, \\ 1 - \dfrac{\alpha}{4(\alpha + 1)}, & 1/2 < c < 1. \end{cases} \tag{16}$$

Every member has $\mathbb{E}Z = 1/2$ and

$$G(z) = \tfrac{1}{2}(4z)^\alpha, \qquad 0 < z \le 1/4.$$

These channels share the tag marginal and erased-score law. For $0 < \alpha < 1$, their regret exponents follow (15); $\alpha > 1$ supplies a bounded-spectrum control. This is the continuous family used in Section 6. On the noisy two-action family of Proposition 1, let e_{nat} be the clipped inverse on $[0, 1]$ of

$$g \longmapsto g^2 \mathsf{H}(g).$$

For every fixed channel, the minimax regret is

$$\Theta\Big(T \min\{T^{-1/2}, e_{\mathrm{nat}}(1/B)\}\Big). \tag{17}$$

For $m = \lfloor B \rfloor \ge 1$ paired requests, weighted EO serves the action favored by the past weighted difference, with fair ties, using

$$\frac{\varepsilon_m}{\varepsilon_m + Z}, \qquad \varepsilon_m = e_{\mathrm{nat}}(1/m).$$

After the prefix, it uses uniform-past continuation. Knowing the full channel, a selector chooses between this rule and MOSS before play and attains (17) up to constants, without knowing the gap or ordering. The paired and marginal scales cross at $B_\star = T/\mathsf{H}(T^{-1/2})$. Appendix D.7 gives the proof and continuation rule, and Figure 1 evaluates the corresponding complete-feedback information and finite-family bounds. The equal-accuracy binary separation is a specialization (Corollary 3).

The learner must be able to reuse provenance semantics across observations. For an even tag alphabet of size $M_{\mathrm{tag}} \ge 16T^2$, Appendix D.8 constructs unknown label meanings for which a single learner incurs $\Omega(\sqrt{T})$ worst-case regret, although every separately calibrated channel admits $O(1)$ regret. A clean-to-dirty ordering avoids this failure by supplying reusable semantics across tag values.

For best-action certification, consider the subclass with a unique optimum and every nonzero gap at least $0 < \Delta \le 1/512$. A policy with shadow budget $B \le T_{\mathrm{cert}}$, incorrect-certificate probability at most δ_0, and noncompletion probability by T_{cert} at most δ_0, uniformly over this subclass, must satisfy, for $0 < \delta_0 \le 1/10$,

$$T_{\mathrm{cert}} + B\mathsf{H}(S(\Delta)) \gtrsim \frac{N \log(1/\delta_0)}{\Delta^2}. \tag{18}$$

Calibrated confidence elimination attains this boundary up to logarithmic factors; Appendix D.9 gives the validity and completion guarantees.

5 Designing Target-Preserving Provenance

The provenance interface must preserve the quantity being compared. For each selected pair and completed history, the returned contrast D_{ab} must satisfy, almost surely under the conditional tag law,

$$\mathbb{E}[D_{ab} \mid C = c, \mathcal{H}_{t-1}, \{a, b\}] = \mu_a - \mu_b. \tag{19}$$

Two constructed interfaces meet this requirement.

A shared evaluation context becomes ready after $L_{\mathrm{r}} \sim \mathrm{Exp}(\lambda_{\mathrm{r}})$, and an independent scheduler supplies a wait allowance $U \sim \mathrm{Exp}(\rho_{\mathrm{w}})$. The rates are positive, and both clocks are independent of the request and scoring randomness. After the two endpoints are selected, a fresh target request is drawn at time U. If $L_{\mathrm{r}} \le U$, both endpoints are evaluated on that request; otherwise the shadow is evaluated on an independent target request. The interface exposes the wait allowance and both scores while keeping the readiness state internal. With tag $C = -U$,

$$Z = e^{-\lambda_{\mathrm{r}} U}, \qquad G(z) = z^{\rho_{\mathrm{w}}/\lambda_{\mathrm{r}}}, \qquad \mathbb{P}(Z = 0) = 0.$$

Thus the preparation and waiting rates determine the near-clean tail and its calibration, while the order of wait allowances alone determines a quality ordering. Readiness (gRPC Authors, 2023) and deadline APIs (gRPC Authors, 2025) provide primitives for such a wrapper. Exposing the readiness bit yields a stronger channel with an exactly clean stratum. The expected serial waiting time over the experiment is $\mathbb{E}M_T/\rho_{\mathrm{w}}$, and a hard upper cap on U creates a positive contamination floor. Appendix E.1 derives these tradeoffs and reports a controlled generator check.

An offline mirrored evaluator can instead bind scores to immutable request instances, model versions, and references. Independent routing to the production request or a fresh target request, outcome-blind auditing, complete retention, and a barrier that waits for both scores preserve (19), including under score-dependent arrival order. A three-state readout distinguishes verified same-request, verified different-request, and unaudited pairs, yielding an exactly clean stratum while supporting coarser binary encodings when needed. Envoy separates mirroring and shadow-span sampling (Envoy Project, n.d.); OpenTelemetry distinguishes early from outcome-dependent sampling (OpenTelemetry Authors, 2025). Istio's default mirroring discards responses (Istio Authors, n.d.), so score capture requires an explicit wrapper. Appendix E.2 specifies the request-ID audit and accounts separately for matching, binding, and audit work.

A fixed nonnegative tag-based weight $w(C)$ with $\mathbb{E}w(C) > 0$ preserves the target in (19), while misspecification changes its precision. Let $\widehat{Z}$ be an estimated contamination probability. If $0 < \ell \le u$ and $\ell(s+Z) \le s+\widehat{Z} \le u(s+Z)$, then the effective weighted precision is at least $4\ell u/(\ell+u)^2$ times the oracle spectrum $\mathsf{H}(s)$. A fixed upward calibration shift can instead cap the effective precision; retaining only the quality order avoids this numerical calibration error. Misordering reduces the quality of usable cuts. Target misspecification is different. For this fixed weight function, write $\mathbb{E}[D_{ab} \mid C = c] = d_{ab} + b_{ab}(c)$. Then a weighted average targets

$$d_{ab} + \frac{\mathbb{E}[w(C)b_{ab}(C)]}{\mathbb{E}w(C)}. \tag{20}$$

A nonzero conditional shift $b_{ab}(c)$ therefore changes the estimand even under exact contamination calibration. The selection and logging mechanism must preserve (19) before the spectral precision analysis applies. For a fixed channel and N, v, a positive base floor $\nu > 0$ restores square-root minimax order as T grows, even with full shadow budget. Appendix E.1 gives the transition scale for general γ and the corresponding calibration and ordering sensitivity bounds.

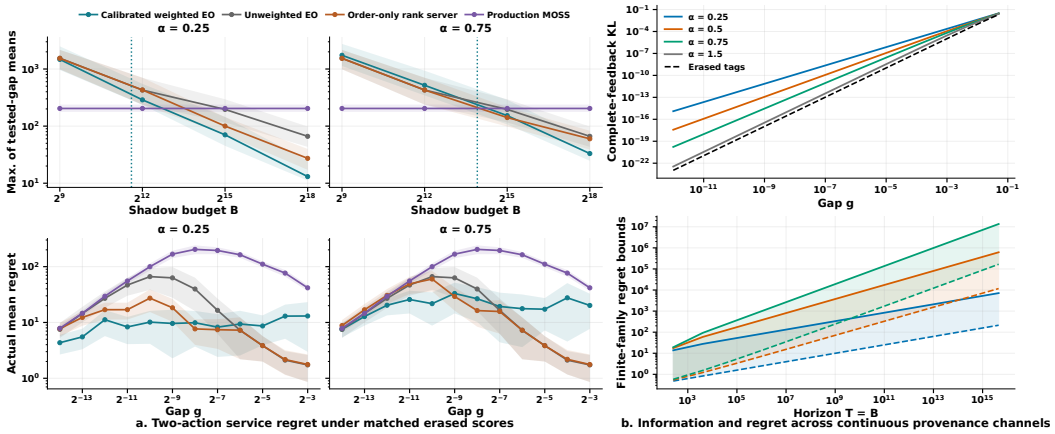

Figure 1: Measured service regret and analytic spectral bounds.

6 Experiments

We test whether provenance improves service at matched score access, whether confidence banks improve certification, and whether the proposed interfaces preserve the target. Experiments were run on a machine with dual NVIDIA RTX 5090 GPUs, Intel Xeon Silver 4410Y CPUs, and 62 GB of memory. PCE uses $\delta = (T+1)^{-2}$. Paired comparisons share erased-score streams; production-only MOSS (Audibert & Bubeck, 2009) observes served scores. We report 95% intervals and use Holm correction (Holm, 1979) within each family; maximum-over-gap inference is reported separately.

6.1 Two-Action Service and Spectrum

Service design. Using the two-action family in Section 2 and channels (16), we cross $T \in \{2^{10}, 2^{12}, \ldots, 2^{18}\}$, $g \in \{2^{-3}, 2^{-4}, \ldots, 2^{-14}\}$, $\alpha \in \{1/4, 1/2, 3/4, 3/2\}$, and $B = 0$ or $\lfloor T^{\beta} \rfloor$ for $\beta \in \{1/2, 2/3, 5/6, 1\}$. Each cell uses 32 seed clusters, each averaging both action orderings. Calibrated weighted empirical orientation (EO), unweighted EO, and the order-only rank server share $m = \lfloor B \rfloor$ paired requests and uniform-past continuation; zero-budget controls use MOSS. The rank bank checks dyadic counts and prefixes, including whole tied groups, with $L_{\mathrm{srv}} = \log[256(m+1)^3(T+1)^2]$. Cell means use t_{31} intervals; paired tests apply correction over 1,200 cells per contrast. Maxima over gaps use 2,000 paired cluster-bootstrap resamples (Efron, 1979). Appendix F.1 specifies sampling and policy controls.

Service results. Figure 1 and Table 2 report the $T = 2^{18}$ results. At full budget, weighted/unweighted/rank/MOSS maximum tested-gap mean regrets are $(13.06, 66.28, 27.28, 203.86)$ for $\alpha = 1/4$ and $(33.08, 66.28, 59.99, 203.86)$ for $\alpha = 3/4$. At $\alpha = 1/4$, the weighted-minus-unweighted difference is -53.21, with nominal 95% interval $[-85.27, -23.19]$ from a 20,000-resample paired bootstrap; simultaneous projected intervals still contain zero (Table 3). These point estimates favor provenance-aware service at larger budgets, whereas MOSS performs better at smaller budgets. The marked $B_\star$ compares theoretical rate scales.

Spectrum design and results. For the same four channels, the right block evaluates tagged and erased-score KL on 241 logarithmically spaced gaps in $[10^{-20}, 1/8]$, with $T = B \in \{2^8, 2^{12}, \ldots, 2^{52}\}$ for regret bounds and 64-point Gauss–Legendre quadrature. The analytic upper policy uses cutoff $z_T = \min\{1/4, T^{-1/(1+\alpha)}\}$; the lower bound uses complete-feedback KL. The displayed $\alpha < 1$ regret bounds grow below the square-root scale, consistent with Corollary 1. Appendix F.2 gives the calculations; the binary-channel controls in Appendix F.6 examine the equal-accuracy separation.

6.2 Certification and Budget Handoff

Binary-tag design. The pooled/tagged comparison isolates the confidence bank at fixed service orientation. We use $T = 2^{24}$ and shared-noise scores $Y_j = E \oplus W_j$, with independent $E \sim \mathrm{Bernoulli}(1/4)$, $W_1 = 0$, and $W_j \sim \mathrm{Bernoulli}(2g)$ for $j \geq 2$, where $g = 1/64$. Mix the joint law equally with its product marginals. The rare-pure, conservative, residual-floor, and aggressive channels use clean-component acceptance and false-clean rates $(\tau, \phi) = (1/4, 0), (1/2, 0), (1/2, 1/50), (1, 1/2)$.

We cross them with $(N, B/T) \in \{(4, 1/4), (4, 3/8), (4, 1), (8, 3/8), (8, 1)\}$ and 64 matched seeds per cell. Pooled PCE, Tagged-PCE, the known-label reference, and PCE-H use $L_{\text{tag}} = \log[256N^2K(T+1)/\delta]$, pooled EO, and active-set MOSS handoff; PCE-H refreshes anchors within phases. Certification requires a singleton before handoff or horizon end; handoff censors incomplete runs. Rates use Wilson intervals (Wilson, 1927) and exact discordant tests.

Binary-tag results. For conservative tags, pooled-to-tagged certificate counts rise from 8 to 35, 31 to 55, and 19 to 54 at $(N, B/T) = (4, 1/4), (4, 3/8), (8, 1)$; all survive 20-cell correction. Four-action full-budget shadow use decreases by 26.6%, while every matched production-regret difference is zero. The measured gain is in certification and shadow use (Table 4; Figure 4).

Ordered-cut design and results. Holding T, g fixed, we use $N = 4$, the four continuous channels, $B/T \in \{3/8, 1\}$, and 32 matched seeds. The order-only bank uses $C \le 2^{-j}$, $j = 0, \ldots, 10$, and $L_{\text{cut}} = L_{\text{tag}} + \log 12$, retaining phase-fixed anchors, pooled EO, and the same MOSS handoff as pooled PCE. At $B/T = 3/8$, certificates rise from $(15, 16, 17)$ to $(32, 29, 28)$ for $\alpha = (1/4, 1/2, 3/4)$, all significant after eight-cell correction; at $\alpha = 3/2$, they change from 15 to 13. All 256 matched production-regret differences are zero. Table 5 reports shadow use; these results evaluate the eleven-cut bank, while the growing rank bank has the general order-only guarantee.

Handoff design. To isolate post-budget learning, we cross $N \in \{4, 8\}$, $T = 2^{24}$, $\alpha \in \{1/4, 3/4\}$, and $B/T \in \{1/4, 1\}$. The shared-noise generator has nonoptimal gaps $g_j = 2^{-8+4(j-1)/(N-2)}$, $j = 1, \ldots, N-1$. Thirty-two clusters balance the optimal index and permute gaps. Pooled, calibrated, and rank PCE use L_{sp}, phase-fixed anchors, phase-local paired banks refreshed at dyadic counts, and marginal continuation. Pooled handoff switches continuation to warm-start MOSS; production-only MOSS is the fifth policy. Intervals use 4,000 optimal-index-stratified bootstrap resamples.

Handoff results. No provenance-versus-pooled regret contrast survives eight-cell correction. At $B = T/4$, marginal continuation gives mean regret 49,472 and 107,521 for four and eight actions; changing the tail to MOSS gives 204.43 and 491.94, also the full-budget PCE means. The largest service effect here comes from continuation, rather than the provenance bank (Figure 7; Table 6).

6.3 Target-Preservation Audit

Design. The request-ID audit in Appendix E.2 uses the shared-noise generator with gaps $(0, 0.03, 0.07, 0.12)$, independent audit and replay probabilities $1/2$, all six action pairs, and two binary audit encodings. Exact enumeration checks 24 conditional gaps; 120,000 simulated jobs use score-dependent arrival order and complete retention. A selection control at $g = 0.05$ accepts with probability 0.9 when $Y_{\text{good}} = 1$ and 0.1 otherwise. The separate two-clock check in Appendix E.1 uses $\lambda_{\text{r}} = 1$, $\rho_{\text{w}} \in \{1/4, 3/4, 3/2\}$, 400,000 jobs per rate, gaps $(0, 0.01, 0.03, 0.06)$, ten quantile bins, and nine quality thresholds per channel.

Results. Enumeration confirms all 24 request-ID identities, while the packet- and clock-audit estimates remain within their respective simultaneous Hoeffding intervals. The maximum empirical discrepancies are 0.01379 for the packet audit and 0.00682 for the clock conditional-gap checks, with every request-ID identity check agreeing with its routing bit. Outcome-dependent selection reverses the gap from 0.05 to -0.05, despite correct request matching.

7 Conclusion

Under marginal-preserving independent resampling, the decision value of provenance is characterized by its posterior-contamination spectrum. The same spectrum governs attainable comparison precision and the hard-family information bound for complete adaptive transcripts, yielding minimax regret bounds that match up to logarithmic factors. It explains constant, intermediate, and square-root regimes, and supports calibrated weighting, order-only learning with the same polynomial rates, and best-action certification. In controlled two-action experiments, provenance-aware service has lower maximum tested-gap mean regret in the reported high-budget near-clean settings. Multi-action experiments separately demonstrate certification and shadow savings, while exposing the importance of post-budget continuation. The constructed clock and request-lineage interfaces preserve the comparison target. These results connect provenance quality to the comparison information available across statistical scales, with target preservation as the condition that makes the connection valid.

AI Use Statement

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

Reproducibility Statement

The observation model and the information supplied to each policy are specified in Sections 2–5. Appendix A.1 maps notation and proof organization; Appendix B collects the model constructions; Appendices C and D provide the proofs, confidence rules, stopping conditions, and budget accounting; and Appendix E gives the provenance-generation and request-lineage details. Section 6 specifies the main experimental grids, confidence settings, replication counts, and hardware. Appendix F supplements these with sampling details, random seeds, numerical calculations, censoring rules, and complete results. The Supplementary Material also includes all code used to run the experiments.

References

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, volume 40 of *Proceedings of Machine Learning Research*, pp. 23–35, 2015.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, pp. 217–226, 2009.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Orly Avner, Shie Mannor, and Ohad Shamir. Decoupling exploration and exploitation in multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pp. 409–416, 2012.

Vinay Praneeth Boda and Prashanth L.A. Correlated bandits or: How to minimize mean-squared error online. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 686–694, 2019.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388, 2024.

Rémy Degenne, Evrard Garcelon, and Vianney Perchet. Bandits with side observations: Bounded vs. logarithmic regret. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 467–476, 2018.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid LLM: Cost-efficient and quality-aware query routing. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Envoy Project. HTTP route components: RequestMirrorPolicy. Envoy Proxy Documentation, n.d. URL https://www.envoyproxy.io/docs/envoy/latest/api-v3/config/route/v3/route_components.proto. Accessed September 25, 2026.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

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

gRPC Authors. Wait-for-Ready. gRPC Documentation, 2023. URL https://grpc.io/docs/guides/wait-for-ready/. Last modified August 22, 2023; accessed September 25, 2026.

gRPC Authors. Deadlines. gRPC Documentation, 2025. URL https://grpc.io/docs/guides/deadlines/. Last modified July 7, 2025; accessed September 25, 2026.

Samarth Gupta, Shreyas Chaudhari, Gauri Joshi, and Osman Yağan. Multi-armed bandits with correlated arms. *IEEE Transactions on Information Theory*, 67(10):6711–6732, 2021.

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

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

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

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

Istio Authors. Mirroring. Istio Documentation, n.d. URL https://istio.io/latest/docs/tasks/traffic-management/mirroring/. Accessed September 25, 2026.

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

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, volume 75 of *Proceedings of Machine Learning Research*, pp. 358–384, 2018.

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

Zhou Lu, Qiuyi Zhang, Xinyi Chen, Fred Zhang, David P. Woodruff, and Elad Hazan. Adaptive regret for bandits made possible: Two queries suffice. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Proceedings of the 25th Conference on Neural Information Processing Systems (NeurIPS)*, volume 24, pp. 684–692, 2011.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34 (5):2326–2366, 2006.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, pp. 73–83, 2009.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

OpenTelemetry Authors. Sampling. OpenTelemetry Documentation, 2025. URL https://opentelemetry.io/docs/concepts/sampling/. Last modified October 16, 2025; accessed September 25, 2026.

Riccardo Poiani, Rémy Degenne, Emilie Kaufmann, Alberto Maria Metelli, and Marcello Restelli. Optimal multi-fidelity best-arm identification. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 121882–121927, 2024.

Chloé Rouyer and Yevgeny Seldin. Tsallis-INF for decoupled exploration and exploitation in multi-armed bandits. In *Proceedings of the 33rd Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3227–3249, 2020.

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

El Mehdi Saad and Gilles Blanchard. Constant regret for sequence prediction with limited advice. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory (ALT)*, volume 201 of *Proceedings of Machine Learning Research*, pp. 1343–1386, 2023.

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

Itai Shufaro, Nadav Merlis, Nir Weinberger, and Shie Mannor. On bits and bandits: Quantifying the regret-information trade-off. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Tobias Sommer Thune and Yevgeny Seldin. Adaptation to easy data in prediction with limited advice. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 2914–2923, 2018.

Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16(54):1793–1861, 2015.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *Proceedings of the 1st Conference on Language Modeling (COLM)*, 2024.

Heyang Zhao, Tianyuan Jin, Weixin Wang, Vincent Y. F. Tan, Pan Xu, and Quanquan Gu. Breaking the total variance barrier: Sharp sample complexity for linear heteroscedastic bandits with fixed action set. In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026.

Appendix

Table of Contents

A Sec. 1: Overview, notation, and related work **15**

A.1 Notation and proof roadmap 15

A.2 Related work and theorem-level comparison 16

B Sec. 2: Settings and preliminaries **16**

B.1 The natural coupled family 17

B.2 A direct proof of the coupling-only separation 17

B.3 A classification realization 18

C Sec. 3: Same-request paired-feedback guarantees **18**

C.1 Probability and change of measure (used in Sections 3–4) 18

C.2 The base two-resource frontier 19

C.3 Service and finite-budget policies 20

C.4 Full algorithm specification 21

C.5 Paired Confidence Bounds 23

C.6 Analysis of All PCE Service Variants 24

C.7 Budget handoff and anchor refresh 27

C.8 Noisy-Feedback Lower Bound 28

C.9 An exact hard-budget lifting 29

C.10 Proofs for information-equivalent service control 30

D Sec. 4: Imperfect-provenance guarantees **30**

D.1 The canonical provenance experiment 31

D.2 An oracle upper bound from weighted prediction 31

D.3 Weighted comparison confidence intervals 32

D.4 Spectrum quotas and the adaptive upper bounds 33

D.5 The complete-transcript spectrum lower bound 34

D.6 Near-clean tails and the regret phase diagram 35

D.7 An exact spectrum frontier on the noisy two-action family 35

D.8 Unknown partitions of a growing tag alphabet 37

D.9 Identification and valid certificates 37

D.10 Ordered confidence banks and policy correspondence 38

D.11 Binary provenance as a specialization 40

D.12 Tagged-PCE upper bound 41

D.13 Converse for binary provenance 43

D.14 Equal accuracy with different decision complexity 44
D.15 Proof of the hidden-resampling frontier 44

E Sec. 5: Target-preserving provenance **45**
E.1 Generating and approximating provenance 45
E.2 Target-preserving provenance and the request-ID protocol 47

F Sec. 6: Experiments and numerical evaluation **48**
F.1 Unknown-gap service under a continuous provenance spectrum 48
F.2 Continuous-spectrum information and risk bounds 49
F.3 Direct Tagged-PCE mechanism experiment 51
F.4 Fixed-grid ordered-cut evaluation 52
F.5 Complete multi-action policies with balanced optimal indices 53
F.6 Decision value of provenance at a fixed score channel 55

A Supplement to Section 1: Overview, notation, and related work

A.1 Notation and proof roadmap

Appendix A.2 collects related work; later appendices place proofs and technical guarantees before implementation and experimental details.

Table 1: Principal notation.

Symbol	Meaning
$N, T, B, \overline{B}, B_\star$	Actions, production horizon, shadow cap, $B \vee 1$, and natural-family budget crossover.
$Y_t, Y, P, \mu_a, \Delta_a, a_*$	I.i.d. base loss vector, generic draw $Y \sim P$, base law, mean, gap, and witnessing optimum.
$\widetilde{Y}_{a,t}, D_{ab}$	Returned score of action a and returned pair contrast $\widetilde{Y}_a - \widetilde{Y}_b$.
$A_t, J_t, M_T, \mathcal{H}_{t-1}$	Production action, optional shadow, shadow count, and completed history.
$d_{ab}, q_{ab}, \mathcal{V}_\gamma(v, \nu)$	Mean difference, squared-difference moment, and paired Bernstein class.
$\mathcal{R}_T, R_T, \mathfrak{R}, \mathfrak{R}_{\mathrm{ch}}$	Realized pseudo-regret, expected regret, and minimax regret with its specified class/channel.
$F, \mathcal{C}, C, m_C,$ $\zeta(c), Z, G$	Resampling bit, tag space, public tag, tag law, conditional resampling probability, $Z = \zeta(C)$, and its CDF.
$\mathsf{H}, \mathsf{A}, S, \mathsf{J}, e$	Spectrum, best-threshold precision, clipped structural moment scale, information-resolution function, and its clipped inverse.
$\alpha, \beta, p_\alpha, \rho$	Near-clean tail exponent, budget exponent, effective Bernstein exponent, and regret exponent.
$K, L_{\mathrm{sp}}, \Lambda_{\mathrm{sp}}, \delta$	Phase depth, spectrum confidence logarithm, KL_{sp}, and failure probability.
$w_j, \lambda_j, W_n, U_n, V_n$	Comparison weights/scales; weighted count, weighted difference sum, and weighted square sum.
$\mathcal{S}_r, p_r, \varepsilon_r$	Active set, phase anchor, and dyadic precision.
$n_a, \widehat{\mu}_a, D_i, I_{a,p}$	Marginal count and mean, a paired difference, and its final confidence interval.
ph, all	Phase-local and historical data streams.
$P_0, P_j, \mathbb{P}_i^{\mathcal{A}}, D_{\mathrm{KL}}, \mathrm{TV}$	Reference, alternatives, observable transcript law, relative entropy, and total variation.
$\bar{\zeta}, r_\ell, \zeta_\ell, \nu_\ell$	Overall resampling rate, binary-label frequency, label-conditional contamination, and effective floor.
τ, ϕ in the binary channel	Clean-component acceptance and false-clean probability.
$T_{\mathrm{cert}}, \delta_0$ in certification	Certification horizon and error/noncompletion tolerance.
$L, \Lambda, L_{\mathrm{tag}}, \Lambda_{\mathrm{tag}}$	Baseline and binary-tag confidence logarithms and their phase factors.
$L_{\mathrm{srv}}, L_{\mathrm{cut}}$	Confidence logarithms for two-action rank service and the eleven-cut experiment.
R, E_0, V_a, s in the hard family	Rare-component coin, shared Bernoulli score, coordinate scores, and mixture probability.
E, W, g in the natural family	Label-noise bit, disagreement bit, and action gap.
$L_{\mathrm{r}}, U, \lambda_{\mathrm{r}}, \rho_{\mathrm{w}}$	Context readiness delay, wait allowance, and their exponential rates.
$\widehat{Z}, \mathsf{H}_{\widehat{Z}}, \ell, u$	Estimated contamination, its weighted precision, and relative denominator bounds.
$\mathcal{L}, c_0, \xi$	Available tag cuts, retained-mass fraction, and cut contamination from misordering.
$\varepsilon_\nu, B_\nu(k)$	Noise-floor resolution and its paired-budget scale.

A.2 Related work and theorem-level comparison

Observation budgets and correlated feedback. Classical adversarial bandits (Auer et al., 2002; Bubeck & Cesa-Bianchi, 2012) and expert prediction (Cesa-Bianchi & Lugosi, 2006) motivate the production-only baseline. Label-efficient prediction controls purchased labels (Cesa-Bianchi et al., 2005); decoupled bandits separate the action played from the action observed (Avner et al., 2012; Rouyer & Seldin, 2020). Feedback graphs encode side observations (Mannor & Shamir, 2011; Alon et al., 2015), while free side observations can change stochastic regret from logarithmic to bounded (Degenne et al., 2018). Correlated-arm bandits exploit additional cross-arm structure (Gupta et al., 2021); correlated-bandit sampling also supports a distinct mean-squared-error objective (Boda & L.A., 2019). Here the observation endpoints and their cost remain fixed while provenance changes the information in their joint scores. Fixed-confidence identification (Garivier & Kaufmann, 2016; Kaufmann et al., 2016) provides the testing perspective for our separate certification endpoint.

Difference-adaptive prediction. For bounded scores, Lemma 1 and equation (9) of Thune & Seldin (2018) bound the fixed-rate comparator second moment by $(N-1)\sum_t \mathbb{E}(Y_{A_t,t} - Y_{a_*,t})^2$. Substitution of (5), Jensen's inequality for exponent $2-\gamma \in [0,1]$, and a fixed learning rate give (6). Optimizing the rate and applying Appendix C.9 recover the clean paired exponent and its additive floor. The spectrum's oracle upper bound uses precisely this second-order argument on weighted differences. Its matching converse evaluates the complete provenance experiment, including all tag values and the free production stream.

Covariance-adaptive identification. For bounded rewards, Theorem 4.1 of Saad et al. (2023) gives, up to logarithms, a total-evaluation complexity with terms

$$\sum_{i \neq i_*} \min_{j:\mu_j > \mu_i} \left\{ \frac{\mathrm{Var}(X_i - X_j)}{(\mu_j - \mu_i)^2} + \frac{1}{\mu_j - \mu_i} \right\},$$

where X_i denotes rewards and i_* maximizes their mean. Their Theorem 6.1 provides a corresponding Bernoulli lower complexity involving optimal-reference comparison variance and inverse gaps. These results already turn covariance structure into sample savings; replacing the comparison variance by $v\Delta^{2-\gamma} + \nu$ gives the established $v\Delta^{-\gamma} + \nu\Delta^{-2}$ opportunity. Their interface counts all evaluated coordinates and permits querying subsets of a request. The present converse retains which evaluations were paired and which provenance accompanied them, producing $\Delta^2[T + B\mathsf{H}(s)]/N$. This is the constraint used both for production regret and for the certification boundary.

Task information and extra queries. Proposition 3.2 of Shufaro et al. (2025) relates Bayes regret to the mutual information between a separated decision index and the observation history, as in information-theoretic bandit analysis (Russo & Van Roy, 2016). Our equal-accuracy construction holds $I(F;C)$ fixed while the task-relevant information in the joint scores and provenance changes. Theorem 1 of Lu et al. (2024) obtains expected strongly adaptive regret $O(\sqrt{NI \log N} \log^{3/2} T)$ on intervals of length I using two queries per round; at $I = T$ its whole-horizon scale is square-root.

Heteroscedastic precision. Heteroscedastic weighting (Kirschner & Krause, 2018) and multi-fidelity sample allocation (Poiani et al., 2024) exploit observations of unequal quality. For fixed-action linear heteroscedastic bandits, Zhao et al. (2026) obtain simple-regret bounds with nearly harmonic-mean dependence on the noise variances and a nearly matching lower bound. Their protocol reveals the current variance after the action and reward. Both analyses exploit accumulated inverse-noise precision from post-decision information. Here Z describes a marginal-preserving mixture of coupled and independently resampled scores, and the intrinsic scale s varies with the paired Bernstein gap. The resulting $\mathsf{H}(s)$ enters a cumulative production-regret bound with a separate shadow budget and its complete-transcript converse, as well as a certification boundary. The equal-accuracy separation and order-only policy address which provenance semantics make this comparison precision accessible.

B Supplement to Section 2: Settings and preliminaries

The constructions below instantiate the score model and low-noise condition used in Section 2, and isolate the marginal-versus-coupled distinction used throughout the paper.

B.1 The natural coupled family

The following two-action experiment isolates the value of the joint loss law. For $0 < g \le 1/8$, let $E \sim \mathrm{Bernoulli}(1/4)$ and $W \sim \mathrm{Bernoulli}(2g)$ be independent. Define $P_{+,g}$ as the law of $(E, E \oplus W)$, where $\oplus$ is binary exclusive-or, and obtain $P_{-,g}$ by swapping its coordinates. Let $Q_{\sigma,g} = P_{\sigma,g,1} \otimes P_{\sigma,g,2}$, and set $\mathcal{F}_{\mathrm{pair}} = \{P_{\sigma,g} : \sigma \in \{+,-\},\ 0 < g \le 1/8\}$ and $\mathcal{F}_{\mathrm{ind}} = \{Q_{\sigma,g} : \sigma \in \{+,-\},\ 0 < g \le 1/8\}$. Every matched pair $P_{\sigma,g}, Q_{\sigma,g}$ has identical one-action marginal distributions. Both classes use the same production-and-shadow interface and budget.

Proposition 1 (The minimax value of coupling)**.** *For numerical constants $0 < c < C < \infty$, all integers $T \ge 1$, and all $0 \le B \le T$,*

$$c \min\{\sqrt{T}, T/B\} \le \mathfrak{R}(T, B, 2; \mathcal{F}_{\mathrm{pair}}) \le C \min\{\sqrt{T}, T/B\}, \tag{21}$$

$$c\sqrt{T} \le \mathfrak{R}(T, B, 2; \mathcal{F}_{\mathrm{ind}}) \le C\sqrt{T}. \tag{22}$$

The paired family lies in $\mathcal{V}_1(2, 0)$ and has positive mass on all four binary loss vectors under both orderings. The upper bounds use at most $\lfloor B \rfloor$ shadows and require neither the gap g nor the ordering σ.

Thus a linear shadow budget yields constant minimax production regret on $\mathcal{F}_{\mathrm{pair}}$ and square-root regret on its memberwise marginal-preserving independent version. The budget crossover is $B \asymp \sqrt{T}$, with a strict order improvement when $B/\sqrt{T} \to \infty$. Any adaptive experiment observing only one coordinate per independent request has identical transcript laws under matched $P_{\sigma,g}$ and $Q_{\sigma,g}$. Thus paired evaluation supplies the information needed to distinguish these investment regimes, even when marginal benchmarking is arbitrarily extensive. This family also has a classification realization with conditional Bayes error $1/4$: take a latent feature $X = W$, classifiers $h_1(X) = 0$, $h_2(X) = X$, and label E or $E \oplus X$ for the two orderings. The learner receives the score feedback specified above. Appendix B.2 gives a direct budgeted information proof.

B.2 A direct proof of the coupling-only separation

The paired loss distributions in Proposition 1 are

(y_1, y_2)	$(0,0)$	$(1,1)$	$(0,1)$	$(1,0)$
$P_{+,g}$	$\frac{3}{4}(1-2g)$	$\frac{1}{4}(1-2g)$	$\frac{3}{2}g$	$\frac{1}{2}g$
$P_{-,g}$	$\frac{3}{4}(1-2g)$	$\frac{1}{4}(1-2g)$	$\frac{1}{2}g$	$\frac{3}{2}g.$

(23)

Consequently, the means under $P_{+,g}$ are $(1/4, 1/4+g)$, the means under $P_{-,g}$ are reversed, and $q_{12} = 2g$. Every displayed mass is positive for $0 < g \le 1/8$. For the classification realization, under $+$ the Bayes classifier is h_1 and under $-$ it is h_2; the conditional label error is $1/4$ at each feature value, so the label margin is $1/2$.

Proof of Proposition 1. We first prove the paired upper bound. Write $n = \lfloor B \rfloor$ and $c_0 = 2 - \sqrt{3}$. When $B < \sqrt{T}$, production-only MOSS gives $C\sqrt{T}$ regret. When $B \ge \sqrt{T}$, query both actions on the first n rounds. On each of these rounds, serve the action with smaller cumulative loss on the preceding paired rounds, breaking ties fairly. After round n, serve the action with smaller cumulative loss on the n paired samples for the rest of the horizon. This rule uses n shadows and does not depend on g or σ. Under either ordering, let $D = Y_{\mathrm{bad}} - Y_{\mathrm{good}}$. Its probabilities at $1, -1, 0$ are $3g/2, g/2, 1-2g$, respectively. For $\lambda = \frac{1}{2}\log 3$,

$$\mathbb{E}e^{-\lambda D} = 1 - c_0 g. \tag{24}$$

The probability of serving the worse action after m paired observations is therefore at most $(1-c_0 g)^m$. This bound includes the randomized treatment of a zero empirical difference. The regret during the paired prefix is at most

$$g \sum_{m=0}^{n-1} (1 - c_0 g)^m \le c_0^{-1}. \tag{25}$$

The remaining regret is at most

$$(T-n)g(1-c_0 g)^n \le Tge^{-c_0 ng} \le \frac{T}{ec_0 n}. \tag{26}$$

Since $n \geq B/2$ and $T/n \geq 1$ in this case, the total is at most CT/B. The two cases prove (21)'s upper bound. Production-only MOSS proves the independent upper bound for every budget.

For the lower bounds, fix any policy satisfying the expected-budget constraint on the relevant class, and compare the transcript laws H_+ and H_- at a fixed gap g. For either ordering of the Bernoulli means,

$$\mathrm{kl}(p, q) \leq \frac{(p-q)^2}{q(1-q)} \leq 6g^2, \qquad \{p, q\} = \{1/4, 1/4 + g\}. \tag{27}$$

A paired observation under the clean experiment has divergence

$$D_{\mathrm{KL}}(P_{+,g}, P_{-,g}) = \frac{3g}{2} \log 3 + \frac{g}{2} \log(1/3) = g \log 3. \tag{28}$$

The adaptive transcript chain rule applies conditionally on the past, including the policy's randomization and its decision to use a shadow. Using budget admissibility under $P_{+,g}$ yields

$$D_{\mathrm{KL}}(\mathsf{H}_+, \mathsf{H}_-) \leq 6Tg^2 + (\log 3) g \mathbb{E}_{+,g} M_T \leq 6Tg^2 + (\log 3) Bg. \tag{29}$$

Choose $g = \frac{1}{64} \min\{T^{-1/2}, B^{-1}\}$. Then $g \leq 1/8$ and the divergence is at most $6/64^2 + (\log 3)/64 < 1/8$, hence $\mathrm{TV}(\mathsf{H}_+, \mathsf{H}_-) \leq 1/4$ by Pinsker's inequality. If N_1 counts production selections of action 1, boundedness of N_1/T gives

$$R_T(\mathcal{A}; P_{+,g}) + R_T(\mathcal{A}; P_{-,g}) = g\{T - \mathbb{E}_{+,g} N_1 + \mathbb{E}_{-,g} N_1\} \tag{30}$$

$$\geq gT\{1 - \mathrm{TV}(\mathsf{H}_+, \mathsf{H}_-)\} \geq \frac{3}{4} gT. \tag{31}$$

The larger regret is at least $3gT/8$, proving the paired lower bound with constant $3/512$.

For the independent experiment, each paired observation consists of two independent Bernoulli coordinates. Equation (27) gives

$$D_{\mathrm{KL}}(\mathsf{H}_+, \mathsf{H}_-) \leq 6g^2\{T + \mathbb{E}_{+,g} M_T\} \leq 12Tg^2. \tag{32}$$

Taking $g = 1/(16\sqrt{T})$ makes this divergence less than $1/8$. The same testing argument gives a regret of at least $3\sqrt{T}/128$ on one of $Q_{+,g}, Q_{-,g}$. All lower bounds allow history-dependent randomized use of the budget. □

B.3 A classification realization

Let (X, Z) be a labeled request with $\mathsf{Z} \in \{0, 1\}$, and let action a be a deterministic classifier f_a. Set $Y_a = \mathbf{1}\{f_a(X) \neq \mathsf{Z}\}$ and $\eta(X) = \mathbb{P}(\mathsf{Z} = 1 \mid X)$. Suppose the action set contains a Bayes classifier $f_*(X) = \mathbf{1}\{\eta(X) \geq 1/2\}$ and $|2\eta(X) - 1| \geq m_0 > 0$ almost surely. Let a_* be the index of this classifier and write $Y_* = Y_{a_*}$.

For binary predictions, $(Y_a - Y_*)^2 = \mathbf{1}\{f_a(X) \neq f_*(X)\}$ and $\mathbb{E}[Y_a - Y_* \mid X] = |2\eta(X) - 1| \mathbf{1}\{f_a(X) \neq f_*(X)\}$. Taking expectations gives $\Delta_a \geq m_0 \mathbb{P}(f_a(X) \neq f_*(X)) = m_0 q_{aa_*}$, hence $(\gamma, v, \nu) = (1, 1/m_0, 0)$ in Definition 1. The low-noise classification condition (Tsybakov, 2004; Massart & Nédélec, 2006) thus gives a paired-feedback Bernstein condition (van Erven et al., 2015).

C Supplement to Section 3: Same-request paired-feedback guarantees

The shared change-of-measure tools are stated first; the remaining subsections give the algorithms, confidence rules, regret proofs, budget accounting, and service guarantees for Section 3.

C.1 Probability and change of measure (used in Sections 3–4)

The change-of-measure argument uses standard bandit transcript accounting (Kaufmann et al., 2016; Lattimore & Szepesvári, 2020). We use natural logarithms. For $p, q \in (0, 1)$, $\mathrm{kl}(p, q) = p \log(p/q) + (1-p) \log((1-p)/(1-q))$. For $|u| \leq 1/4$,

$$\mathrm{kl}(1/2, 1/2 - u) = -\tfrac{1}{2} \log(1 - 4u^2) \leq 4u^2. \tag{33}$$

The inequality follows from $-\log(1-x) \leq x/(1-x)$ and $4u^2 \leq 1/4$.

Lemma 1 (Transcript accounting). *Let P_0, P_j be two environments with the same public information, and let $\mathbb{P}_i^{\mathcal{A}}$ denote the law of the full T-round observable transcript under policy $\mathcal{A}$ and environment i. Let $Q_{i,t}(\cdot \mid \mathcal{H}_{t-1}, a, j)$ be the conditional observation kernel under environment i, for production endpoint a and optional shadow endpoint j. For any common nonanticipating randomized policy,*

$$D_{\mathrm{KL}}(\mathbb{P}_0^{\mathcal{A}} \| \mathbb{P}_j^{\mathcal{A}}) = \mathbb{E}_0 \sum_{t=1}^{T} D_{\mathrm{KL}}\big(Q_{0,t}(\cdot \mid \mathcal{H}_{t-1}, A_t, J_t) \| Q_{j,t}(\cdot \mid \mathcal{H}_{t-1}, A_t, J_t)\big). \tag{34}$$

If only observations containing j differ, and their singleton and pair KL are at most d_s, d_p, respectively, then

$$D_{\mathrm{KL}}(\mathbb{P}_0^{\mathcal{A}} \| \mathbb{P}_j^{\mathcal{A}}) \le d_s \mathbb{E}_0 U_j + d_p \mathbb{E}_0 V_j, \tag{35}$$

where U_j counts singleton rounds observing j, and V_j counts shadow rounds whose pair contains j. If the reference is in the class defining budget admissibility, then

$$\sum_{j=1}^{N} \mathbb{E}_0 U_j \le T, \qquad \sum_{j=1}^{N} \mathbb{E}_0 V_j \le 2B. \tag{36}$$

Proof. Factor the transcript law into the policy's action kernels and the environment's observation kernels. At a fixed history, the action kernel is identical under the two environments, so it contributes zero conditional KL. Iterating the relative-entropy chain rule gives (34); integrating internal randomization gives the same formula for randomized policies. The conditional KL is zero on unchanged observations, and is bounded by d_s or d_p otherwise, giving (35). Pathwise, a non-shadow round contributes to one singleton count and a shadow round to two pair counts, so $\sum_j U_j \le T$ and $\sum_j V_j = 2M_T$. Apply the budget bound under the admissible reference. All constructions below have finite forward KL because each alternative assigns positive probability to the reference support. □

Lemma 2 (From average KL to production regret). *Suppose $P_0, P_1, \ldots, P_N$ have common public information and under P_j action j is optimal with every other action at least Δ worse. If*

$$\frac{1}{N} \sum_{j=1}^{N} D_{\mathrm{KL}}(\mathbb{P}_0^{\mathcal{A}} \| \mathbb{P}_j^{\mathcal{A}}) \le \frac{1}{8},$$

then $\max_j R_T(\mathcal{A}; P_j) \ge \Delta T/4$.

Proof. Let $N_j^{\mathrm{prod}} = \sum_{t=1}^{T} \mathbf{1}\{A_t = j\}$ be the production count of arm j. Since $N_j^{\mathrm{prod}} \in [0, T]$, the variational bound and Pinsker's inequality give

$$\mathbb{E}_j N_j^{\mathrm{prod}} \le \mathbb{E}_0 N_j^{\mathrm{prod}} + T\mathrm{TV}(\mathbb{P}_0^{\mathcal{A}}, \mathbb{P}_j^{\mathcal{A}}) \le \mathbb{E}_0 N_j^{\mathrm{prod}} + T\sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(\mathbb{P}_0^{\mathcal{A}} \| \mathbb{P}_j^{\mathcal{A}})}.$$

Average, apply concavity of the square root, and use $\sum_j \mathbb{E}_0 N_j^{\mathrm{prod}} = T$:

$$\frac{1}{N} \sum_j \mathbb{E}_j N_j^{\mathrm{prod}} \le \frac{T}{N} + \frac{T}{4} \le \frac{3T}{4}.$$

Average regret is at least $\Delta N^{-1} \sum_j (T - \mathbb{E}_j N_j^{\mathrm{prod}}) \ge \Delta T/4$, and the maximum is at least the average. □

C.2 The base two-resource frontier

Set $K = \lceil \log_2(2T) \rceil$, $L = \log(16N^2K(T+1)/\delta)$, and $\Lambda = KL$, for $\delta \in (0,1)$. PCE uses precisions $\varepsilon_r = 2^{-r}$, a phase-fixed active anchor, and complete target cycles. It intersects paired second-moment and marginal intervals, eliminating midpoint differences more than $2\varepsilon_r$ above the minimum. Appendix C.4 specifies the radii, cycle tests, singleton continuation, and inconsistent-interval fallback.

PCE-R serves the target and uses the smallest active anchor. PCE-S chooses its anchor by historical marginal mean and serves it on paired rounds. PCE-O uses the same query scheduler and chooses the served endpoint by the historical pair difference. The latter two reuse historical confidence intervals; all three continue learning from singleton production scores.

Theorem 2 (Structure-adaptive budgeted regret). *PCE-R, PCE-S, and PCE-O use at most $\lfloor B \rfloor$ shadows pathwise. For a numerical C_{reg}, every bounded i.i.d. score law satisfies*

$$R_T \le C_{\mathrm{reg}} \min\{T, \sqrt{NTL}\} + \delta T. \tag{37}$$

For $P \in \mathcal{V}_\gamma(v, \nu)$, each policy also satisfies

$$R_T \le C_{\mathrm{reg}} \min\left\{T, \sqrt{NTL}, T\left[\left(\frac{vN\Lambda}{B}\right)^{1/\gamma} + \sqrt{\frac{\nu N\Lambda}{B}}\right]\right\} + \delta T. \tag{38}$$

The policy inputs exclude γ, v, ν, a_*. The bounds hold for any phase-start measurable active anchor, either choice of historical-interval reuse, and any predictable orientation of the scheduled pairs.

On the joint event, an optimum survives, active gaps are at most $8\varepsilon_r$, and $q_{ap} \le 32v\varepsilon_r^{2-\gamma} + 4\nu$. The marginal and paired cycle quotas are $O(L\varepsilon_r^{-2})$ and $O(L[v\varepsilon_r^{-\gamma} + \nu\varepsilon_r^{-2}])$. Appendices C.5–C.6 sum these quotas and production loss, including interrupted phases and singleton continuation.

Theorem 3 (Matching noisy-information lower bound). *For all $\gamma \in [1, 2]$, $v \ge 1$, $\nu \in [0, 1]$, and admissible N, T, B,*

$$\mathfrak{R}(T, B, N; \mathcal{V}_\gamma(v, \nu)) \ge c_* \min\left\{T, \sqrt{NT}, T\left[\left(\frac{vN}{B}\right)^{1/\gamma} + \sqrt{\frac{\nu N}{B}}\right]\right\}. \tag{39}$$

Here $c_* > 0$ is numerical. With $\delta = (T+1)^{-2}$, the bounds match up to the displayed logarithms throughout $v \ge 1$.

Why the two resources have different values. For $v \ge 8$ and $0 < \Delta \le 1/64$, the hard family has

$$\begin{gathered} Y_a = (1-R)E_0 + RV_a, \quad R \sim \mathrm{Bernoulli}(s), \quad E_0 \sim \mathrm{Bernoulli}(1/2), \\ s = \min\{1/4, v\Delta^{2-\gamma} + \nu\}. \end{gathered} \tag{40}$$

The variables $R, E_0, V_1, \ldots, V_N$ are mutually independent. In reference 0, the Bernoulli means of V_a are $1/2 - \Delta/s$ for action 1 and $1/2$ otherwise; alternative $j \ge 2$ changes only action j's bias to $1/2 - 2\Delta/s$. Write $\mathbb{P}_j^{\mathcal{A}}$ for a policy's full transcript law in environment j. The construction keeps every environment in the structural class and gives

$$\frac{1}{N-1}\sum_{j=2}^{N} D_{\mathrm{KL}}(\mathbb{P}_0^{\mathcal{A}} \| \mathbb{P}_j^{\mathcal{A}}) \le \frac{32\Delta^2}{N-1}\left(T + \frac{B}{s}\right). \tag{41}$$

The singleton and pair contributions are respectively Δ^2 and Δ^2/s. Their separate counts yield (39); scaling extends the construction from $v \ge 8$ to $1 \le v < 8$ (Appendices C.8 and C.8.5).

C.3 Service and finite-budget policies

A paired observation serves one queried action and exposes the other. This separates two design choices: which set to query and which endpoint to serve. For queried sets $\mathcal{Q}_t \subset [N]$, $|\mathcal{Q}_t| \in \{1, 2\}$, define

$$R_T^{\mathrm{query}} = \mathbb{E}\sum_{t=1}^{T} \min_{a \in \mathcal{Q}_t} \Delta_a. \tag{42}$$

Even an oracle choosing the better endpoint pays this query-set cost. A *role-blind scheduler* chooses sets, stopping decisions, and budget use from arm-labeled past observations without using the service labels. On a pair $e = \{a, b\}$, $a < b$, *empirical orientation* (EO) serves the endpoint favored by the mean of its previous canonical differences $Y_a - Y_b$, using an independent fair coin at a tie. This is two-expert follow-the-leader on each pair.

Theorem 4 (Service cost on a fixed query schedule). *Predictable orientations of a common role-blind scheduler can be coupled to have identical queried sets, arm-labeled observations, stopping decisions, and shadow counts. If M_e counts queries of pair e and $g_e = |\mu_a - \mu_b|$ and $q_e = q_{ab}$ for $e = \{a, b\}$, then*

$$0 \le R_T^{\mathrm{EO}} - R_T^{\mathrm{query}} \le \sum_{e: g_e > 0} \min\left\{g_e \mathbb{E}M_e, \frac{2q_e}{g_e} + 3\right\}. \tag{43}$$

Under (88), replace q_e by its observed moment q_e^{ξ}.

A summable Bernstein bound on each pair's wrong-sign probability proves the inequality, including adaptive interruption; Appendix C.10 gives the proof. PCE-O applies EO to PCE-S and inherits Theorem 2.

Refreshing the query schedule and handing off service. Finite budgets expose two costs: an anchor can remain fixed throughout a long phase, and shadow exhaustion turns paired cycles into singleton exploration. The *active-set MOSS handoff* freezes the current active set immediately after the final affordable shadow and uses the marginal index

$$I_t(a) = \widehat{\mu}_a^{\mathrm{all}}(t) - \sqrt{\frac{\log_+\big(T/(N n_a^{\mathrm{all}}(t))\big)}{n_a^{\mathrm{all}}(t)}}. \tag{44}$$

Here $n_a^{\mathrm{all}}(t)$ and $\widehat{\mu}_a^{\mathrm{all}}(t)$ count and average every earlier observation of a, $\log_+(x) = \max\{0, \log x\}$, and an unobserved action has index $-\infty$. Serve the active action minimizing I_t and continue updating its marginal statistics. All deterministic ties use the arm index. At $B = 0$, this is production-only MOSS.

PCE-H combines this handoff with geometric within-phase anchor refresh: partition complete cycles into blocks of lengths $1, 2, 4, \ldots$; at each block boundary, test phase completion and, if the phase continues, reset the anchor to the active action with the smallest historical marginal mean while retaining all confidence statistics.

Corollary 2 (Budget handoff preserves the frontier)**.** *PCE-H satisfies the pathwise shadow budget and the bounds of Theorem 2, with a different numerical constant. It uses neither (γ, v, ν) nor an optimal reference action.*

Every cycle observes each active action. A sufficiently long geometric block meets the paired quota, with less than a factor-four total overhead. The budget-feasible phase controls surviving gaps; the inherited-data index supplies the marginal bound. Appendix C.7 handles adaptive handoff times and inherited counts.

C.4 Full algorithm specification

PCE-S maintains $\mathcal{S}_r$, a phase-fixed anchor p_r, and phase-local and historical confidence intervals.

C.4.1 Sampling, confidence intervals, and elimination

The inputs are N, T, B and $\delta \in (0, 1)$. Set

$$K = \lceil \log_2(2T) \rceil, \quad L = \log \frac{16N^2K(T+1)}{\delta}, \quad \Lambda = KL, \quad \varepsilon_r = 2^{-r}. \tag{45}$$

Let $\widehat{\mu}_a^{\mathrm{all}}$ be the mean of all previously revealed scores of arm a, assigning value zero before its first observation. At the start of phase r, choose

$$p_r = \min \arg\min_{a \in \mathcal{S}_r} \widehat{\mu}_a^{\mathrm{all}}. \tag{46}$$

The anchor is chosen from the phase-start history and stays fixed until the phase ends.

One cycle. Visit each active target once in fixed order. For $a \neq p_r$ with budget, serve p_r and shadow a; otherwise serve a alone. A k-action cycle uses k requests and at most $k - 1$ shadows, observing every action even after exhaustion. Affordable cycles sample each anchor–target pair once and the anchor k times. Store $Y_a - Y_b$ for $a < b$, reversing signs as needed.

Two confidence sources. Within a phase, write $p = p_r$, let n_a count all revealed scores of arm a, and let $\widehat{\mu}_a$ be their mean. The marginal difference interval is

$$I_{a,p}^{\mathrm{marg,ph}} = \left[\widehat{\mu}_a - \widehat{\mu}_p \pm \left(\sqrt{\frac{L}{2n_a}} + \sqrt{\frac{L}{2n_p}}\right)\right]. \tag{47}$$

For the m same-request differences $D_i = Y_{a,i} - Y_{p,i}$ observed in this phase, set

$$\overline{D}_m = \frac{1}{m}\sum_{i=1}^{m} D_i, \quad \widehat{Q}_m = \frac{1}{m}\sum_{i=1}^{m} D_i^2, \quad r_m = 2\sqrt{\frac{\widehat{Q}_m L}{m}} + \frac{4L}{m}, \quad I_{a,p}^{\mathrm{pair,ph}} = [\overline{D}_m \pm r_m]. \tag{48}$$

Algorithm 1 Service-aware paired-confidence elimination (PCE-S)

Require: N, T, B, δ; compute K, L from (45)
$\mathcal{S}_1 \leftarrow [N]$; $b \leftarrow \lfloor B \rfloor$; $t \leftarrow 1$; initialize historical statistics
for $r = 1, \ldots, K$ **do**
  if $|\mathcal{S}_r| = 1$ **then**
    Serve its arm without shadows through T; **return**
  end if
  Choose p_r by (46); reset phase-local statistics
  repeat
    for $a \in \mathcal{S}_r$ in increasing order **do**
      if $t > T$ **then return**
      end if
      if $a \neq p_r$ and $b > 0$ **then**
        $(A_t, J_t) \leftarrow (p_r, a)$; $b \leftarrow b - 1$
      else
        $(A_t, J_t) \leftarrow (a, \varnothing)$
      end if
      Reveal selected scores; update phase-local and historical statistics
      $t \leftarrow t + 1$
    end for
    Form (49) at the complete-cycle boundary
    if an intersection is empty **then**
      Run production-only MOSS through T; **return**
    end if
  until every interval has diameter at most $2\varepsilon_r$
  Form $\mathcal{S}_{r+1}$ by (50)
end for
Serve $\min \mathcal{S}_{K+1}$ without shadows for the remaining rounds

Here i indexes observations of the specified pair, and $[x \pm r] = [x - r, x + r]$. Apply the same formulas to all revealed marginal scores and all observations of the specified pair across phases to obtain $I_{a,p}^{\mathrm{marg,all}}$ and $I_{a,p}^{\mathrm{pair,all}}$. An interval with an unavailable sample mean is $\mathbb{R}$, with diameter $+\infty$. Define

$$I_{a,p}^{\mathrm{ph}} = [-1, 1] \cap I_{a,p}^{\mathrm{marg,ph}} \cap I_{a,p}^{\mathrm{pair,ph}},$$
$$I_{a,p} = I_{a,p}^{\mathrm{ph}} \cap I_{a,p}^{\mathrm{marg,all}} \cap I_{a,p}^{\mathrm{pair,all}}, \qquad I_{p,p} = \{0\}. \tag{49}$$

The historical intervals retain information across phases; the phase-local intervals provide a uniform completion bound. The four intervals may be statistically dependent. Their simultaneous coverage suffices for a valid intersection.

Phase transition. Test only after complete cycles. If an intersection is empty, stop shadowing and run production-only MOSS for the remaining horizon. Once every interval has diameter at most $2\varepsilon_r$, let $\widehat{d}_a$ be its midpoint and retain

$$\mathcal{S}_{r+1} = \left\{ a \in \mathcal{S}_r : \widehat{d}_a \leq \min_{b \in \mathcal{S}_r} \widehat{d}_b + 2\varepsilon_r \right\}. \tag{50}$$

A sole survivor is served thereafter. After phase K, serve the smallest-index survivor. If the horizon ends during a cycle, play ends immediately. Budget exhaustion leaves the marginal observations and phase transitions active.

Cyclic-service specialization. PCE-R uses the same phases and stopping rule, chooses $p_r = \min \mathcal{S}_r$, and uses $I_{a,p} = I_{a,p}^{\mathrm{ph}}$. On target a, it serves a and shadows p_r whenever $a \neq p_r$ and budget remains. Both policies are covered by Theorem 2.

C.4.2 The service cost of a comparison cycle

Proposition 2 (Role exchange preserves observations). *Fix a phase-start history, an active set $\mathcal{S}$ of size k, an anchor $p \in \mathcal{S}$, and a complete cycle with at least $k-1$ shadows available. Under ideal feedback, the anchor-service and cyclic-service allocations of that cycle have the same joint law of arm-labeled observed scores. Each uses k requests and $k-1$ shadows. Their conditional expected production regrets are respectively*

$$\ell_{\text{anchor}} = k\Delta_p, \qquad \ell_{\text{cyclic}} = \sum_{a \in \mathcal{S}} \Delta_a. \tag{51}$$

Proof. At target $a \neq p$, both allocations reveal the coordinates $\{Y_a, Y_p\}$ of the same fresh score vector; at target p, both reveal Y_p. The target order is fixed and request vectors are independent, so the complete arm-labeled observation experiments coincide. Linearity of conditional expectation gives (51): the first allocation serves p on all k requests, while the second serves every active arm once. □

Consequently, serving the anchor has smaller cycle cost exactly when its gap is below the active-set average. Equation (46) uses historical scores to favor low-loss active arms. The regret guarantee uses the weaker property that the anchor belongs to the active set. Thus service decisions can use empirical quality while preserving the comparison argument.

C.4.3 PCE-O: a matched service modification

In PCE-S, orient each scheduled pair by its historical canonical-difference EO rule and shadow the other endpoint. Leave target order, singleton rounds, anchors, confidence tests, budget accounting, and stopping unchanged. The scheduler ignores service labels, so matched PCE-S/O runs share all observations and eliminations. EO history persists when confidence reuse is ablated; experiments label these separately. PCE-R remains a distinct fixed-anchor, phase-local scheduler.

C.5 Paired Confidence Bounds

The interval uses empirical-second-moment control, as in variance-adaptive bandits and empirical Bernstein methods (Audibert et al., 2009; Maurer & Pontil, 2009). The simultaneous event below uses bounded-increment concentration and supermartingale stopping (Hoeffding, 1963; Freedman, 1975); it is a finite-horizon counterpart of time-uniform confidence sequences (Howard et al., 2021).

Lemma 3 (Second-moment confidence and shrinkage). *Let $X_1, \dots, X_m$ be i.i.d. in $[-1, 1]$, let $d = \mathbb{E}X_1$, $q = \mathbb{E}X_1^2$, and write $\overline{X} = m^{-1}\sum_i X_i$, $\widehat{Q} = m^{-1}\sum_i X_i^2$. For every $x > 0$, with probability at least $1 - 4e^{-x}$,*

$$|\overline{X} - d| \le 2\sqrt{\widehat{Q}x/m} + 4x/m, \tag{52}$$

$$2\sqrt{\widehat{Q}x/m} + 4x/m \le 3\sqrt{qx/m} + 7x/m. \tag{53}$$

Proof. We first state the elementary concentration step with conservative constants. Put $u = x/m$. Bernstein's inequality applied to $X_i - d$, which has absolute value at most 2 and variance at most q, and to $X_i^2 - q$, which has absolute value at most 1 and variance at most q, gives simultaneously, except on a set of probability at most $4e^{-x}$,

$$|\overline{X} - d| \le \sqrt{2qu} + \tfrac{4}{3}u, \qquad |\widehat{Q} - q| \le \sqrt{2qu} + \tfrac{2}{3}u. \tag{54}$$

For completeness, the scalar Bernstein step follows by expanding the moment generating function of a centered U with $|U| \le b$: for $0 \le \lambda < 3/b$, the bounds $\mathbb{E}|U|^k \le b^{k-2}\mathbb{E}U^2$ and $k! \ge 2\,3^{k-2}$ give

$$\mathbb{E}e^{\lambda U} \le \exp\left(\frac{\lambda^2 \mathbb{E}U^2}{2(1 - \lambda b/3)}\right).$$

Multiplication for independent variables and Chernoff optimization yield the upper tail at threshold $\sqrt{2m\sigma^2 x} + bx/3$, where σ^2 bounds each variance. Apply the same argument to $-U$. Increasing the linear constants gives (54).

Now $q \le \widehat{Q} + \sqrt{2qu} + 2u/3$. Since $\sqrt{2qu} \le q/2 + u$, we obtain $q \le 2\widehat{Q} + 10u/3$. Substitution into the first inequality gives

$$|\overline{X} - d| \le 2\sqrt{\widehat{Q}u} + (\sqrt{20/3} + 4/3)u \le 2\sqrt{\widehat{Q}u} + 4u.$$

For shrinkage, $\sqrt{2qu} \le q + u/2$ yields $\widehat{Q} \le 2q + 7u/6$. Consequently,

$$2\sqrt{\widehat{Q}u} + 4u \le 2\sqrt{2qu} + (2\sqrt{7/6} + 4)u \le 3\sqrt{qu} + 7u.$$

These are exactly (52) and (53). □

Lemma 4 (Uniform coverage under adaptive sampling and reuse)**.** *With K, L as in (45), there is an event $\mathcal{G}$ with probability at least $1 - \delta$ on which every phase-local and historical interval used by PCE-S, PCE-R, or PCE-O covers its population difference at every complete-cycle check. All their paired radii also satisfy (53) with $x = L$ and population moment q_{ab}. The event remains valid for every predictable active-anchor rule, including changes within a phase.*

Proof. We first justify concentration at a fixed sample count under predictable observation choices. Fix an arm a, or a canonically oriented pair (a, b) with $a < b$. Whether it is observed on request t is decided before Y_t. Conditional on the completed history and this decision, its new scalar score or difference has its fixed population law. The centered-observation moment-generating bounds used in Hoeffding's inequality and Lemma 3 therefore hold conditionally at each selected request.

More explicitly, let Z_t^{obs} indicate the selected stream, let X_t be its value, and let $\psi(\lambda)$ bound $\log \mathbb{E} \exp(\lambda(X_t - \mathbb{E}X_t))$. Then

$$\exp\left\{\lambda \sum_{s=1}^{t} Z_s^{\mathrm{obs}}(X_s - \mathbb{E}X_s) - \psi(\lambda) \sum_{s=1}^{t} Z_s^{\mathrm{obs}}\right\}$$

is a nonnegative supermartingale. To evaluate it at the nth selected observation, extend the experiment after the horizon with independent draws of that stream. The nth observation then occurs by time $T + n$. Optional sampling at this bounded stopping time and Chernoff's argument give the usual fixed-n concentration bound. On paths having at least n actual observations, the extended and actual prefixes coincide. Apply the same argument to $X_t^2 - \mathbb{E}X_t^2$ for the second-moment bounds. Thus the fixed-sample failures are at most $2e^{-L}$ for a marginal mean and $4e^{-L}$ for the paired coverage and shrinkage event, even when observation times are adaptive.

Distinct endpoints imply at most T samples per historical arm or pair stream. Union-bounding over counts, means, and ordered pairs costs $\{2N + 4N(N-1)\}Te^{-L} \le 4N^2Te^{-L}$, including canonical sign reversal. Marginal difference intervals add the two armwise radii.

For each phase, condition on its starting history. The active set is fixed until its next elimination, future score vectors are independent of that history, and all observation decisions remain predictable even if the anchor changes. The same conditional argument bounds failure of every phase-local arm and pair stream, uniformly over sample counts, by $4N^2Te^{-L}$. The conditional bound is uniform in the starting history, so taking expectations and summing over at most K phases is valid. Including the historical event, the total failure probability is at most

$$4N^2(K+1)Te^{-L} = \delta \frac{(K+1)T}{4K(T+1)} \le \delta/2 \le \delta.$$

The union bound applies jointly to the dependent intervals. It covers repeated checks, adaptive phase starts, data reuse, and budget-dependent stopping of the paired streams. □

C.6 Analysis of All PCE Service Variants

We establish pathwise bounds on the event $\mathcal{G}$ from Lemma 4, then integrate these bounds to obtain expected regret.

C.6.1 Correct elimination and active-set geometry

Lemma 5 (Safe phase transition). *An optimal action is never eliminated on $\mathcal{G}$. After completion of phase r, every survivor has $\Delta_a \le 4\varepsilon_r$. At the start of phase r, every active action satisfies $\Delta_a \le 8\varepsilon_r$.*

Proof. Fix an optimal action a_* in $\mathcal{S}_r$ and anchor p. Put $d_a = \mu_a - \mu_p$. Since all intersections contain their true differences, none is empty. When a phase ends, their diameters are at most $2\varepsilon_r$, so their midpoints satisfy $|\widehat{d}_a - d_a| \le \varepsilon_r$. The minimum true difference is d_{a_*}. Hence

$$\min_b \widehat{d}_b \ge d_{a_*} - \varepsilon_r, \qquad \widehat{d}_{a_*} \le d_{a_*} + \varepsilon_r,$$

which proves that a_* satisfies the elimination threshold. For any survivor a,

$$d_a \le \widehat{d}_a + \varepsilon_r \le \widehat{d}_{a_*} + 3\varepsilon_r \le d_{a_*} + 4\varepsilon_r.$$

Thus $\Delta_a \le 4\varepsilon_r$. Induct from $\mathcal{S}_1 = [N]$. Its gaps are at most 1, which is at most $8\varepsilon_1 = 4$; in later phases $4\varepsilon_{r-1} = 8\varepsilon_r$. The same fixed optimal action may be followed throughout the induction. □

Fix an optimal action a_* witnessing Definition 1; safe elimination retains this action. For any two active actions a, p at the start of phase r,

$$q_{ap} = \mathbb{E}(Y_a - Y_p)^2 \le 2\mathbb{E}(Y_a - Y_{a_*})^2 + 2\mathbb{E}(Y_p - Y_{a_*})^2 \tag{55}$$

$$\le 4v(8\varepsilon_r)^{2-\gamma} + 4\nu \le 32v\varepsilon_r^{2-\gamma} + 4\nu. \tag{56}$$

The inequality holds at $\gamma = 2$ with the stated power convention and also for phase 1. The retained optimum provides a common reference for this bound on all active-action pairs.

C.6.2 Number of cycles at a given precision

Lemma 6 (Phase completion bounds). *On $\mathcal{G}$, a phase with precision ε and k active arms completes after at most*

$$m^{\mathrm{marg}}(\varepsilon) = \lceil 2L\varepsilon^{-2} \rceil \tag{57}$$

complete cycles, irrespective of the remaining shadow budget, unless the horizon ends first. If enough shadow budget remains for these paired cycles, it also completes after at most

$$m^{\mathrm{pair}}(\varepsilon) = \left\lceil C_{\mathrm{ph}} L\left(v\varepsilon^{-\gamma} + \nu\varepsilon^{-2}\right) \right\rceil, \qquad C_{\mathrm{ph}} = 4096. \tag{58}$$

Before such completion, its production regret is at most 8ε times the number of production rounds. A cycle uses k production rounds and at most $k-1$ shadows.

Proof. After m complete cycles each marginal mean has at least m samples, so the marginal difference half-width is at most $\sqrt{2L/m}$; at (57) it is at most ε. Intersecting with other nonempty covering intervals can only reduce diameter.

While all scheduled shadows are affordable, each nonanchor–anchor pair has one sample per cycle. By (56) and Lemma 3, at any $m \ge m^{\mathrm{pair}}(\varepsilon)$ its radius is at most

$$3\sqrt{\frac{(32v\varepsilon^{2-\gamma} + 4\nu)L}{m}} + \frac{7L}{m} \le \left(3\sqrt{32/C_{\mathrm{ph}}} + 7/C_{\mathrm{ph}}\right)\varepsilon < \varepsilon.$$

Here $v \ge 1$, $\gamma \ge 1$, and $\varepsilon \le 1$ imply $L/m \le \varepsilon^{\gamma}/(C_{\mathrm{ph}} v) \le \varepsilon/C_{\mathrm{ph}}$. Both completion tests are at full-cycle boundaries, so these integer cycle counts suffice without an additional cycle. Every served action belongs to the active set under any of the three service allocations, and every cycle observes each target at least once. Lemma 5 therefore gives the regret and budget statements for all three allocations. Historical intersections preserve coverage and can only decrease interval diameters, so the phase-local quotas apply with or without data reuse. A final partial cycle costs no more than the full cycle used in the bound. □

The constant C_{ph} bounds the phase length in the analysis. Algorithm 1 implements the data-dependent stopping rule in (48).

C.6.3 The production-only guarantee

Let $u = \sqrt{NL/T}$. If $u \geq 1/2$, the pathwise bound $\mathcal{R}_T \leq T$ already implies the desired estimate with a numerical constant. Otherwise choose a phase index r_* with $u \leq \varepsilon_{r_*} < 2u$. Such an index exists with $r_* \leq K$, since $u \geq 1/T > 2^{-K}$ when $T > 1$; the case $T = 1$ is trivial. Until this phase completes, Lemma 6 bounds the regret, even if the horizon truncates an earlier phase, by

$$\begin{aligned}\sum_{r=1}^{r_*} 8\varepsilon_r N\lceil 2L\varepsilon_r^{-2}\rceil &\leq 16NL\sum_{r=1}^{r_*}\varepsilon_r^{-1} + 8N\sum_{r=1}^{r_*}\varepsilon_r \\ &\leq 32NL/u + 8N.\end{aligned}$$

If the target phase completes, all subsequent active arms have gap at most $4\varepsilon_{r_*} < 8u$ because the active sets only shrink. The remaining regret is therefore at most $8Tu$. Since $u < 1/2$ implies $4NL < T$, the term $8N$ is absorbed in a numerical multiple of $\sqrt{NTL}$. The result is $C\sqrt{NTL}$ on $\mathcal{G}$, clipped by T. This production-marginal argument applies to every bounded score law and continues after the shadow budget is exhausted.

C.6.4 A budget-feasible target for the paired guarantee

Assume $B \geq 1$; otherwise the proposed paired term is at least T. Set $C_0 = 64C_{\mathrm{ph}}$ and define the following *analysis-only* precision:

$$u = \left(\frac{C_0 vNLK}{B}\right)^{1/\gamma} + \sqrt{\frac{C_0\nu NLK}{B}}. \tag{59}$$

If $u \geq 1/2$, the trivial regret bound is a numerical multiple of Tu, so suppose $u < 1/2$. Choose $r_* \leq K$ with $u \leq \varepsilon_{r_*} < 2u$. Existence follows as above: $u \geq 1/T$ because $B \leq T$, $v, N, L, K \geq 1$, and $\gamma \in [1, 2]$.

We first show that the available budget covers every phase through this target. Since the quantity inside the ceiling in (58) is at least 1, $m^{\mathrm{pair}}(\varepsilon) \leq 2C_{\mathrm{ph}}L(v\varepsilon^{-\gamma} + \nu\varepsilon^{-2})$. Define W_* as the sum of the per-phase shadow quotas through r_*:

$$W_* = \sum_{r=1}^{r_*} Nm^{\mathrm{pair}}(\varepsilon_r) \tag{60}$$

$$\leq 4C_{\mathrm{ph}}NL(vu^{-\gamma} + \nu u^{-2}) \leq \frac{8C_{\mathrm{ph}}B}{C_0K} = \frac{B}{8K}. \tag{61}$$

We used the geometric sums $\sum_r \varepsilon_r^{-\gamma} \leq 2\varepsilon_{r_*}^{-\gamma}$ and $\sum_r \varepsilon_r^{-2} \leq 2\varepsilon_{r_*}^{-2}$, and the two separate lower bounds on u in (59). Because $\lfloor B\rfloor \geq B/2$, this upper bound is smaller than the implemented budget. Induct over phases: if earlier phases have ended within their paired completion bounds, enough budget remains to finish the next such bound; hence the budget cannot be exhausted before r_* completes. If the horizon ends earlier, all preceding partial work is still bounded by the same quotas. This induction establishes phase completion and the resource bound jointly.

Let $\mathcal{R}_{\mathrm{coarse}}$ be the realized pseudo-regret incurred up to the target or the end of the horizon. Then

$$\mathcal{R}_{\mathrm{coarse}} \leq 16C_{\mathrm{ph}}NL\left(v\sum_{r=1}^{r_*}\varepsilon_r^{1-\gamma} + \nu\sum_{r=1}^{r_*}\varepsilon_r^{-1}\right) \tag{62}$$

$$\leq 16C_{\mathrm{ph}}NL\left(vKu^{1-\gamma} + 2\nu/u\right) \tag{63}$$

$$\leq \frac{48C_{\mathrm{ph}}}{C_0}Bu \leq Tu. \tag{64}$$

For the last inequality use $u^\gamma \geq C_0 vNLK/B$ and $u^2 \geq C_0\nu NLK/B$, treating the latter term as zero when $\nu = 0$. If the target phase completes, surviving gaps are below $8u$ and remain so; the remainder contributes at most $8Tu$, regardless of subsequent shadow depletion. Thus regret on $\mathcal{G}$ is at most $9Tu$.

Since $C_0^{1/\gamma} \leq C_0$, $\sqrt{C_0} \leq C_0$, and the constants are numerical, this is the paired term of (38). The factor K was retained to cover $\gamma = 1$ uniformly: $\sum_r \varepsilon_r^{1-\gamma} \leq Ku^{1-\gamma}$.

Completion of Theorem 2. The budget counter decreases only on shadow rounds and never below zero, giving the pathwise cap. Both bounds on $\mathcal{R}_T$ hold on the same event $\mathcal{G}$, so their minimum and the trivial bound T hold there simultaneously. Since $R_T = \mathbb{E}\mathcal{R}_T$ and $\mathcal{R}_T \le T$ on $\mathcal{G}^c$, taking expectations adds at most δT. Single-survivor stopping contributes zero regret on $\mathcal{G}$. The phase cap is harmless: after phase K, the surviving gaps are at most $4 \cdot 2^{-K} \le 2/T$, and both target-precision arguments reach their relevant target no later than K. Empty intersections occur only outside $\mathcal{G}$. This proves (37) and (38), including budget exhaustion and horizon truncation. □

C.7 Budget handoff and anchor refresh

PCE-H uses PCE-O's confidence intersections, elimination rule, and empirical orientation, adds geometric within-phase anchor refresh, and applies the active-set MOSS handoff of Appendix C.3 at shadow exhaustion. This subsection proves Corollary 2.

Lemma 7 (Phase quotas under geometric refresh)**.** *On $\mathcal{G}$, PCE-H retains an optimum and the active-gap invariant of Lemma 5. Its marginal phase quota is $m^{\mathrm{marg}}(\varepsilon)$. Given sufficient shadow budget, its paired phase quota is less than $4m^{\mathrm{pair}}(\varepsilon)$.*

Proof. Every complete cycle observes each active arm at least once, irrespective of the anchor. Hence the original marginal quota still suffices. The first refresh block whose length is at least $m = m^{\mathrm{pair}}(\varepsilon)$ has length less than $2m$. During that block every current anchor–target pair acquires m observations, so Lemma 6 forces completion. Preceding blocks have total length one less than this block length, giving fewer than $4m$ cycles. All anchors remain active, so (56) applies throughout. Phase-local pair streams retain their observations when the anchor changes; the original simultaneous confidence event covers each such predictably sampled stream. The phase-completion and elimination tests are unchanged, proving the remaining claims. □

Lemma 8 (Marginal learning with inherited observations)**.** *Let τ be a stopping time and S a history-measurable action set containing a fixed optimum on an event G. After τ, use the MOSS index with global parameters T, N, all earlier marginal observations, and actions restricted to S. For bounded stochastic scores,*

$$\mathbb{E}\left[\mathbf{1}_G \sum_{t>\tau} \Delta_{A_t}\right] \le C\sqrt{NT}.$$

If $\mathbb{P}(G^c) \le \delta$, the expected remaining regret is at most $C\sqrt{NT} + \delta T$.

Proof. The case $N > T$ follows from regret at most T. Otherwise set $d = \sqrt{N/T}$ and $b(n) = \sqrt{\log_+(T/(Nn))/n}$. Complete each adaptively revealed arm stream with independent draws after the horizon. Write $\widehat{\mu}_a(n)$ for its first-n mean and

$$Z_{\mathrm{idx}} = \max_{1 \le n \le T} (\widehat{\mu}_{a_*}(n) - \mu_* - b(n))_+.$$

Maximal Hoeffding peeling gives $\mathbb{E}Z_{\mathrm{idx}} \le Cd$. Here is an explicit summability argument. Put $h = T/N$. For sample counts above h, geometrically spaced blocks $[m, (3/2)m]$ give upper-tail bounds $\exp(-cmz^2)$, whose integrals sum to $C/\sqrt{h}$. Below h, use blocks $[h/(3/2)^{j+1}, h/(3/2)^j]$. If the lower endpoint is m, the bonus is at least $\sqrt{j\log(3/2)/((3/2)m)}$. The maximal bounded-increment inequality at threshold $m(z + b_{\min})$ gives $\exp[-(2m/(3/2))(z + b_{\min})^2]$. Its integral is at most $Ch^{-1/2}(3/2)^{j(1/2-2/(3/2)^2)}$, which is summable. Rounding endpoints changes only the numerical constant.

On G, a chosen action with positive current count n satisfies $\widehat{\mu}_a(n) - \mu_a \le -\Delta_a + Z_{\mathrm{idx}} + b(n)$. Actions with $\Delta_a \le 2Z_{\mathrm{idx}} + d$ contribute at most $T(2Z_{\mathrm{idx}} + d)$. Each selection of any remaining action corresponds to a distinct count n, even when its inherited count and membership in S depend on the earlier history. Its number of selections is bounded by

$$1 + \sum_{n=1}^{T} \mathbf{1}\{\widehat{\mu}_a(n) - \mu_a \le -\Delta_a/2 + b(n)\}.$$

For $n \geq \lceil 32\Delta_a^{-2}[1 + \log_+(T\Delta_a^2/N)]\rceil$, one has $b(n) \leq \Delta_a/4$. Hoeffding's inequality and a geometric sum therefore bound the expected count by $C[1 + \log_+(T\Delta_a^2/N)]/\Delta_a^2$. After multiplying by Δ_a and summing over $\Delta_a > d$, boundedness of $(1+2\log x)/x$ for $x \geq 1$ gives $CN/d = C\sqrt{NT}$. Together with $T\mathbb{E}(2Z_{\mathrm{idx}} + d)$, this proves the claim. The argument uses the unconditioned completed sample streams and covers arbitrary inherited counts. The complement of G contributes at most δT. □

Proof of Corollary 2. Before handoff, the original marginal phase argument bounds prefix regret by $C\sqrt{NTL}$ on $\mathcal{G}$. Lemma 8 bounds the remaining expected regret, giving the marginal branch. For the paired branch, apply the original target-precision induction with the paired phase constant multiplied by four. If the horizon ends first, its entire regret is covered by the prefix quota. Otherwise the target phase finishes before shadow exhaustion, and all subsequent active gaps are at most $C\varepsilon_B$, where

$$\varepsilon_B = \left(\frac{vN\Lambda}{B}\right)^{1/\gamma} + \sqrt{\frac{\nu N\Lambda}{B}}.$$

The inherited active set thus bounds all later regret by $CT\varepsilon_B$. The original geometric-sum argument bounds the prefix at the same order. The event $\mathcal{G}^c$ contributes at most δT once. Every shadow still decrements the unchanged budget counter, proving pathwise feasibility. □

C.8 Noisy-Feedback Lower Bound

C.8.1 Reference law inside the moment class

Fix $v \geq 8$, $\gamma \in [1, 2]$, $\nu \in [0, 1]$, and $0 < \Delta \leq 1/64$. Let s be as in (40). Because $\Delta^{2-\gamma} \geq \Delta$ and $1/4 \geq 8\Delta$, we have $s \geq 8\Delta > 0$. Draw $R \sim \mathrm{Bernoulli}(s)$, $E_0 \sim \mathrm{Bernoulli}(1/2)$, and independent V_a, independent of R, E_0. Set $Y_a = (1-R)E_0 + RV_a$.

Under P_0, let $\theta_1 := \mathbb{P}(V_1 = 1) = 1/2 - \Delta/s, \qquad \theta_a = 1/2\ (a \neq 1)$. Under P_j for $j \in \{2, \ldots, N\}$, change only θ_j to $1/2 - 2\Delta/s$. All bit parameters lie in $[1/4, 1/2]$. Under P_0, $\mu_1 = 1/2 - \Delta$ and all other means are $1/2$. Under P_j, $\mu_j = 1/2 - 2\Delta$, $\mu_1 = 1/2 - \Delta$, and other means remain $1/2$.

For every pair, $\mathbb{E}(Y_a - Y_b)^2 \leq \mathbb{P}(R = 1) = s$. Under each environment there is a unique optimum, and each nonzero optimality gap is at least Δ. Consequently,

$$\mathbb{E}(Y_a - Y_{a_*})^2 \leq s \leq v\Delta^{2-\gamma} + \nu \leq v\Delta_a^{2-\gamma} + \nu$$

for every nonoptimal arm; for the optimal arm the left side is zero. Hence the reference and all alternatives belong to $\mathcal{V}_\gamma(v, \nu)$. Budget admissibility therefore gives $\mathbb{E}_0 M_T \leq B$. The learner observes the selected score coordinates; R is latent.

C.8.2 Information bounds

Only observations containing j differ between P_0 and P_j. A singleton of j has marginal $\mathrm{Bernoulli}(1/2)$ versus $\mathrm{Bernoulli}(1/2 - 2\Delta)$, so

$$d_s \leq 16\Delta^2. \tag{65}$$

For a pair containing j, reveal R in an augmented experiment. Data processing and conditional independence when $R = 1$ give

$$d_p \leq s\,\mathrm{kl}(1/2, 1/2 - 2\Delta/s) \leq \frac{16\Delta^2}{s}. \tag{66}$$

When $R = 0$, the shared bit has the same distribution under both laws and contributes no KL. Data processing bounds the score-feedback information by this augmented experiment that reveals R.

Let $d = N - 1$. Averaging Lemma 1 over $j = 2, \ldots, N$ yields

$$\frac{1}{d}\sum_{j=2}^{N} D_{\mathrm{KL}}(\mathbb{P}_0^{\mathcal{A}} \| \mathbb{P}_j^{\mathcal{A}}) \leq 32\left(\frac{T\Delta^2}{d} + \frac{B\Delta^2}{ds}\right). \tag{67}$$

The same inequality without averaging holds for $N = 2$ and $d = 1$.

C.8.3 Choosing the gap, including the clipped mixture

Define

$$x = \left(\frac{vd}{\overline{B}}\right)^{1/\gamma}, \qquad z = \sqrt{\frac{\nu d}{\overline{B}}}, \qquad \Delta = \kappa \min\{1, \sqrt{d/T}, \max\{x, z\}\},$$

with $\kappa = 2^{-14}$. Then $T\Delta^2/d \le \kappa^2$. If $s = 1/4$, $B\Delta^2/(ds) \le 4\kappa^2$ because $B \le T$. If $s < 1/4$ and $x \ge z$, use $s \ge v\Delta^{2-\gamma}$ and $\Delta \le \kappa x$ to obtain $\frac{B\Delta^2}{ds} \le \frac{B\Delta^\gamma}{vd} \le \kappa^\gamma \frac{B}{\overline{B}} \le \kappa$. If $s < 1/4$ and $z > x$, necessarily $\nu > 0$, and $s \ge \nu$ gives $\frac{B\Delta^2}{ds} \le \frac{B\Delta^2}{d\nu} \le \kappa^2 \frac{B}{\overline{B}} \le \kappa^2$. Thus (67) is below $1/8$ uniformly over all cases. The construction is valid because $\Delta \le \kappa \le 1/64$.

C.8.4 Converting information to regret

For $N \ge 3$, $d \ge 2$ and $\sum_{j=2}^N \mathbb{E}_0 N_j^{\mathrm{prod}} \le T$. The same Pinsker calculation as in Lemma 2 gives

$$\frac{1}{d}\sum_{j=2}^N \mathbb{E}_j N_j^{\mathrm{prod}} \le \frac{T}{d} + \frac{T}{4} \le \frac{3T}{4}.$$

Under P_j, every arm other than j has gap at least Δ, so some alternative has regret at least $\Delta T/4$. For $N = 2$, use the reference as well. Its regret is $R_0 = \Delta \mathbb{E}_0 N_2^{\mathrm{prod}}$, and the alternative's regret is $R_2 = \Delta(T - \mathbb{E}_2 N_2^{\mathrm{prod}})$. Their transcript total variation is at most $1/4$, hence

$$R_0 + R_2 \ge \Delta T(1 - \mathrm{TV}(\mathbb{P}_0^{\mathcal{A}}, \mathbb{P}_2^{\mathcal{A}})) \ge 3\Delta T/4.$$

At least one is $\ge 3\Delta T/8$. Finally, $N/2 \le d \le N$, and $\max\{x, z\} \le x + z \le 2\max\{x, z\}$. Replacing d by N and the maximum by the sum changes the rate only by numerical factors, proving Theorem 3.

C.8.5 Extending the moment constant by rescaling

For $1 \le v \le 8$, define $S(P)$ as the law of $Y/8$ for $Y \sim P$. If $P \in \mathcal{V}_\gamma(8, \nu)$, then for its witnessing optimum

$$q'_{aa_*} = q_{aa_*}/64 \le 8^{1-\gamma}(\Delta'_a)^{2-\gamma} + \nu/64 \le v(\Delta'_a)^{2-\gamma} + \nu.$$

Thus $S(P) \in \mathcal{V}_\gamma(v, \nu)$ for the entire source class. Given any target-class budget-admissible policy $\mathcal{A}$, define a source-class policy $\mathcal{A}'$ that feeds every revealed score divided by eight to $\mathcal{A}$ and takes its chosen actions. All observation times and shadow counts agree in law, so $\mathbb{E}_P M_T(\mathcal{A}') = \mathbb{E}_{S(P)} M_T(\mathcal{A}) \le B$ for every source-class law. Also $R_T(\mathcal{A}'; P) = 8R_T(\mathcal{A}; S(P))$. Applying the proved source-class lower bound to $\mathcal{A}'$ and taking the infimum over target policies gives one eighth of the $v = 8$ rate. That rate is nondecreasing in v, so it is at least one eighth of the desired $v \in [1, 8]$ rate. The constants remain numerical and uniform in γ, ν, N, T, B. The scaled family has common support $\{0, 1/8\}^N$ and lies within the bounded-loss model.

C.9 An exact hard-budget lifting

Lemma 9 (Paired-prefix continuation)**.** *Let an algorithm use one shadow on each of $m \ge 1$ i.i.d. requests and incur expected production regret $f(m)$. For $T \ge m$, run that algorithm for the first m requests. On each later request, independently choose a uniform index in $[m]$ and serve the production action recorded at that index. This policy uses exactly m shadows and has expected regret $Tf(m)/m$.*

Proof. Conditional on the first m requests, every continuation action has expected gap $m^{-1}\sum_{s \le m} \Delta_{A_s}$. Its unconditional value is $f(m)/m$, by stationarity. Adding the m prefix requests and the $T - m$ continuation requests proves the equality. The selected past index is drawn before the current score vector, so the policy is admissible. □

Taking $m = \lfloor B \rfloor \ge 1$ gives the exact hard-budget continuation used by the spectrum oracle. At $B < 1$, production-only MOSS supplies the marginal guarantee.

C.10 Proofs for information-equivalent service control

C.10.1 Coupling the complete query experiment

Couple orientations using identical scheduler randomization and latent score vectors. Equal arm-labeled histories imply equal next query sets, stopping, and shadow decisions; orientation changes none of the revealed coordinates. Induction therefore preserves histories through adaptive anchors, partial cycles, stopping, and budget depletion. Under contamination, use the equivalent i.i.d. vectors Q from Appendix D.15, whose pair law is also orientation-invariant.

For a realized query path, the pseudo-regret has the exact decomposition

$$\mathcal{R}_T = \sum_{t=1}^{T} \min_{a \in \mathcal{Q}_t} \Delta_a + \sum_{t:\,|\mathcal{Q}_t|=2} g_{\mathcal{Q}_t} \mathbf{1}\{A_t \text{ has the larger mean in } \mathcal{Q}_t\}.$$

The second term is zero at ties and nonnegative otherwise, yielding the oracle lower comparison for the realized query schedule.

C.10.2 Expected orientation error on adaptive pair streams

Fix a pair e with gap $g > 0$ and moment q. Reverse its canonical sign for the analysis if necessary so that its difference has positive mean g. Its orientation is wrong only when its previous empirical difference is nonpositive. With $n \geq 1$ previous observations, Bernstein's inequality gives

$$\mathbb{P}\{\overline{D}_n \leq 0\} \leq \exp\left\{-\frac{ng^2}{2q + 4g/3}\right\}. \tag{68}$$

The centered difference is bounded in absolute value by two and its variance is at most q. Using the exact variance instead could only sharpen the bound.

To analyze adaptive stopping, append independent draws of a fixed pair after the horizon until n observations exist. The decision to observe a pair at an actual request is predictable, and every new observed value has the fixed population law conditionally on the completed past. The extended fixed-count stream obeys the usual concentration bound, as in Lemma 4. The event that query $n+1$ actually occurs and has a wrong previous sign is a subset of the fixed-count bad-sign event in this extended stream. Its probability is bounded by (68), regardless of adaptive interruption. At $n = 0$, bound the probability by one.

Writing $c = g^2/(2q + 4g/3)$ and summing nonnegative contributions,

$$\mathbb{E}[\text{orientation cost on } e] \leq g \sum_{n=0}^{T-1} e^{-nc} \leq \frac{g}{1 - e^{-c}} \leq g + \frac{2q}{g} + \frac{4}{3} \leq \frac{2q}{g} + 3.$$

We used $1/(1 - e^{-c}) \leq 1 + 1/c$ and $g \leq 1$. Pathwise the same cost is at most gM_e, giving the other side of the minimum in (43). Summation over unordered pairs proves Theorem 4. Replacing q by $q^{\bar{\zeta}}$ proves the contaminated version.

C.10.3 Application to PCE

PCE-S uses only arm-labeled scores and budget, hence is role-blind. EO orientation preserves its observation paths and phase transitions. Both endpoints remain active, so $\Delta_{A_t} \leq 8\varepsilon_r$ still holds on the joint event, extending Theorem 2 to PCE-O.

The coupling holds for orientations of the same scheduler. Historical-interval reuse changes the scheduler and is a separate design choice. Singleton rounds contribute directly to R_T^{query}.

D Supplement to Section 4: Imperfect-provenance guarantees

This appendix collects the provenance-spectrum characterization, calibrated and ordered confidence constructions, lower bounds, certification results, and binary-provenance specializations supporting Section 4.

D.1 The canonical provenance experiment

A channel is a fixed probability law m_C on the standard Borel tag space $\mathcal{C}$ and a measurable calibration $\zeta : \mathcal{C} \to [0,1]$. Conditional on $C = c$, the observed pair has law (7). Its single-coordinate marginals are P_a, independently of c. Drawing fresh tags and conditional pairs on every request realizes the stipulated post-decision experiment. The shadow counter increments on every paired request.

Posterior sufficiency. Given $Z = \zeta(C)$, the conditional score law is independent of the remaining tag information. The standard Borel assumption supplies an environment-independent regular conditional kernel $m(dc \mid Z = z)$. Reconstructing C from this kernel, conditionally independently of scores, simulates any original policy. Induction preserves the joint law of its actions, scores, stopping decisions, and shadow counts. Conversely, C determines Z. Hence calibrated Z preserves every decision guarantee.

The spectrum and its inverse. Cauchy–Schwarz gives $(\mathbb{E}w)^2 \le \mathbb{E}[w^2(s+Z)]\,\mathbb{E}[(s+Z)^{-1}]$, and equality holds for $w = s/(s+Z)$. This proves (8). For $a = 2-\gamma \in [0,1]$, clipping preserves $c^a S(\varepsilon) \le S(c\varepsilon) \le S(\varepsilon), \qquad 0 < c \le 1$. Integrating the corresponding reciprocal inequalities gives

$$c^2 \mathsf{J}(\varepsilon) \le \mathsf{J}(c\varepsilon) \le c^\gamma \mathsf{J}(\varepsilon). \tag{69}$$

The function J is continuous and strictly increasing on $(0,1]$. Since $S(\varepsilon) \ge \varepsilon$, we have $\mathsf{J}(\varepsilon) \le \varepsilon$, proving continuity at zero. Its inverse e in (9) is therefore well defined. Multiplying its argument by a fixed numerical constant changes the clipped inverse by at most a numerical factor, uniformly for $\gamma \in [1,2]$.

D.2 An oracle upper bound from weighted prediction

This policy is supplied with the full channel $G, \zeta(C)$ and the structural parameters γ, v, ν. It is a fixed-rate weighted version of SODA's difference estimator, followed by the budget continuation already proved in Appendix C.9.

Let $m \ge 1$, $d = N-1$, $k_N = N\log(2N)$, and $\varepsilon = e(k_N/m)$. If $\varepsilon = 1$, the trivial paired-prefix regret bound $R_m \le m\varepsilon$ suffices. Otherwise the inverse is interior and $m\mathsf{J}(\varepsilon) = k_N$. First consider the untruncated case

$$\lambda = v\varepsilon^{2-\gamma} + \nu < 1, \qquad w(z) = \frac{\lambda}{\lambda + z}, \qquad A = \mathbb{E}w(Z) = \lambda \mathsf{H}(\lambda).$$

In particular, $\lambda \ge \varepsilon$ and $\lambda \ge \nu$. Initialize weights $W_{a,0} = 1/N$ and use $p_{a,t} = W_{a,t-1} / \sum_b W_{b,t-1}$. Choose production $A_t \sim p_t$ and a uniform shadow $J_t \ne A_t$ before observing the current tag. Set

$$z_{a,t} = d\mathbf{1}\{J_t = a\}\frac{w(Z_t)}{A}(\widetilde{Y}_{J_t,t} - \widetilde{Y}_{A_t,t}), \quad W_{a,t} = W_{a,t-1}e^{-\eta z_{a,t} - \eta^2 z_{a,t}^2}, \quad \eta = \frac{A\varepsilon}{16 d\lambda}. \tag{70}$$

The $\widetilde{Y}$ symbols denote returned scores under the resampling channel. Conditional marginal preservation and the action-independent current tag law imply

$$\mathbb{E}[z_{a,t} \mid \mathcal{H}_{t-1}] = \mu_a - \sum_b p_{b,t}\mu_b.$$

Also $|z_{a,t}| \le d/A$, so $|\eta z_{a,t}| \le \varepsilon/(16\lambda) \le 1/16$. The standard fixed-rate SODA potential argument, using $e^{-x-x^2} \le 1-x$ on this range and $\sum_a p_{a,t}\mathbb{E}[z_{a,t} \mid \mathcal{H}_{t-1}] = 0$, gives for a witnessing optimum

$$R_m \le \frac{\log N}{\eta} + \eta \sum_{t \le m} \mathbb{E} z_{a_*,t}^2. \tag{71}$$

For completeness, the total weight is a supermartingale starting at one. Its comparator weight gives $\mathbb{E}\log \sum_a W_{a,m} \ge -\log N + \eta R_m - \eta^2 \sum_t \mathbb{E} z_{a_*,t}^2$, whereas Jensen bounds the left side by zero, proving (71).

The conditional squared difference is bounded by $v\Delta_a^{2-\gamma} + \nu + Z$. Since $w^2 \le w$ and $Zw^2 \le \lambda w$,

$$\frac{\mathbb{E}[w(Z)^2 D_{aa_*}^2]}{A^2} \le \frac{v\Delta_a^{2-\gamma} + \nu + \lambda}{A} \le \frac{v\Delta_a^{2-\gamma} + 2\lambda}{A}.$$

Consequently, Jensen's inequality for the concave power $2-\gamma \in [0,1]$ gives

$$\sum_{t\le m} \mathbb{E} z_{a_*,t}^2 \le \frac{d}{A}\{v m^{\gamma-1} R_m^{2-\gamma} + 2\lambda m\}. \tag{72}$$

At $\gamma = 2$, use the convention $R_m^0 = 1$ in this upper bound. Combining the last two displays with $m\varepsilon^2 \mathsf{H}(\lambda) = k_N \ge d \log N$ and $\lambda \ge v\varepsilon^{2-\gamma}$ yields

$$R_m \le 16 m\varepsilon + \frac{(m\varepsilon)^{\gamma-1} R_m^{2-\gamma}}{16} + \frac{m\varepsilon}{8}.$$

For $x = R_m/(m\varepsilon)$, this reads $x \le 16 + 1/8 + x^{2-\gamma}/16$. If $x \ge 1$, use $x^{2-\gamma} \le x$ to obtain $x \le 17.2$; otherwise $x < 1$. Thus $R_m \le 18 m\varepsilon$.

If the structural expression is clipped, $S(\varepsilon) = 1$, then $\varepsilon^2 = k_N/[m\mathsf{H}(1)] \ge k_N/m$. Since $m \le T$ and $k_N \ge N$, this gives $T\varepsilon \ge \sqrt{NT}$. The production-only MOSS branch therefore supplies the claimed minimum. The clipped regime is therefore covered by the marginal branch.

For $B \ge 2$, choose $m = \lfloor B \rfloor$ and use uniform-past continuation, giving $R_T = (T/m) R_m \le 18 T e(k_N/m)$. By (69) and $m \ge B/2$, $e(k_N/m) \le 2e(k_N/B)$. A priori selection among this policy, MOSS, and the trivial bound is allowed because the channel and class parameters are supplied. For $B < 2$, $e(k_N/B) = 1$ and the marginal branch suffices. This proves (11) with a universal constant and the pathwise shadow cap.

Full-class classification at $\gamma = 1$. For fixed N, v, zero floor, and full budget, the two inverse arguments N/T and $N\log(2N)/T$ differ only by a fixed factor. Theorem 1 and (69) therefore determine the minimax rate up to constants. Write $\varepsilon_T = e(N/T)$; it is interior and untruncated for all sufficiently large T. Then

$$T\varepsilon_T = \frac{N}{\varepsilon_T \mathsf{H}(v\varepsilon_T)}, \qquad \frac{T\varepsilon_T}{\sqrt{T}} = \sqrt{\frac{N}{\mathsf{H}(v\varepsilon_T)}}.$$

Dominated convergence gives $\varepsilon_T \mathsf{H}(v\varepsilon_T) \to \mathbb{P}(Z=0)/v$, and monotone convergence gives $\mathsf{H}(v\varepsilon_T) \to \mathbb{E}[1/Z]$. The marginal cap changes this scale by at most a fixed factor. Thus a clean atom is equivalent to $\Theta(1)$ regret; a finite inverse moment is equivalent to $\Theta(\sqrt{T})$; and no clean atom with an infinite inverse moment gives both $\omega(1)$ and $o(\sqrt{T})$. These are statements about the entire $\mathcal{V}_1(v,0)$ class for its known channel and parameters.

D.3 Weighted comparison confidence intervals

The calibrated and ordered policies use the original PCE phase schedule, active sets, complete-cycle checks, inconsistent-interval fallback, and marginal intervals. Only the paired confidence bank changes. It suffices to use phase-local paired data. At every phase start, freeze the active anchor as in PCE; scores selected by any predictable orientation have the same arm-labeled pair law.

Lemma 10 (A weighted second-moment interval)**.** *Let a fixed pair stream have differences $D_i \in [-1,1]$ and fresh tags C_i, with $\mathbb{E}[D_i \mid C_i, \mathcal{H}_{i-1}] = d$. For a fixed measurable $w : \mathcal{C} \to [0,1]$, put*

$$W_n = \sum_{i\le n} w(C_i), \qquad U_n = \sum_{i\le n} w(C_i) D_i, \qquad V_n = \sum_{i\le n} w(C_i)^2 D_i^2.$$

With the finite grid $\eta_k = 2^{-k-1}$, $0 \le k \le K$, a simultaneous event of probability at least $1 - 2(K+1)e^{-L}$ satisfies

$$|U_n - dW_n| \le 3\sqrt{LV_n} + 4L \qquad \text{for every } n \le T. \tag{73}$$

For $W_n > 0$, dividing this bound by W_n gives a confidence interval centered at U_n/W_n, intersected with $[-1,1]$. For $W_n = 0$, use $[-1,1]$.

Proof. For $|x| \le 1/2$, $e^{x-x^2} \le 1 + x$. Therefore, for either sign of η with $|\eta| \le 1/2$,

$$\mathbb{E}\Big[e^{\eta w(C_i)(D_i - d) - \eta^2 w(C_i)^2 D_i^2} \mid C_i, \mathcal{H}_{i-1}\Big] \le e^{-\eta w(C_i) d}\{1 + \eta w(C_i) d\} \le 1.$$

Taking a conditional expectation over the current tag shows that $\exp\{\eta(U_n - dW_n) - \eta^2 V_n\}$ is a nonnegative supermartingale. Ville's inequality and a union bound over signs and grid points yield $|U_n - dW_n| \le L/\eta + \eta V_n$ for every positive grid point, simultaneously in n. If $V_n \le 4L$, choose $\eta = 1/2$ to get $4L$. Otherwise choose a grid point between $\frac{1}{2}\sqrt{L/V_n}$ and $\sqrt{L/V_n}$, giving $3\sqrt{LV_n}$. Since $V_n \le T$ and $K = \lceil \log_2(2T) \rceil$, this grid point exists. These estimates imply (73). □

The calibrated confidence bank. For $0 \le j \le K$, set $\lambda_j = 2^{-j}$ and $w_j(c) = \lambda_j/(\lambda_j + \zeta(c))$. Compute the preceding interval separately for each j and intersect these intervals with the phase marginal interval. PCE's midpoint and width tests use the resulting intersection, with its existing fallback if inconsistent. This specifies the policy without the channel law m_C, structural parameters γ, v, ν, or an optimal action. Its input includes the numerical function $\zeta(c)$, which is evaluated only after the current tag arrives.

Adaptive pair streams. At each phase start, assign every possible pair an independent infinite tape of marks (C_i, D_i) with its required law. A predictable query reveals the next entry. Because only one pair is queried per request and current scores are independent of the past, this preserves the transcript law. Concentration at every deterministic tape count covers adaptive interruption; union bounds require no independence between marginal and paired streams.

The ordered confidence bank. When the numerical calibration is unknown but nondecreasing in a real tag C, compute the interval of Lemma 10 for every observed lower-threshold subset $\{i \le n : C_i \le c\}$ of each phase-local pair stream, using indicator weights. Use complete groups at tied tag values and intersect all intervals with the marginal interval. There are at most n nonempty distinct sets at count n.

Fix deterministic n and condition on its tags. The differences are conditionally independent, bounded, and mean d; their order defines at most n subsets. Apply the exponential bound to each subset and test rate, integrate out tags, and union-bound over $n \le T$, pairs, and phase starts. Adaptive stopping selects a member of this event, without conditioning on an unextended stopped history. The stated L_{sp} also covers the calibrated bank, quota events, and marginal events, with total failure probability at most δ.

D.4 Spectrum quotas and the adaptive upper bounds

On the joint confidence event for all intersected intervals, a witnessing optimum survives and every phase-r active gap is at most $8\varepsilon_r$. Comparing both active endpoints to that optimum gives

$$\mathbb{E}[D_{ab}^2 \mid C = c] \le C_0\{S(\varepsilon_r) + \zeta(c)\} \tag{74}$$

for a numerical C_0. When the unclipped structural expression is below one, use the existing bound $q_{ab}^P \le 32v\varepsilon_r^{2-\gamma} + 4\nu$ and the independent squared-difference bound one. When it is at least one, $D_{ab}^2 \le 1$ gives (74) with the clipped $S = 1$.

Calibrated quota. Write $S = S(\varepsilon)$ and choose an analysis-only grid point $\lambda \in [S, 2S]$, using $\lambda = 1$ at the upper endpoint. The phase grid covers it because $S \ge \varepsilon \ge 2^{-K}$. Its weight satisfies

$$\mathbb{E}w_\lambda \ge S\mathsf{H}(S), \qquad \mathbb{E}[w_\lambda^2 D^2] \le 4C_0 S^2 \mathsf{H}(S). \tag{75}$$

Bounded-variable concentration, uniformly over the finite grid and raw counts, yields

$$W_n \ge \tfrac{1}{2} nS\mathsf{H}(S), \qquad V_n \le 8C_0 nS^2\mathsf{H}(S) + c_0 L_{\mathrm{sp}}$$

when $nS\mathsf{H}(S) \ge c_1 L_{\mathrm{sp}}$. Substituting into (73) bounds the radius by

$$C_1\sqrt{\frac{L_{\mathrm{sp}}}{n\mathsf{H}(S)}} + \frac{C_2 L_{\mathrm{sp}}}{nS\mathsf{H}(S)}.$$

Consequently,

$$n \ge \frac{C_3 L_{\mathrm{sp}}}{\varepsilon^2 \mathsf{H}(S(\varepsilon))} \tag{76}$$

makes the interval sufficiently short for PCE's phase test, after increasing the numerical C_3. Both the range term and count startup are absorbed because $S(\varepsilon) \ge \varepsilon$. This is a raw shadow-call quota, including tags assigned small weights.

Ordered quota. Use $\mathsf{A}(S) = \sup_{0\le z\le 1} G(z)/(S+z)$ from the threshold comparison. The set $\{\zeta(C) \le z\}$ is a lower threshold set in C, with an open or closed endpoint as necessary. For $r = G(z) > 0$, its accepted differences have second moment at most $C_0(S+z)$. Bernoulli thinning and the second-moment interval show that $n \ge CL_{\mathrm{sp}}(S+z)/(r\varepsilon^2)$ raw calls suffice; the startup is absorbed by $S \ge \varepsilon$. For each analysis precision choose one deterministic population threshold within a factor two of the displayed supremum. Quota concentration then requires a finite union over phase precisions and pairs. The confidence event itself already covers every observed threshold.

For $0 < S \le 1$, partition Z into $[0,S], (S,2S], (2S,4S], \ldots$, truncating at one. Each bin contributes at most $2\mathsf{A}(S)$ to the spectrum, whereas restricting to any threshold gives the reverse bound. This proves (13). Since $S \ge \varepsilon \ge 2^{-K}$, the ordered quota is at most $C(K+1)$ times (76). The argument pools arbitrarily many tag values and has no smallest-label-probability requirement.

Completing the regret proof. A phase needs at most $CNL_{\mathrm{sp}}/\mathsf{J}(\varepsilon)$ raw shadows for the calibrated bank, or $CN(K+1)L_{\mathrm{sp}}/\mathsf{J}(\varepsilon)$ for the ordered bank. The original marginal cycle quota remains $CNL_{\mathrm{sp}}\varepsilon^{-2}$. Optimum survival and active production gaps are unchanged because all intervals estimate the same differences.

By (69), successive shadow quotas at least double, while $\varepsilon_r/\mathsf{J}(\varepsilon_r)$ is nondecreasing. Total shadow cost is therefore a constant multiple of the final quota, and production cost at most K times the final phase cost. Choosing $\mathsf{J}(\varepsilon) \ge CN\Lambda_{\mathrm{sp}}/\overline{B}$ (or $CN(K+1)\Lambda_{\mathrm{sp}}/\overline{B}$ for ordered PCE) bounds prefix and remaining regret by $CT\varepsilon$. Dyadic rounding costs a constant. Interrupted cycles retain the active-gap bound; singleton continuation gives $C\sqrt{NTL_{\mathrm{sp}}}$ on the same event. Failure contributes δT. Since $\mathsf{J}(\varepsilon) \le \varepsilon$, the phase grid suffices whenever the bound is nontrivial. Charging each raw shadow establishes the pathwise cap and the parameter-adaptive bounds.

D.5 The complete-transcript spectrum lower bound

First let $v \ge 8$, $d = N-1$, and $0 < \Delta \le 1/64$. Use the reference and alternatives of Appendix C.8, with $s = \min\{1/4, S(\Delta)\}$. Their base laws lie in $\mathcal{V}_\gamma(v,\nu)$ and all nonzero optimum gaps are at least Δ. Attach the same fixed provenance channel to every environment. The reference therefore satisfies the class's expected shadow budget.

For a pair containing alternative coordinate j, the four-cell calculation in Appendix D.13, conditional on $C = c$, gives

$$D_{\mathrm{KL}}(Q_0^c \| Q_j^c) \le \frac{256\Delta^2}{s+\zeta(c)}.$$

Pairs excluding j are unchanged. Since the tag law does not depend on the base environment, KL disintegration, also for continuous tags, gives

$$D_{\mathrm{KL}}(\text{labeled pair}_0 \| \text{labeled pair}_j) = \mathbb{E}_m D_{\mathrm{KL}}(Q_0^C \| Q_j^C) \le 256\Delta^2 \mathsf{H}(s).$$

Singleton KL is at most $16\Delta^2$. The adaptive chain rule and the fact that one singleton contains at most one alternative coordinate and a pair at most two imply (14).

Choose $\kappa = 2^{-18}$ and $\Delta = \kappa \min\{1, \sqrt{d/T}, e(d/\overline{B})\}$. The singleton information is at most $16\kappa^2$. If $s = 1/4$, then $\mathsf{H}(s) \le 4$ and $B \le T$ make the pair contribution at most $2048\kappa^2$. Otherwise $s = S(\Delta)$, and (69) gives $\frac{B\Delta^2\mathsf{H}(s)}{d} \le \frac{B}{d}\kappa^\gamma \mathsf{J}(e(d/\overline{B})) \le \kappa^\gamma B/\overline{B} \le \kappa$. Thus the average transcript KL is below $1/8$. For $N \ge 3$, Pinsker and $\sum_{j\ge 2} \mathbb{E}_0 N_j^{\mathrm{prod}} \le T$ give an alternative with regret at least $\Delta T/4$. For $N = 2$, the reference and alternative have regret sum at least $3\Delta T/4$. Replacing d by N costs a numerical factor by (69).

For $1 \le v \le 8$, apply the entire-class scaling $Y \mapsto Y/8$ from Appendix C.8.5. It commutes with every conditional pair kernel and leaves $m, \zeta(C)$ and all tags unchanged. The source and target policies have identical shadow counts, while source regret is eight times target regret. The clipped source $S_8(\varepsilon)$ is at least $S_v(\varepsilon)$, so its spectrum resolution is at least as large. The source lower rate therefore dominates the desired target rate up to the numerical factor eight. This proves (10) uniformly for $v \ge 1$.

D.6 Near-clean tails and the regret phase diagram

For any $Z \in [0,1]$, integration by parts gives

$$\mathsf{H}(s) = \frac{1}{s+1} + \int_0^1 \frac{G(z)}{(s+z)^2}\, dz. \tag{77}$$

Suppose $c_- z^\alpha \le G(z) \le c_+ z^\alpha$ for $0 < z \le z_0$, with $0 < \alpha < 1$ and positive fixed constants. Restricting to $Z \le s$ gives the lower bound $G(s)/(2s)$. For the upper bound in (77), split at z_0 and substitute $z = su$. Since $\int_0^\infty u^\alpha/(1+u)^2\, du < \infty$, this proves $\mathsf{H}(s) \asymp s^{\alpha-1}$. If $G(z) \sim cz^\alpha$, the constant is $c\alpha\pi/\sin(\pi\alpha)$.

On the unclipped range, the reciprocal of $\mathsf{J}(\varepsilon)$ is consequently comparable to

$$\frac{(v\varepsilon^{2-\gamma} + \nu)^{1-\alpha}}{\varepsilon^2} \asymp v^{1-\alpha}\varepsilon^{-p_\alpha} + \nu^{1-\alpha}\varepsilon^{-2}, \qquad p_\alpha = \gamma + (2-\gamma)\alpha.$$

The constants may depend on the tail parameters. When clipping occurs, the paired branch is within a fixed factor of the marginal branch because $B \le T$. The log-free spectrum rate is therefore equivalent to

$$\min\left\{T, \sqrt{NT}, T\left[\left(\frac{v^{1-\alpha}N}{B}\right)^{1/p_\alpha} + \sqrt{\frac{\nu^{1-\alpha}N}{B}}\right]\right\}. \tag{78}$$

The oracle factor $\log(2N)$ is constant for fixed N, and the parameter-adaptive confidence logarithms are subpolynomial for $\delta = (T+1)^{-2}$. Setting $\nu = 0$ and $B = T^{\beta+o(1)}$ proves Corollary 1. If instead $\nu_T = T^{-\kappa+o(1)}$, the exponent is

$$\min\left\{\tfrac{1}{2}, \max\left\{1 - \frac{\beta}{p_\alpha}, 1 - \frac{\beta + \kappa(1-\alpha)}{2}\right\}\right\}.$$

Polynomial improvement requires both $\beta > p_\alpha/2$ and $\beta + \kappa(1-\alpha) > 1$.

At $\alpha = 1$, (77) gives $\mathsf{H}(s) \asymp \log(1/s)$. For fixed N, v, zero floor, and fixed $\gamma < 2$, both oracle bounds give exact full-budget order $\sqrt{T/\log T}$, since $\log(1/S(\varepsilon)) \sim (2-\gamma)\log(1/\varepsilon)$. The parameter-adaptive PCE policies retain their displayed logarithmic overhead. For $\alpha > 1$, $\mathbb{E}[1/Z] < \infty$ and $\mathsf{H}(s)$ approaches that finite number. At $\gamma = 2$, $S(\varepsilon) = 1$ for all positive ε, so the marginal rate determines the order. An atom of mass $r_0 > 0$ at zero gives $\mathsf{H}(s) \sim r_0/s$, recovering the clean branch with effective budget $r_0 B$.

Matched erased experiments. For (16), the conditional mean of the lower-half calibration is $\alpha/[4(\alpha+1)]$. The upper-half value is its complement, proving $\mathbb{E}Z = 1/2$. Direct inversion of the lower branch gives $G(z) = \frac{1}{2}(4z)^\alpha$ on $(0, 1/4]$ and no atom at zero. The tag law is uniformly distributed for every α, and the erased kernel is always $(P_{ab} + P_a \otimes P_b)/2$.

An atom-free slowly varying example sets $\zeta(c) = \frac{1}{4}\exp(1 - 1/(2c))$ on $0 < c \le 1/2$ and sets it to $1 - m_{\text{low}}$ on the upper half, where m_{low} is its lower-half conditional mean. Its overall mean remains $1/2$, and

$$G(z) = \frac{1}{2\{1 + \log((1/4)/z)\}}, \qquad 0 < z \le 1/4.$$

Equation (77), or splitting its integral at $z = s$, gives $\mathsf{H}(s) \asymp 1/[s\log(1/s)]$. At $\gamma = 1, \nu = 0, B = T$ and fixed N, v, the known-channel, known-class minimax rate is exactly $\Theta(\log T)$; the parameter-adaptive policies have polylogarithmic upper bounds. Thus an exactly clean atom is unnecessary even for that regime.

D.7 An exact spectrum frontier on the noisy two-action family

Use the family $P_{\sigma,g}$ of Proposition 1, with $0 < g \le 1/8$, and any fixed provenance channel. Put

$$\mathsf{J}_{\text{nat}}(x) = x^2\mathsf{H}(x), \qquad e_{\text{nat}}(u) = \sup\{x \in [0,1] : \mathsf{J}_{\text{nat}}(x) \le u\},$$

where $\mathsf{J}_{\text{nat}}(0) = 0$. Its scaling satisfies $c^2\mathsf{J}_{\text{nat}}(x) \le \mathsf{J}_{\text{nat}}(cx) \le c\mathsf{J}_{\text{nat}}(x)$ for $0 < c \le 1$.

Proposition 3 (Exact natural-family spectrum). *For every fixed channel, every $T \geq 1$, and $0 \leq B \leq T$, the natural-family minimax regret is*

$$\Theta\Big(T \min\{T^{-1/2}, e_{\mathrm{nat}}(1/\overline{B})\}\Big), \tag{79}$$

with universal constants. A policy supplied with the full channel, including the distribution G of Z, attains the upper bound with at most $\lfloor B \rfloor$ shadows, without knowing g or σ.

Proof. Set $a_g = 3/8 - 3g/2 \in [3/16, 3/8]$. Conditional on posterior contamination z, the two off-diagonal cell probabilities are $(3g + a_g z)/2$ and $(g + a_g z)/2$ under the positive ordering, and are exchanged under the negative ordering. Diagonal cells agree. Thus the complete labeled-pair KL is exactly

$$d(g) = g\mathbb{E} \log \frac{3g + a_g Z}{g + a_g Z} = \frac{g^2}{a_g} \int_1^3 \mathsf{H}(gu/a_g)\, du. \tag{80}$$

For $t \geq 1$, $\mathsf{H}(g)/t \leq \mathsf{H}(tg) \leq \mathsf{H}(g)$. Since $u/a_g \geq 1$, this yields the uniform bounds

$$(\log 3) g^2 \mathsf{H}(g) \leq d(g) \leq \frac{32}{3} g^2 \mathsf{H}(g). \tag{81}$$

Singleton KL is at most $6g^2$. The complete transcript of any policy with the class-wide expected budget therefore has KL at most $6Tg^2 + (32/3)B\mathsf{J}_{\mathrm{nat}}(g)$. Choose $g = \kappa \min\{T^{-1/2}, e_{\mathrm{nat}}(1/\overline{B})\}$ with a sufficiently small universal $\kappa \leq 1/8$. The two terms are at most $6\kappa^2$ and $(32/3)\kappa B/\overline{B}$, respectively, including when the inverse is clipped at one. The two-ordering Pinsker argument proves the lower bound.

For $m \geq 1$ paired calls, set $\varepsilon_m = e_{\mathrm{nat}}(1/m)$ and use $w(z) = \varepsilon_m/(\varepsilon_m + z)$. In this two-action construction define the oriented contrast $D_i = Y_{2,i} - Y_{1,i}$, so a positive mean favors action 1. Define

$$A = \mathbb{E}w(Z) = \varepsilon_m \mathsf{H}(\varepsilon_m), \qquad X_i = w(Z_i) D_i.$$

Serve action 1 when the past sum of X_i is positive and action 2 when it is negative, with fair ties. The policy uses the full channel to set ε_m, and never uses the unknown gap or ordering. Under the positive ordering, $\mathbb{E}X_i = Ag$. The exact conditional second moment is $\mathbb{E}[D_i^2 \mid Z_i = z] = 2g + a_g z$, so $\mathbb{E}X_i^2 \leq 2gA + \frac{3}{8}\varepsilon_m A$, using $w^2 \leq w$ and $zw^2 \leq \varepsilon_m w$. The centered increment is bounded in absolute value by two. Bernstein therefore gives $\mathbb{P}(\text{wrong sign after } n) \leq \exp\{-cnAg^2/(g + \varepsilon_m)\}$. Summing this geometric bound, and also using the trivial regret bound, yields

$$R_m \leq \min\{mg,\ g + C/A + C\varepsilon_m/(Ag)\}. \tag{82}$$

If $g \leq \varepsilon_m$, use $mg \leq m\varepsilon_m$. If $g > \varepsilon_m$, then $\varepsilon_m < 1/8$, so the inverse is interior and $m\varepsilon_m^2 \mathsf{H}(\varepsilon_m) = m\varepsilon_m A = 1$. The second bound is therefore at most $Cm\varepsilon_m$, since $1/A = m\varepsilon_m$ and $g \leq 1 \leq 1/A$. If the inverse saturates at one for a small m, only the first case is needed. This proves $R_m \leq Cme_{\mathrm{nat}}(1/m)$ uniformly in g and in the channel.

For $B \geq 2$, run this policy for $m = \lfloor B \rfloor$ paired calls and apply the exact uniform-past continuation of Appendix C.9. Its regret is at most $CTe_{\mathrm{nat}}(1/m)$. Since $m \geq B/2$, inverse scaling gives $e_{\mathrm{nat}}(1/m) \leq 2e_{\mathrm{nat}}(1/B)$. Production-only MOSS supplies $O(\sqrt{T})$, and covers $B < 2$. From the known channel and T, B, select the smaller uniform rate before play. This establishes (79) with universal constants. □

A complete full-budget classification. For all sufficiently large T, $\varepsilon_T = e_{\mathrm{nat}}(1/T)$ is interior and decreases to zero. Since $\mathsf{H}(\varepsilon_T) \geq 1/(1 + \varepsilon_T) \geq 1/2$, the marginal cap in (79) changes $T\varepsilon_T$ by at most a factor $\sqrt{2}$. The inverse identity gives

$$T\varepsilon_T = \frac{1}{\varepsilon_T \mathsf{H}(\varepsilon_T)}, \qquad \frac{T\varepsilon_T}{\sqrt{T}} = \frac{1}{\sqrt{\mathsf{H}(\varepsilon_T)}}. \tag{83}$$

Dominated and monotone convergence give $\varepsilon \mathsf{H}(\varepsilon) \to \mathbb{P}(Z = 0)$ and $\mathsf{H}(\varepsilon) \to \mathbb{E}[1/Z]$, respectively. Thus (83) gives the same exhaustive constant, square-root, and intermediate classification as the full-class argument in Appendix D.2. For $G(z) \asymp z^\alpha$, $0 < \alpha < 1$, the exact rate is $\Theta(\min\{\sqrt{T}, T/\overline{B}^{1/(1+\alpha)}\})$. At $\alpha = 1$ it is $\Theta(\sqrt{T/\log T})$ at full budget; the slowly varying example in Appendix D.6 gives $\Theta(\log T)$. These exact rates concern the known-channel oracle; adaptive PCE retains its displayed logarithms.

D.8 Unknown partitions of a growing tag alphabet

Proposition 4 (Unknown label meanings)**.** *For every integer $T \geq 1$ and even $M \geq 16T^2$, there is a family of channels on $[M]$ such that each fixed calibrated channel permits regret at most $2/(2-\sqrt{3})$ on the two-action family at full budget, whereas any single policy required to handle every unknown channel has worst-case regret at least $5\sqrt{T}/256$.*

Proof. Choose an unknown subset $\mathcal{U} \subset [M]$ of size $M/2$. Draw tags uniformly and set $F = 0$ precisely when $C \in \mathcal{U}$. Each channel has contamination $1/2$ and an exactly clean stratum of mass $1/2$. If $\mathcal{U}$ is supplied, the clean-stratum empirical-orientation proof in Appendix D.14, with acceptance probability $1/2$, gives the constant bound.

For the unknown-channel lower bound, put a uniform prior on $\mathcal{U}$, independently of the action ordering, and grant both scores and a tag on every round. Compare this experiment to independent uniform tags and independent $F \sim \mathrm{Bernoulli}(1/2)$. The probability of any repeated tag is at most $T(T-1)/(2M)$. At the ith fresh label, after $k = i-1$ membership bits have been exposed, its conditional probability of corruption is $(M/2 - n_1)/(M-k)$, differing from $1/2$ by at most $k/[2(M-k)]$. Sequential maximal coupling adds failure probability at most

$$\sum_{k=0}^{T-1} \frac{k}{2(M-k)} \leq \frac{T(T-1)}{4(M-T+1)} \leq \frac{T(T-1)}{2M}.$$

Thus the joint laws of all tags and latent corruption bits differ in total variation by at most $T(T-1)/M$. Coupling conditional scores identically and applying data processing gives the same bound for every policy's transcript. Latent memberships are exposed only for constructing this coupling, never to the learner.

Let $g = 1/(16\sqrt{T})$. In the comparison experiment, the hidden-contamination pair has KL at most $32g^2$: only the two off-diagonal cells differ, each by g, and both cell probabilities are at least $1/16$. The full KL is at most $1/8$, so the two-ordering Bayes regret is at least $3gT/8$. Since each policy's regret lies in $[0, gT]$ and $T(T-1)/M \leq 1/16$, transfer to the unknown-partition experiment loses at most $gT/16$. Its Bayes regret is at least $5gT/16 = 5\sqrt{T}/256$, implying the stated worst-case bound for a fixed partition and ordering. □

The lower bound concerns one policy uniform over unknown partitions of an even alphabet with $M \geq 16T^2$. For each separately supplied fixed meaning map, the calibrated guarantee remains constant. Fixed alphabets allow repeated observations of each label and form a separate statistical regime. An ordering supplies reusable semantics across distinct observed labels.

D.9 Identification and valid certificates

Consider the unique-optimum subclass with every nonzero gap at least $0 < \Delta \leq 1/512$, a certification horizon T_{cert}, and shadow budget $0 \leq B \leq T_{\mathrm{cert}}$. Let $0 < \delta_0 \leq 1/10$, $S_\Delta = S(\Delta)$, and let L_{id} denote L_{sp} evaluated at $T = T_{\mathrm{cert}}$ and $\delta = \delta_0$. Run the confidence-elimination policy through T_{cert} and issue a certificate only when its active set is a singleton.

Proposition 5 (Spectrum certification boundary)**.** *With probability at least $1-\delta_0$, every issued certificate at any stopping time up to T_{cert} is correct. For the calibrated policy, certification by T_{cert} also holds with probability at least $1-\delta_0$ if either*

$$T_{\mathrm{cert}} \geq \frac{CNL_{\mathrm{id}}}{\Delta^2}, \qquad \text{or} \qquad B \geq \frac{CNL_{\mathrm{id}}}{\Delta^2 \mathsf{H}(S_\Delta)}, \quad T_{\mathrm{cert}} \geq B, \tag{84}$$

Ordered tags add a factor $1 + \log(1/\Delta)$ to the paired quota. Conversely, any policy with expected shadow budget at most B uniformly over this class, incorrect-certificate probability at most δ_0, and noncompletion probability by T_{cert} at most δ_0, satisfies

$$T_{\mathrm{cert}} + B\mathsf{H}(S_\Delta) \geq \frac{cN\log(1/\delta_0)}{\Delta^2}. \tag{85}$$

Proof. On the simultaneous confidence event the optimum is never eliminated; any singleton active set certifies it, also at data-dependent stopping times. At a precision below a numerical multiple of Δ, the active-gap invariant forces a singleton. Marginal sample quotas sum geometrically. Weighted quotas also sum geometrically by (69), proving (84); the factor K needed to sum production regret is unnecessary for the total sample quota. Since $\mathsf{H}(S_\Delta) \le 1/S_\Delta \le 1/\Delta$, a nonvacuous sufficient paired quota requires $\Delta \ge CNL_{\mathrm{id}}/T_{\mathrm{cert}}$. Adjusting C ensures the fixed-horizon grids reach the certifying precision; the algorithm does not take Δ as an input. The marginal condition supplies the same coverage. The ordered claim follows from (13).

For necessity, use the in-class hard reference and alternatives with gaps Δ and 2Δ. Under the reference, certification of action 1 occurs with probability at least $1 - 2\delta_0$, and under each alternative with probability at most δ_0. Binary data processing gives KL at least $\mathrm{kl}(1 - 2\delta_0, \delta_0) \ge c_0 \log(1/\delta_0)$. Average and apply (14) with $T = T_{\mathrm{cert}}$. If $s = S_\Delta$, this gives the stated result directly. If $s = 1/4$, then $\mathsf{H}(s) \le 4$ and $B \le T_{\mathrm{cert}}$ imply $T_{\mathrm{cert}} + B\mathsf{H}(s) \le 5T_{\mathrm{cert}} \le 5\{T_{\mathrm{cert}} + B\mathsf{H}(S_\Delta)\}$. This proves necessity for $v \ge 8$. For $v < 8$, use source gap 8Δ in the entire-experiment scaling. The source parameter $\min\{1, 8(8\Delta)^{2-\gamma} + \nu\}$ and S_Δ differ by at most a numerical factor, and the gap changes by eight. The restriction on Δ ensures validity of the source family; all information constants remain numerical. □

A sufficiently large sum $T_{\mathrm{cert}} + B\mathsf{H}(S_\Delta)$ implies one of the two sufficient conditions within a factor two. Thus (85) and (84) characterize the resource boundary up to logarithms. When the observations are insufficient for completion, the policy returns the remaining active set without a certificate.

For validity over an unbounded horizon, allocate summable confidence weights over pairs, phase indices, regularization indices, and test-rate indices, proportional to the inverse squares of one plus each index. Ville's inequality is uniform in each raw stream count. For the ordered bank also allocate inverse-square weights over deterministic counts and union-bound over that count's tag prefixes. Every issued certificate is then correct on a single event of probability at least $1 - \delta_0$, with only additional logarithmic terms in the finite-time quotas.

D.10 Ordered confidence banks and policy correspondence

This subsection gives two ordered confidence banks realizing the ordered-confidence argument. Each phase-local pair record contains (C_i, D_i), with canonical arm-labeled difference D_i. The anchor is fixed within a phase and the record resets at phase start. Historical confidence reuse is optional for these quotas.

Exact lower-threshold banks with doubling checks. Maintain the observations in a balanced tree keyed by C, grouping equal tags and storing each group's count, difference sum, and square sum. An insertion takes $O(\log(n+1))$ work and the record uses $O(n)$ space. At raw pair counts $n = 1, 2, 4, \ldots$, scan groups in increasing tag order and compute every lower-prefix interval. A prefix with count k, sum U, and square sum V has interval

$$\left[\frac{U}{k} \pm \frac{3\sqrt{L_{\mathrm{sp}} V} + 4L_{\mathrm{sp}}}{k}\right] \cap [-1, 1]. \tag{86}$$

Intersect these intervals, and cache the result until the next refresh. Marginal intervals continue to update. The scan takes $O(n)$ work; geometric refresh counts make total scanning work $O(n)$. Total insertion and scanning work is $O(n \log(n+1))$. Sorting the full record at each doubling gives the same total order and is another exact implementation. These are aggregate costs; a refresh request may require a linear scan.

Cached intervals cover the same fixed means; elimination uses complete cycles. Before interruption, each anchor–target pair receives one sample per cycle, so required refreshes are available at cycle ends. The next dyadic count above n_* is below $2n_*$, increasing the phase quota only by a constant. Budget/horizon interruption uses the original active-gap and marginal bounds.

A smaller growing bank using only tag order. At a refresh count n, sort the tags and consider ranks $k \in \{1, 2, 4, \ldots, 2^{\lfloor \log_2 n \rfloor}\}$. For each rank, include the prefix ending at the last observation tied with that rank's tag. Include the whole record as well, and remove duplicate prefixes. This gives at

most $2 + \lfloor \log_2 n \rfloor$ intervals. The bank uses neither numerical calibration nor the tag distribution. It approximates the full lower-threshold bank while preserving its quota up to constants.

To verify the quota, fix a phase precision ε and one deterministic population cut of mass $r > 0$ whose contamination is at most z. On a fresh pair tape at a deterministic count n, let M be the number accepted by this cut. Bernoulli thinning gives $M \geq nr/2$ once the stated quota is met. Choose $k = 2^{\lfloor \log_2 M \rfloor}$. Closing its tie group gives a candidate prefix of size k' with $M/2 \leq k' \leq M, \qquad k' \geq nr/4$. The upper bound follows because a population threshold includes complete tie groups. Every observation in this candidate still has $\zeta(C_i) \leq z$. Conditional on the first n tape tags, its indices depend only on the tags; its differences have mean d_{ab} and second moments at most $C_0(S(\varepsilon) + z)$. Conditional bounded-variable concentration yields $V \leq 2C_0 k'(S(\varepsilon) + z) + cL_{\mathrm{sp}}$ except on the allocated failure event. Equation (86) therefore has radius at most

$$C\sqrt{\frac{L_{\mathrm{sp}}(S(\varepsilon) + z)}{nr}} + \frac{CL_{\mathrm{sp}}}{nr}.$$

Thus $n \geq CL_{\mathrm{sp}}(S(\varepsilon) + z)/(r\varepsilon^2)$ suffices; $S(\varepsilon) \geq \varepsilon$ absorbs the range term and thinning startup. For each precision choose one deterministic population cut within a factor two of $\mathsf{A}(S)$ from Appendix D.4. This is a finite union over precisions and pairs. Coverage of the data-selected rank prefixes follows from the existing uniform event over deterministic tape counts and all lower prefixes. Consequently the empirical-rank bank and doubling schedule preserve the ordered bound with a different numerical constant. Conditioning on a fixed tape prefix, followed by a union over counts, also covers adaptive queries and phase starts.

Finite numerical grids. A fixed collection of tag cutoffs has the best quota among those cutoffs. Monotone reparameterization can place the useful near-clean mass outside a fixed grid's resolution. If the continuous tag CDF F_C is supplied, the transformed tag $U = F_C(C)$ is uniform, and the grid $U \leq 2^{-j}$, $0 \leq j \leq \lceil \log_2 T \rceil$, supplies a horizon-dependent alternative. Every population cut with mass $r \geq q_{\min} = 2^{-\lceil \log_2 T \rceil}$ contains a grid prefix with mass in $[r/2, r]$ and no worse quality. Its quota loses at most a factor two. A smaller cut has $r/(S + z) \leq q_{\min}/S$. Every cut whose quota can complete within T satisfies

$$r \geq \frac{CL_{\mathrm{sp}}(S + z)}{T\varepsilon^2} \geq \frac{CL_{\mathrm{sp}}}{T},$$

so the grid covers all such useful cuts. At atoms, independent uniform randomization within an equal-tag group implements the quantile split and preserves its conditional target. The CDF supplies the quantile map, whereas the empirical-rank construction uses tag order directly. The eleven-cut experiment therefore has the finite-grid guarantee above.

Calibration error and target validity. Under $\mathbb{E}[D \mid C, \text{past}] = d$, any fixed bounded nonnegative tag weight preserves coverage, although misspecification changes efficiency. Oracle SODA also needs the true mean weight for normalization. Learned weights must be predictable before their scores and retained thereafter; retrospective fitting requires a separate argument. Score-dependent tags can change the target.

Two-action ordered-rank service. For the natural family, use the oriented contrast $D_i = Y_{2,i} - Y_{1,i}$ from Appendix D.7; a positive contrast serves action 1. Let $m = \lfloor B \rfloor \geq 1$, $\delta = (T + 1)^{-2}$, and $L_{\mathrm{srv}} = \log\{256(m + 1)^3(T + 1)^2\}$. At paired counts $n = 1, 2, 4, \ldots$, form the empirical-rank intervals with L_{srv} in (86), and intersect them with valid pooled and marginal intervals and $[-1, 1]$. If the intersection excludes zero, lock its certified action ordering. Before certification, serve by the empirical sign of a smallest-radius prefix, with fair zero ties. Its cutoff may be cached and its empirical service sum updated between checkpoints; these updates issue no new certificate. If the intersection is empty, continue with the empirical-sign service rule without certification. Every production choice uses completed past data. After m paired requests, use the same uniform-past continuation as weighted and unweighted EO. With $m = 0$, use production-only MOSS.

The confidence event is valid because, at each deterministic $n \leq m$, there are at most n lower prefixes and $1 + \lceil \log_2(2m) \rceil$ test rates. Unioning their two signs, all counts, and marginal events is covered by L_{srv}. On that event, every locked ordering is correct. At a natural-family gap g, the accepted second moment is at most $2g + (3/8)z$. The preceding quota argument and (13) imply certification within

$$n_g \leq \frac{CL_{\mathrm{srv}}(1 + \log_2(m + 1))}{g^2 \mathsf{H}(g)}$$

when this is affordable and $g \geq 1/m$, including the constant overshoot from doubling checks. Prior to certification each request incurs at most g regret. Smaller gaps use the bound mg directly. Since $g\mathsf{H}(g)$ is increasing, splitting at the analysis-only value $\varepsilon = e_{\mathrm{nat}}(CL_{\mathrm{srv}}(1+\log_2(m+1))/m)$ gives

$$R_m \leq Cm\, e_{\mathrm{nat}}\left(\frac{CL_{\mathrm{srv}}(1+\log_2(m+1))}{m}\right) + \delta m. \tag{87}$$

For $g \leq \varepsilon$ use mg; for $g > \varepsilon$ use $n_g g$ and the inverse identity. Since $\mathsf{J}_{\mathrm{nat}}(x) \leq x$, this second case has $g \geq 1/m$ when the inverse is interior. A saturated inverse makes the bound immediate. Uniform-past continuation multiplies prefix regret by T/m. This is a logarithmic paired-prefix guarantee. The marginal/paired minimum in the general PCE theorem concerns its separate algorithm.

A channel-based oracle branch choice. For weighted EO, choosing its paired-prefix policy when $Te_{\mathrm{nat}}(1/m) \leq \sqrt{T}$ and production-only MOSS otherwise uses only the supplied channel, horizon, and budget. The choice is fixed before observing scores and is shared across all gaps and action orderings for those inputs. It gives the natural-family minimum up to a numerical constant.

D.11 Binary provenance as a specialization

Specialize (7) to hidden Bernoulli contamination F of rate $\bar{\zeta}$, independent of the past and base scores. Production marginals remain unchanged. Erasing F gives

$$P_{ab}^{\bar{\zeta}} = (1-\bar{\zeta})P_{ab} + \bar{\zeta}(P_a \otimes P_b), \qquad q_{ab}^{\bar{\zeta}} = (1-\bar{\zeta})q_{ab} + \bar{\zeta}\{\mathrm{Var}(Y_a) + \mathrm{Var}(Y_b) + d_{ab}^2\}. \tag{88}$$

Let $\mathcal{D}_\gamma(v,\nu,\bar{\zeta})$ denote these experiments with $P \in \mathcal{V}_\gamma(v,\nu)$.

Theorem 5 (Hidden-resampling frontier)**.** *For $\nu_{\bar{\zeta}} = (1-\bar{\zeta})\nu + \bar{\zeta}$, the minimax regret on $\mathcal{D}_\gamma(v,\nu,\bar{\zeta})$ satisfies (39) and (38) with ν replaced by $\nu_{\bar{\zeta}}$, uniformly for $v \geq 1$ and $\bar{\zeta} \in [0,1]$. The upper policy requires neither $\bar{\zeta}$ nor the base structure.*

Thus a fixed positive contamination rate imposes the marginal-learning order when $B \leq T$. We next keep the entire kernel (88) fixed and change the provenance information delivered with each pair.

Labels received after the decision. On a shadow round, after the endpoints have been chosen, the learner receives their scores and a label $C \in \{0,1\}$. Conditional on F, this label is independent of the scores and the past; its channel is fixed across actions, requests, and base environments. Put

$$r_\ell = \mathbb{P}(C=\ell), \qquad \zeta_\ell = \mathbb{P}(F=1 \mid C=\ell), \qquad \nu_\ell = (1-\zeta_\ell)\nu + \zeta_\ell. \tag{89}$$

Labels of probability zero are omitted. We have $\sum_\ell r_\ell = 1$ and $\sum_\ell r_\ell \zeta_\ell = \bar{\zeta}$. The conditional pair law is

$$Q_{ab}^{\ell} = (1-\zeta_\ell)P_{ab} + \zeta_\ell(P_a \otimes P_b). \tag{90}$$

Every conditional one-action marginal equals P_a. Every shadow consumes budget, including calls with unhelpful labels. Write $\mathfrak{R}_{\mathrm{tag}}$ for the minimax regret over base laws in $\mathcal{V}_\gamma(v,\nu)$ with this fixed channel, and define

$$\Psi_{\mathrm{tag}}(u) = \min_{\ell: r_\ell > 0}\left\{\left(\frac{uvN}{r_\ell \overline{B}}\right)^{1/\gamma} + \sqrt{\frac{u\nu_\ell N}{r_\ell \overline{B}}}\right\}. \tag{91}$$

Theorem 6 (Binary provenance frontier)**.** *Let $K = \lceil \log_2(2T) \rceil$, $L_{\mathrm{tag}} = \log(256N^2K(T+1)/\delta)$, and $\Lambda_{\mathrm{tag}} = KL_{\mathrm{tag}}$, with $0 < \delta < 1$. Universal constants $c, C > 0$ satisfy*

$$\mathfrak{R}_{\mathrm{tag}} \geq c\min\{T, \sqrt{NT}, T\Psi_{\mathrm{tag}}(1)\}, \tag{92}$$

$$R_T^{\text{tagged-PCE}} \leq C\min\{T, \sqrt{NTL_{\mathrm{tag}}}, T\Psi_{\mathrm{tag}}(\Lambda_{\mathrm{tag}})\} + \delta T. \tag{93}$$

Both bounds hold uniformly over the stated base class. The lower bound grants the learner the channel parameters and allows an expected shadow budget. Tagged-PCE uses at most $\lfloor B \rfloor$ shadows and requires neither these parameters nor γ, v, ν.

Tagged-PCE intersects label-specific difference intervals with all-label marginal intervals. Label ℓ supplies frequency r_ℓ and floor ν_ℓ; a common event permits adaptation to the better stream while retaining every production observation. Appendix D.12 proves concentration and budget accounting.

The converse uses the in-class family (40). With $d = N - 1$, its complete labeled transcript satisfies

$$\frac{1}{d}\sum_{j=2}^{N} D_{\mathrm{KL}}(\mathbb{P}_0^{\mathcal{A}} \| \mathbb{P}_j^{\mathcal{A}}) \le \frac{16T\Delta^2}{d} + \frac{512B\Delta^2}{d}\sum_{\ell: r_\ell > 0} \frac{r_\ell}{s + \zeta_\ell}. \tag{94}$$

The two information contributions sum to at most twice their maximum. Gap optimization gives (92), including its singleton branch; Appendix D.13 covers every policy and $v \ge 1$.

Equal score laws, accuracy, and mutual information. Consider the two-action family of Proposition 1, with $\bar{\zeta} = 1/2$. Interpret $C = 0$ as a clean flag. A conservative channel has $(\mathbb{P}(C = 0 \mid F = 0), \mathbb{P}(C = 0 \mid F = 1)) = (1/2, 0)$; an aggressive channel has $(1, 1/2)$. Both have accuracy $\mathbb{P}(C = F) = 3/4$. For binary entropy $h_2(p) = -p \log_2 p - (1-p)\log_2(1-p)$, both also satisfy

$$I(F; C) = h_2(1/4) - 1/2 \quad \text{bits}. \tag{95}$$

Their unlabeled score laws, covariances, action marginals, and evaluation costs agree member by member.

Corollary 3 (Equal mutual information, different regret)**.** *Let $\mathcal{F}_{\mathrm{cons}}$ and $\mathcal{F}_{\mathrm{aggr}}$ be these conservative and aggressive tagged families. For all $T \ge 1$,*

$$\mathfrak{R}(T, T, 2; \mathcal{F}_{\mathrm{cons}}) = \Theta(1), \qquad \mathfrak{R}(T, T, 2; \mathcal{F}_{\mathrm{aggr}}) = \Theta(\sqrt{T}), \tag{96}$$

with numerical constants. The policies need neither the gap nor the ordering.

The conservative flag certifies an exactly coupled stratum of mass $1/4$. The aggressive flag certifies an independently resampled stratum of mass $1/4$; its two contamination floors are $1/3$ and 1. This channel still identifies the original joint law at the population level:

$$P_{ab} = \tfrac{3}{2} Q^0_{ab} - \tfrac{1}{2} Q^1_{ab}. \tag{97}$$

The square-root lower bound applies to estimators using this known inverse as well as to arbitrary policies on the labeled scores. Appendix D.14 gives the exact constant-regret upper bound and the common-support converse.

A criterion for useful provenance. Write $\tau = \mathbb{P}(C = 0 \mid F = 0)$ and $\phi = \mathbb{P}(C = 0 \mid F = 1)$. Then $r_0 = (1 - \bar{\zeta})\tau + \bar{\zeta}\phi$ and $\nu_0 = ((1 - \bar{\zeta})\tau\nu + \bar{\zeta}\phi)/r_0$. False-clean errors add a noise floor; false-dirty errors reduce the usable count. At $\gamma = 1$, $\nu = 0$, and $B = T$, with fixed $N, v, \bar{\zeta} \in (0, 1)$, and r_0 bounded below, the clean-label branch has order $1 + \sqrt{T\phi}$ up to logarithms. Thus provenance quality is assessed through the frequency and residual contamination of a usable stratum. The joint score distribution, tag accuracy, and mutual information with the contamination indicator leave this resource unresolved.

D.12 Tagged-PCE upper bound

For Theorem 6, Tagged-PCE keeps all marginal observations and phase-local/historical difference counts, sums, and square sums for each label. It intersects their intervals with the marginal intervals, treating unavailable streams as $\mathbb{R}$. Phase tests and budget accounting are unchanged: endpoints precede the tag, and every pair costs one shadow. The refresh and active-only handoff in Appendix C.3 remain available.

Fix $K = \lceil \log_2(2T) \rceil$, $L_{\mathrm{tag}} = \log(256N^2K(T+1)/\delta)$, and $\Lambda_{\mathrm{tag}} = KL_{\mathrm{tag}}$. Given label ℓ, the pair law has the original marginal means and satisfies

$$q_{aa_*,\ell} \le (1 - \zeta_\ell)v\Delta_a^{2-\gamma} + (1 - \zeta_\ell)\nu + \zeta_\ell \le v\Delta_a^{2-\gamma} + \nu_\ell.$$

Here $q_{ab,\ell} = \mathbb{E}[D_{ab}^2 \mid C = \ell]$ and $\nu_\ell = (1 - \zeta_\ell)\nu + \zeta_\ell$.

Lemma 11 (Coverage and thinning under adaptive pair queries)**.** *With probability at least $1 - \delta$, all marginal intervals and both labels' paired intervals cover their corresponding mean differences, simultaneously across phases, historical streams, and sample counts. Each paired interval also satisfies the empirical-second-moment shrinkage bound. On the same event, for each such pair-query stream, label ℓ appears at least $r_\ell m/2$ times among its first m queries whenever $m \ge 8L_{\mathrm{tag}}/r_\ell$. The latter condition is vacuous when $r_\ell = 0$.*

Proof. At each predictable query time of a fixed pair, the current tag has probability r_ℓ independently of the completed past. Its indicators on this query clock are i.i.d. Bernoulli(r_ℓ). Complete stopped streams after the horizon and apply the fixed-count Chernoff bound $\mathbb{P}\{N_\ell(m) < r_\ell m/2\} \le e^{-r_\ell m/8}$. A union over $m \le T$, phases, pair streams, and the two labels gives the stated thinning property.

Conditional on label ℓ, the next accepted pair difference has a fixed law with mean $\mu_a - \mu_b$. For a selected difference X, an exponential increment calibrated to that conditional law has conditional expectation at most

$$1 - r_\ell + r_\ell \mathbb{E}[\exp\{\lambda(X - \mathbb{E}[X \mid C = \ell]) - \psi_\ell(\lambda)\} \mid C = \ell] \le 1.$$

Thus the original accepted-stream coverage and square-moment proof applies to each label. The tag can affect the conditional score law; its realization follows the predictable pair choice. Completing accepted streams with independent conditional-law draws gives the fixed-count bounds used in Lemma 4. The logarithm above covers their union together with the marginal and thinning failures. Anchor changes preserve the predictability of the query clock, and phase statistics persist across geometric refresh blocks. □

Lemma 12 (Tagged phase quota)**.** *On the event of Lemma 11, an optimal reference survives and the active-gap invariant of Lemma 5 holds. The original marginal phase quota remains valid with L_{tag}. For any $r_\ell > 0$, a fixed-anchor phase of precision ϵ completes within*

$$M_\ell(\epsilon) = \left\lceil \frac{4}{r_\ell} \left\lceil C_{\mathrm{ph}} L_{\mathrm{tag}} (v\epsilon^{-\gamma} + \nu_\ell \epsilon^{-2}) \right\rceil \right\rceil \le \frac{C L_{\mathrm{tag}}}{r_\ell} (v\epsilon^{-\gamma} + \nu_\ell \epsilon^{-2})$$

complete cycles, given sufficient shadow budget. Geometric refresh increases this quota by less than a factor four.

Proof. All label-specific intervals cover the same mean difference. Hence their intersection preserves the original safe elimination argument. For an active pair, the surviving reference gives $q_{ap,\ell} \le 2q_{aa_*,\ell} + 2q_{pa_*,\ell} \le 32v\epsilon^{2-\gamma} + 4\nu_\ell$. The inner ceiling in M_ℓ is therefore a sufficient number of accepted label-ℓ samples by the original empirical-Bernstein shrinkage calculation. Since $v \ge 1$ and $\epsilon \le 1$, it exceeds $2L_{\mathrm{tag}}$. Each nonanchor–anchor pair is queried once per complete cycle; the thinning event supplies at least the required number of accepted samples by M_ℓ cycles. This proves the paired quota with a single factor L_{tag}/r_ℓ. The marginal quota follows because each cycle observes every active action. The geometric-block argument of Lemma 7 proves the final statement. □

Upper bound in Theorem 6. The marginal argument gives $C\sqrt{NTL_{\mathrm{tag}}}$. For any label with $r_\ell > 0$ and $B \ge 1$, define the analysis-only precision

$$u_\ell = \left(\frac{C_0 v N \Lambda_{\mathrm{tag}}}{r_\ell B}\right)^{1/\gamma} + \sqrt{\frac{C_0 \nu_\ell N \Lambda_{\mathrm{tag}}}{r_\ell B}}.$$

The trivial regret bound suffices if $u_\ell \ge 1/2$. Otherwise choose a target phase with precision in $[u_\ell, 2u_\ell)$. The tagged quotas through this phase use at most

$$\frac{CNL_{\mathrm{tag}}}{r_\ell}(vu_\ell^{-\gamma} + \nu_\ell u_\ell^{-2}) \le \frac{B}{8K}$$

shadows for a sufficiently large numerical C_0, including the refresh factor. The original quota induction then completes the target before budget exhaustion, since $\lfloor B \rfloor \ge B/2$. All labels consume budget as they arrive. Horizon truncation is covered by the same partial-work bound.

The exploration-prefix regret is at most

$$\frac{CNL_{\mathrm{tag}}}{r_\ell}(vKu_\ell^{1-\gamma} + \nu_\ell/u_\ell) \le CBu_\ell.$$

After the target phase, all active gaps are at most Cu_ℓ, giving at most CTu_ℓ total regret. This also bounds the remainder under active-only handoff. Lemma 8 supplies its expected marginal remainder guarantee. The paired bounds hold on the same event for both labels, so taking their minimum is valid. Its complement contributes at most δT.

When $B < 1$, no shadow is feasible and the paired expression with $B \vee 1$ is at least T. A zero-mass label contributes $+\infty$. More generally, $r_\ell B < 1$ makes its paired branch trivial, and $\zeta_\ell = 1$ gives $\nu_\ell = 1$, whose floor term is dominated by the marginal guarantee since $r_\ell B \le T$. These cases establish the stated uniform bound. □

D.13 Converse for binary provenance

We prove the lower half of Theorem 6 for a fixed tag channel, granting its parameters to the policy. Let $d = N - 1$. For $v \geq 8$, use the reference P_0 and alternatives P_j of Appendix C.8, with $s = \min\{1/4, v\Delta^{2-\gamma} + \nu\}, \qquad 0 < \Delta \leq 1/64$. Their latent mixture variable R in (40) is independent of the channel contamination bit F. The reference and every alternative belong to $\mathcal{V}_\gamma(v, \nu)$, since $s \geq 8\Delta$ and their nonzero optimality gaps are at least Δ.

Information in an observed label and pair. Fix a pair containing j. Write p_i for its four-cell clean law and u_i for the product of its marginal laws in environment $i \in \{0, j\}$. Conditional on label ℓ, its observed law is $z_i^\ell = (1 - \zeta_\ell)p_i + \zeta_\ell u_i$. Every cell of p_j has mass at least $s/8$. An off-diagonal cell has mass $s\theta_a(1 - \theta_b) \geq s/8$ because $\theta_a, \theta_b \in [1/4, 1/2]$; a diagonal cell also receives $(1 - s)/2$. Every marginal mean lies in $[1/2 - 2\Delta, 1/2] \subset [15/32, 1/2]$, so every product cell has mass at least $1/8$. It follows that

$$z_j^\ell(x) \geq \frac{(1 - \zeta_\ell)s + \zeta_\ell}{8} \geq \frac{s + \zeta_\ell}{16}. \tag{98}$$

Changing θ_j moves each clean cell and each product cell by at most 2Δ. Thus $|z_0^\ell(x) - z_j^\ell(x)| \leq 2\Delta$ and

$$D_{\mathrm{KL}}(z_0^\ell \| z_j^\ell) \leq \sum_x \frac{(z_0^\ell(x) - z_j^\ell(x))^2}{z_j^\ell(x)} \leq \frac{256\Delta^2}{s + \zeta_\ell}. \tag{99}$$

The tag distribution is identical in both environments. Consequently, the KL of the complete labeled observation equals $\sum_\ell r_\ell D_{\mathrm{KL}}(z_0^\ell \| z_j^\ell)$. A pair excluding j is unchanged, and a singleton contributes at most $16\Delta^2$. Lemma 1, applied under the admissible reference, gives

$$\frac{1}{d}\sum_{j=2}^{N} D_{\mathrm{KL}}(\mathbb{P}_0^{\mathcal{A}} \| \mathbb{P}_j^{\mathcal{A}}) \leq \frac{16T\Delta^2}{d} + \frac{512B\Delta^2}{d} \sum_{\ell : r_\ell > 0} \frac{r_\ell}{s + \zeta_\ell}.$$

This is (94). It accounts directly for the observed scores and label, and hence for every statistic computed from them.

One gap that controls both strata. For every positive-mass label, put

$$x_\ell = \left(\frac{vd}{r_\ell \overline{B}}\right)^{1/\gamma}, \qquad z_\ell = \sqrt{\frac{\nu_\ell d}{r_\ell \overline{B}}},$$

$$\Delta = \kappa \min\left\{1, \sqrt{d/T}, \min_{\ell : r_\ell > 0} \max(x_\ell, z_\ell)\right\}, \qquad \kappa = 2^{-18}.$$

The singleton term is at most $16\kappa^2$. If $s = 1/4$, then $\sum_\ell r_\ell/(s + \zeta_\ell) \leq 4$, and $B \leq T$ bounds the paired term by $2048\kappa^2$. Otherwise, $s + \zeta_\ell = v\Delta^{2-\gamma} + \nu + \zeta_\ell \geq v\Delta^{2-\gamma} + \nu_\ell$. When $x_\ell \geq z_\ell$, the contribution before the factor 512 satisfies

$$\frac{Br_\ell\Delta^2}{d(s + \zeta_\ell)} \leq \frac{Br_\ell\Delta^\gamma}{vd} \leq \kappa^\gamma \frac{B}{\overline{B}} \leq \kappa.$$

When $z_\ell > x_\ell$, we have $\nu_\ell > 0$ and the same contribution is at most $\kappa^2 B/\overline{B} \leq \kappa^2$. There are at most two labels, so the average transcript KL is below $1/8$ in every case, uniformly even when one r_ℓ tends to zero.

For $N \geq 3$, Pinsker's inequality and the production-count argument of Appendix C.8 give $\frac{1}{d}\sum_{j=2}^{N} \mathbb{E}_j N_j^{\mathrm{prod}} \leq T/d + T/4 \leq 3T/4$. Some alternative therefore incurs regret at least $T\Delta/4$. For $N = 2$, comparing the reference with its single alternative gives a maximum regret of at least $3T\Delta/8$. Since $N/2 \leq d \leq N$ and $\max(x, z) \leq x + z \leq 2\max(x, z)$, this is the claimed lower rate. The fixed number of labels matters here: their information sum and largest information contribution differ by at most a numerical factor.

The complete scale range. For $1 \leq v \leq 8$, divide every score in the $v = 8$ source family by eight. Appendix C.8.5 proves membership in $\mathcal{V}_\gamma(v, \nu)$. This invertible transformation commutes with both conditional pair kernels and preserves the tag channel, all production decisions, and every shadow count. A target policy induces a source policy with eight times its regret. The source lower rate at $v = 8$ dominates the target rate, proving the result uniformly for all $v \geq 1$. The lower bound is for the complete labeled transcript and therefore applies to the full channel likelihood.

D.14 Equal accuracy with different decision complexity

We prove Corollary 3 on the common-support family (23), with $0 < g \le 1/8$ and either ordering. Fix $\bar{\zeta} = 1/2$. The conservative channel has $(r_0, \zeta_0) = (1/4, 0)$ and $(r_1, \zeta_1) = (3/4, 2/3)$; the aggressive channel has $(r_0, \zeta_0) = (3/4, 1/3)$ and $(r_1, \zeta_1) = (1/4, 1)$. Both satisfy $\mathbb{P}(C = F) = 3/4$. Erasing the label gives exactly the same score law $(P_{\sigma,g} + P_{\sigma,g,1} \otimes P_{\sigma,g,2})/2$ in each channel, where $P_{\sigma,g,a}$ denotes marginal a.

Their tag information also agrees. Define binary entropy in bits by $h_2(x) = -x \log_2 x - (1-x)\log_2(1-x)$, with $0 \log_2 0 = 0$. The marginal tag probabilities are $1/4$ and $3/4$, and each channel has conditional entropy $H(C \mid F) = 1/2$ bit. Therefore $I(F; C) = H(C) - H(C \mid F) = h_2(1/4) - 1/2$ for both channels. The mutual information is computed under the fixed Bernoulli$(1/2)$ contamination prior.

Conservative provenance. Query both actions on every round and serve the action with smaller cumulative loss over the preceding $C = 0$ pairs, breaking ties fairly. The rule needs neither g nor the ordering. Write $D = Y_{\text{bad}} - Y_{\text{good}}$ and $X = \mathbf{1}\{C = 0\}D$. Conditional on $C = 0$, D has probabilities $3g/2, g/2, 1 - 2g$ at $1, -1, 0$. With $\lambda = \frac{1}{2}\log 3$ and $c_0 = 2 - \sqrt{3}$, $\mathbb{E}e^{-\lambda X} = 1 - r_0 c_0 g$. Thus the probability of a wrong service decision after m observations is at most $(1 - r_0 c_0 g)^m$, and

$$R_T \le g \sum_{m=0}^{T-1} (1 - r_0 c_0 g)^m \le \frac{1}{r_0 c_0} = \frac{4}{2 - \sqrt{3}}. \tag{100}$$

For a lower bound, choose $g = 1/8$. Before receiving any observation, the average first-round regret over the two orderings is $g/2 = 1/16$. Hence the minimax regret is bounded above and below by numerical positive constants for every $T \ge 1$.

Aggressive provenance. Under the two orderings, the diagonal cells of both the clean and product laws agree. Their two off-diagonal differences are $+g$ and $-g$. The smaller product off-diagonal mass is $\frac{1}{4}(\frac{3}{4} - g) \ge \frac{5}{32}$. Each conditional mixture therefore has both off-diagonal masses at least $\zeta_\ell(5/32) \ge 5/96$. Its two changing cells differ by exactly g, giving $D_{\mathrm{KL}}(Q^\ell_{+,g} \| Q^\ell_{-,g}) \le \frac{2g^2}{5/96} = \frac{192}{5} g^2 \le 64 g^2$. The label law is the same in both orderings, so the complete pair-and-label KL is at most $64g^2$. Singleton KL is at most $6g^2$ by (27). For any nonanticipating policy, including optional shadows and arbitrary processing of past labels, the transcript divergence is consequently at most $64Tg^2$. Taking $g = 1/(32\sqrt{T})$ makes it at most $1/16$, hence total variation at most $1/4$. The two-ordering production-count argument gives

$$\max_{\sigma \in \{+,-\}} R_T(\mathcal{A}; \sigma, g) \ge \frac{3gT}{8} = \frac{3\sqrt{T}}{256}. \tag{101}$$

Production-only MOSS supplies the matching $C\sqrt{T}$ upper bound. This proves both exact orders. The score kernel, tag accuracy, and mutual information $I(F; C)$ coincide, while the conditional information for deciding the ordering has different gap dependence. In the aggressive channel, the population identity $P_{ab} = (3Q^0_{ab} - Q^1_{ab})/2$ recovers the clean law; the transcript lower bound continues to apply to every estimator based on this identity.

The symmetric 75% channel retained in the experiments has posterior contamination $1/4$ and $3/4$. Replacing $5/96$ by $5/128$ in the same cell bound again gives pair KL at most $64g^2$ and the identical square-root conclusion.

D.15 Proof of the hidden-resampling frontier

Fix a base law P and contamination probability $\bar{\zeta}$. Singleton observations have law P_a, while paired observations have law

$$Q_{ab} = (1 - \bar{\zeta})P_{ab} + \bar{\zeta}(P_a \otimes P_b).$$

These kernels are the corresponding marginals of $Q = (1 - \bar{\zeta})P + \bar{\zeta} \bigotimes_{a=1}^{N} P_a$. Drawing an independent vector from Q on every request therefore gives the same observable transcript law for any nonanticipating two-score policy. Production means and gaps are unchanged. If a_* witnesses the base moment condition and Y'_a, Y''_{a_*} are independent draws from the corresponding marginals, then

$$q^Q_{aa_*} = (1 - \bar{\zeta})q^P_{aa_*} + \bar{\zeta}\mathbb{E}(Y'_a - Y''_{a_*})^2 \le v\Delta_a^{2-\gamma} + \nu_{\bar{\zeta}}, \qquad \nu_{\bar{\zeta}} = (1 - \bar{\zeta})\nu + \bar{\zeta}.$$

Thus $Q \in \mathcal{V}_\gamma(v, \nu_{\bar\zeta})$, and Theorem 2 gives the upper bound with the same pathwise shadow cap.

For the converse, specialize the binary-provenance experiment to a constant label: $r_0 = 1$ and $\zeta_0 = \bar\zeta$. Then $\nu_0 = \nu_{\bar\zeta}$ and the observation kernel is exactly Q_{ab}. The lower bound in Appendix D.13, together with the score-rescaling argument for $1 \le v < 8$, gives the matching lower frontier. This proves Theorem 5.

E Supplement to Section 5: Target-preserving provenance

This appendix collects mechanism-generation, calibration and ordering robustness, noise-floor transitions, and request-lineage audit details supporting Section 5.

E.1 Generating and approximating provenance

A continuous channel from two clocks. For each paired evaluation, choose both endpoints from the past, then draw independent clocks $L_{\mathrm{r}} \sim \mathrm{Exp}(\lambda_{\mathrm{r}})$ and $U \sim \mathrm{Exp}(\rho_{\mathrm{w}})$. The first measures readiness of a shared evaluation context; the second is an exogenous wait allowance. The clocks are independent of endpoint choices and of the request and scoring randomness. At time U, draw a fresh production request from the target population. If $L_{\mathrm{r}} \le U$, evaluate both endpoints on this request, giving P_{ab}; otherwise evaluate the shadow on a fresh independent target request, giving $P_a \otimes P_b$. Retain both scores and report U after endpoint selection. In this interface, the execution component retains the readiness bit $F = \mathbf{1}\{L_{\mathrm{r}} > U\}$.

Conditionally on $U = u$,

$$\mathbb{P}(F = 1 \mid U = u) = e^{-\lambda_{\mathrm{r}} u}, \qquad \mathrm{Law}(\widetilde{Y}_a, \widetilde{Y}_b \mid U = u) = (1 - e^{-\lambda_{\mathrm{r}} u}) P_{ab} + e^{-\lambda_{\mathrm{r}} u}(P_a \otimes P_b).$$

Both conditional marginals preserve their production targets. With $C = -U$, the posterior $\zeta(C) = e^{\lambda_{\mathrm{r}} C}$ is nondecreasing. Writing $\alpha = \rho_{\mathrm{w}}/\lambda_{\mathrm{r}}$ gives

$$G(z) = \mathbb{P}\left(U \ge -\frac{\log z}{\lambda_{\mathrm{r}}}\right) = z^\alpha, \quad 0 < z < 1, \qquad \mathbb{P}(Z = 0) = 0, \qquad \mathbb{E}Z = \frac{\alpha}{1+\alpha}. \tag{102}$$

Consequently

$$\mathsf{H}(s) = \alpha \int_0^1 \frac{z^{\alpha-1}}{s+z}\, dz \sim \frac{\alpha\pi}{\sin(\pi\alpha)} s^{\alpha-1} \quad (0 < \alpha < 1).$$

For $\alpha = 1$, $\mathsf{H}(s) = \log((1+s)/s)$; for $\alpha > 1$, $\mathsf{H}(s) \to \alpha/(\alpha-1)$.

The request-independent clocks and post-wait target draw ensure target preservation. Exposing F supplies a clean mass $1/(1+\alpha)$. For predictable query indicators I_t, $\mathbb{E}\sum_t I_t U_t = \mathbb{E}M_T/\rho_{\mathrm{w}}$ measures serial waiting; parallel time also depends on scheduling. Capping U at $U_{\max}$ gives $Z_{\mathrm{cap}} = \max\{Z, e^{-\lambda_{\mathrm{r}} U_{\max}}\}$, a finite inverse moment, and eventual square-root regret. The fast-rate window ends at this floor. Changing α here changes $\mathbb{E}Z$; the family in (16) additionally keeps the erased law fixed across α.

Relative calibration error. Let $\widehat{Z} = \widehat{\zeta}(C) \in [0,1]$ and $\widehat{w}_s = s/(s+\widehat{Z})$. Define

$$\mathsf{H}_{\widehat{Z}}(s) = \frac{(\mathbb{E}\widehat{w}_s)^2}{\mathbb{E}[\widehat{w}_s^2(s+Z)]}. \tag{103}$$

This is the variational precision of the chosen weights. Cauchy–Schwarz gives $\mathsf{H}_{\widehat{Z}}(s) \le \mathsf{H}(s)$. Suppose $0 < \ell \le u$ and $\ell(s+Z) \le s + \widehat{Z} \le u(s+Z)$ almost surely. Introduce the tilted law $Q_s(dc) = (s+\zeta(c))^{-1} m_C(dc)/\mathsf{H}(s)$ and $r = (s+Z)/(s+\widehat{Z}) \in [a,b] = [1/u, 1/\ell]$. Cancellation yields

$$\frac{\mathsf{H}_{\widehat{Z}}(s)}{\mathsf{H}(s)} = \frac{(\mathbb{E}_{Q_s} r)^2}{\mathbb{E}_{Q_s} r^2}.$$

For $\bar r = \mathbb{E}_{Q_s} r$, the inequality $(r-a)(r-b) \le 0$ gives $\mathbb{E}_{Q_s} r^2 \le (a+b)\bar r - ab$. Minimizing $\bar r^2/((a+b)\bar r - ab)$ over $\bar r \in [a,b]$ gives

$$\frac{4\ell u}{(\ell+u)^2}\, \mathsf{H}(s) \le \mathsf{H}_{\widehat{Z}}(s) \le \mathsf{H}(s). \tag{104}$$

For $(1-\kappa)Z \le \widehat{Z} \le (1+\kappa)Z$, $0 \le \kappa < 1$, the retained fraction is at least $1-\kappa^2$. For $0 < \kappa < 1$, an additive bound $|\widehat{Z} - Z| \le \eta$ gives this fraction whenever $s \ge \eta/\kappa$.

Relative error also controls the range term: $\mathbb{E}\widehat{w}_s \ge s\mathsf{H}(s)/u$. At $s = S(\varepsilon) \ge \varepsilon$, the confidence range contribution $1/(\varepsilon \mathbb{E}\widehat{w}_s)$ is at most $u/(\varepsilon^2 \mathsf{H}(s))$. Together with (104), this retains the spectral confidence quota up to constants depending on ℓ, u.

For a fixed upward shift $\widehat{Z} = \min\{1, Z+\eta\}$, $0 < \eta < 1$, dominated convergence gives

$$\lim_{s \downarrow 0} \mathsf{H}_{\widehat{Z}}(s) = \frac{(\mathbb{E}[1/\widehat{Z}])^2}{\mathbb{E}[Z/\widehat{Z}^2]} < \infty \quad \text{if } \mathbb{P}(Z > 0) > 0. \tag{105}$$

Thus even a divergent true spectrum can yield bounded precision after a fixed positive shift. If $Z = 0$ almost surely, the shifted weights are constant and $\mathsf{H}_{\widehat{Z}}(s) = 1/s$.

Calibration, ordering, and target preservation. Fix a returned pair contrast $D = D_{ab}$ and write $d = d_{ab}$. The conditional identity $\mathbb{E}[D \mid C] = d$ validates tag-based confidence weights independently of their numerical calibration, provided each weight uses the current tag and past information rather than the current score. The weighted-SODA oracle additionally uses the true normalizer $\mathbb{E}w$: using an alternative normalizer $\widetilde{A} > 0$ changes the estimator mean to $(\mathbb{E}w/\widetilde{A})d$, so its stated oracle inputs remain relevant.

For an available family of tag subsets $\mathcal{L}$, indicator weights have precision

$$\mathsf{H}_{\mathcal{L}}(s) = \sup_{\substack{A \in \mathcal{L} \\ \mathbb{P}(C \in A) > 0}} \frac{\mathbb{P}(C \in A)}{s + \mathbb{E}[Z \mid C \in A]} \le \mathsf{H}(s).$$

For $0 < c_0 \le 1$ and $0 \le \xi \le 1$, if a declared lower cut A retains $\mathbb{P}(C \in A) \ge c_0 G(z)$ and has $\mathbb{P}(Z > z \mid C \in A) \le \xi$, then $\mathbb{E}[Z \mid C \in A] \le z + \xi$ and $\mathsf{H}_{\mathcal{L}}(s) \ge c_0 G(z)/(s + z + \xi)$. Thus $\xi = O(s+z)$ with constant retained mass preserves that cut's precision order. For completely reversed quality order, lower cuts are upper-Z tails; including the whole space gives $\mathsf{H}_{\mathcal{L}}(s) = 1/(s + \mathbb{E}Z)$. Retaining the original quality order avoids numerical-shift errors in an ordered policy. In contrast, if $\mathbb{E}[D \mid C] = d + b(C)$, weighting targets

$$d + \frac{\mathbb{E}[w(C)b(C)]}{\mathbb{E}w(C)}.$$

A confidence interval for the production target must account for this bias. Calibration efficiency and target validity therefore require separate checks.

The finite scale induced by $\nu > 0$. Let $a = 2 - \gamma > 0$, $0 < \nu \le 1/2$, and $\varepsilon_\nu = (\nu/v)^{1/a}$. For clarity write $\mathsf{J}_\nu(\varepsilon) = \varepsilon^2 \mathsf{H}(\min\{1, v\varepsilon^a + \nu\})$. For $0 \le \varepsilon \le \varepsilon_\nu$,

$$\tfrac{1}{2}\varepsilon^2 \mathsf{H}(\nu) \le \mathsf{J}_\nu(\varepsilon) \le \varepsilon^2 \mathsf{H}(\nu), \tag{106}$$

because $\nu \le S(\varepsilon) \le 2\nu$ and $\mathsf{H}(2\nu) \ge \mathsf{H}(\nu)/2$. For a dimension/logarithm multiplier $k > 0$, define

$$B_\nu(k) = \frac{k}{\mathsf{J}_\nu(\varepsilon_\nu)} = \frac{k}{\varepsilon_\nu^2 \mathsf{H}(2\nu)} \asymp \frac{k v^{2/(2-\gamma)}}{\nu^{2/(2-\gamma)} \mathsf{H}(\nu)}. \tag{107}$$

The constants in the last comparison lie between one and two. For $\overline{B} \ge B_\nu(k)$, (106) implies

$$\sqrt{\frac{k}{\overline{B}\mathsf{H}(\nu)}} \le e(k/\overline{B}) \le \sqrt{\frac{2k}{\overline{B}\mathsf{H}(\nu)}}.$$

For larger resolved gaps, $v\varepsilon^a \ge \nu$ and the positive-noise precision differs by at most a factor two from its zero-noise value at the same gap. Thus (107) marks the paired branch's transition to square-root scaling. The full regret guarantee also retains its production baseline. In Theorem 1, take $k = N$ for the lower bound and $k = N\log(2N)$ for the oracle upper; adaptive policies retain their stated additional logarithms.

At $\gamma = 1$, the crossover is $B_\nu(k) \asymp kv^2/(\nu^2 \mathsf{H}(\nu))$. If $G(z) \asymp z^\alpha$ with $0 < \alpha < 1$, this becomes $B_\nu(k) \asymp kv^2 \nu^{-(1+\alpha)}$ as $\nu \downarrow 0$. Every fixed $\nu > 0$ gives an eventual square-root rate for fixed channel and fixed N, v. For $\gamma = 2$, $v \ge 1$ gives $S(\varepsilon) = 1$ throughout. All losses are normalized to $[0,1]$; the clocks in (102) account separately for elapsed time.

E.2 Target-preserving provenance and the request-ID protocol

Protocol and information available to the learner. A trusted wrapper binds every score to its immutable i.i.d. request and job ID. Before scores arrive, independent routing F selects the production request or a fresh target request for the shadow, and an independent coin selects identity audits. Join by job ID and retain both packets. The public tag encodes routing and audit status; raw identities stay private. Every issued shadow costs one evaluation.

For predictable endpoints, matching yields P_{ab} and replay yields $P_a \otimes P_b$, both with marginals P_a, P_b. Conditional tag independence and Bayes' rule give (7) and (19). Binding establishes lineage; independent routing, auditing, and complete retention establish the common target.

At audit and contamination probabilities $1/2$, flagging only audited matches as clean gives $(\tau, \phi) = (1/2, 0)$; flagging only audited mismatches as dirty gives $(1, 1/2)$. These are the two channels of Corollary 3, with identical expected audit and score costs.

Preserving the complete audit readout. Let p_{aud} be the probability of an independent, outcome-blind audit and let ζ_{rep} be the independent-replay probability. The states match, different request, and unaudited have masses $p_{\mathrm{aud}}(1-\zeta_{\mathrm{rep}})$, $p_{\mathrm{aud}}\zeta_{\mathrm{rep}}$, and $1-p_{\mathrm{aud}}$, with posterior contaminations $0, 1, \zeta_{\mathrm{rep}}$. Their spectrum is

$$\mathsf{H}_3(s) = \frac{p_{\mathrm{aud}}(1-\zeta_{\mathrm{rep}})}{s} + \frac{p_{\mathrm{aud}}\zeta_{\mathrm{rep}}}{s+1} + \frac{1-p_{\mathrm{aud}}}{s+\zeta_{\mathrm{rep}}}. \tag{108}$$

Different IDs establish lineage; independence comes from fresh i.i.d. replay. Both binary encodings are deterministic functions of the full readout R, so observing R simulates either policy at the same cost. For $C = h(R)$, $Z_C = \mathbb{E}[Z_R \mid C]$, and Jensen's inequality gives

$$\mathsf{H}_C(s) = \mathbb{E}[(s + \mathbb{E}[Z_R \mid C])^{-1}] \le \mathbb{E}[(s+Z_R)^{-1}] = \mathsf{H}_3(s).$$

The binary channels are controlled coarsenings. Envoy exposes mirroring and shadow-span sampling (Envoy Project, n.d.); Istio's default mirroring discards responses (Istio Authors, n.d.). The wrapper adds score capture, identity binding, and complete retention, with outcome-blind audit selection as distinguished from outcome-dependent sampling (OpenTelemetry Authors, 2025).

Pricing provenance work. Let M_T count shadows and A_T^{aud} count audited shadows. In a common work unit, let $c_{\mathrm{eval}} > 0$ price an extra scored call, c_{join} matching and retention, c_{bind} binding, and c_{aud} the incremental detailed audit. All prices are fixed and nonnegative. Then

$$\mathcal{W}_T = (c_{\mathrm{eval}} + c_{\mathrm{join}} + c_{\mathrm{bind}})M_T + c_{\mathrm{aud}}A_T^{\mathrm{aud}}, \qquad \mathbb{E}\mathcal{W}_T = (c_{\mathrm{eval}} + c_{\mathrm{join}} + c_{\mathrm{bind}} + p_{\mathrm{aud}}c_{\mathrm{aud}})\mathbb{E}M_T.$$

The second equality follows by summing the conditionally mean-p_{aud} audit coins over predictable queries, including adaptive stopping. An expected allowance W gives

$$B = \min\left\{T, \frac{W}{c_{\mathrm{eval}} + c_{\mathrm{join}} + c_{\mathrm{bind}} + p_{\mathrm{aud}}c_{\mathrm{aud}}}\right\}.$$

Replacing p_{aud} by one in the denominator and taking the floor ensures a hard work cap. At fixed audit probability, the paired-information value per unit of extra work is proportional to $\mathsf{H}_3(s)/(c_{\mathrm{eval}} + c_{\mathrm{join}} + c_{\mathrm{bind}} + p_{\mathrm{aud}}c_{\mathrm{aud}})$. This accounting keeps metadata costs explicit when translating the score-count frontier into a supplied work budget.

Request-ID audit details. For the request-ID design in Section 6.3, the six action pairs, two public labels, and two encodings give 24 conditional-gap identities. Exact enumeration matches these identities to 2.8×10^{-17}. With seed 20260920, the packet experiment contains 60,181 audited jobs; score-dependent arrival order is joined by job ID with complete pair retention. All 24 empirical gaps lie within the simultaneous Hoeffding intervals, with maximum discrepancy 0.01379, and every audited identity agrees with its routing state.

Selection can change the target despite correct matching. Take a fully clean pair $(Y_{\mathrm{good}}, Y_{\mathrm{bad}}) = (E, E \oplus W)$ with $E \sim \mathrm{Bernoulli}(1/4)$ and $W \sim \mathrm{Bernoulli}(2g)$. Its bad-minus-good difference D has mean g and second moment $2g$. Accept a tag with probability 0.9 when $Y_{\mathrm{good}} = 1$ and 0.1 otherwise. Acceptance has probability 0.3, while $\mathbb{E}[\mathbf{1}\{C=0\}D] = 0.1(1.5g) - 0.9(0.5g) = -0.3g$. Thus the accepted conditional gap is $-g$, reversing the action ordering. At $g = 0.05$, the two

conditional gaps are -0.05 and 0.092857. This example isolates target selection from request matching.

For a known positive selection probability $\pi_c(O) = \mathbb{P}(C = c \mid O)$, where O contains the observed pair, the raw-query estimator $X_c = \mathbf{1}\{C = c\}D/\pi_c(O)$ has $\mathbb{E}X_c = d$, $\mathbb{E}X_c^2 = \mathbb{E}[D^2/\pi_c(O)]$, $|X_c| \le 1/\inf \pi_c$. These are inverse-probability identities (Horvitz & Thompson, 1952). In the preceding example its second moment is $140g/9$, equal to 0.777778 at $g = 0.05$, compared with the unweighted 0.1. Correction restores the target and changes its statistical cost. More generally, a known conditional bias bound $|\mathbb{E}[D \mid C = c] - d| \le b_c$ permits valid intersection after widening that stratum's interval by b_c.

F Supplement to Section 6: Experiments and numerical evaluation

The supplements follow the service, spectrum, certification, and handoff experiments in Section 6. They provide sampling details, random seeds, inference procedures, numerical calculations, and detailed result tables; the final subsection gives fixed-channel calculations for the binary separation.

F.1 Unknown-gap service under a continuous provenance spectrum

Matched sampling. For the service grid in Section 6.1, sample the hidden resampling flag $F \sim \text{Bernoulli}(1/2)$ before the clean or independent-coordinate score vector. Then draw $C \mid F$ by rejection from $\text{Unif}(0, 1)$, accepting with probability $\zeta_\alpha(C)$ for $F = 1$ and $1 - \zeta_\alpha(C)$ for $F = 0$, using (16). Both acceptance probabilities integrate to $1/2$, so the resulting channels have identical erased-score paths. Tag proposals, score primitives, tie coins, and continuation indices use disjoint random streams. Production choices precede current scores and tags.

Policy correspondence and controls. Weighted EO uses the natural-family scale ε_m from Section 4.4; unweighted EO changes only its weight to one. Ordered service uses the dyadic rank bank and L_{srv} specified in Section 6.1, with radius (86). A nonempty intersection with the marginal interval that excludes zero certifies an action. Before certification, the server uses the sign of the cached smallest-radius prefix, updating its sum after each observation; an empty intersection keeps this service rule without issuing a certificate. All sign ties are fair. Appendix D.10 proves the prefix guarantee.

Production-only MOSS uses horizon T, observes only its served coordinate, and spends no shadows. The rate-based selector chooses weighted EO exactly when $m \ge 2$ and $\varepsilon_m \le T^{-1/2}$, otherwise MOSS; this choice is fixed across gaps and orderings. Each paired branch spends m shadows and then samples uniform prefix indices to replay its actual production actions. Recorded regret is g times the full-horizon wrong-action count. Three additional controls replace uniform-past continuation by the same MOSS tail initialized from both prefix coordinates, so their total-regret differences equal their prefix differences.

Seeds and paired inference. Every service cell uses seeds $972000, \ldots, 972031$, with both action orderings averaged within each seed cluster. Cellwise intervals and tests operate on whole clusters. For each maximum over the twelve gap means, resample the same cluster indices jointly across gaps and methods and recompute the maximum; the resampling count is given in Section 6.1. Maximum-risk bands are percentile intervals; gap-wise bands are pointwise t_{31} intervals. Tables retain zero-budget MOSS, while logarithmic budget axes show the four positive budgets. Table 2 and Figure 2 report the complete largest-horizon results.

Aggregate outcomes. All 76,800 evaluated action-ordering paths respect the shadow cap. The experiment records 2152 certified paths, zero incorrect certificates, and zero empty intersections.

Across 1,200 cells, the Holm-corrected wins/losses/unresolved counts for the first method are weighted versus unweighted EO: $1/0/1199$; weighted versus MOSS: $144/1/1055$; order-only versus unweighted EO: $0/0/1200$; order-only versus MOSS: $173/4/1023$; and the rate-based selector versus MOSS: $139/0/1061$. With the common MOSS tail, weighted versus unweighted has $1/0/1199$ and order-only versus unweighted has $0/0/1200$.

The erased-score decision control. At a fixed gap in the two-action family, write $p_{01}(g) = \mathbb{P}_+(Y_1 = 0, Y_2 = 1)$ and $p_{10}(g) = \mathbb{P}_+(Y_1 = 1, Y_2 = 0)$ for the erased experiment. Both are positive,

$p_{01}(g) - p_{10}(g) = g$, and reversing the action ordering exchanges these two cells while preserving the diagonal cells. The full-pair log-likelihood ratio after n observations is therefore

$$\log \frac{p_{01}(g)}{p_{10}(g)} \sum_{i=1}^{n} (Y_{2,i} - Y_{1,i}).$$

The coefficient is positive for every $g > 0$. Consequently unweighted EO uses the equal-prior likelihood-ratio decision sign without knowing the gap. Its comparison with weighted EO isolates provenance at matched score access.

Paired maximum-risk follow-up. For each of the 100 original (α, T, B) configurations, let $X_{i,a,g}$ denote method a's actual production regret averaged over the two action orderings for seed i, and let $\mathcal{G}_{12}$ denote the original twelve-gap grid. We compare $\max_{g\in\mathcal{G}_{12}} \mathbb{E}X_{i,a,g} - \max_{g\in\mathcal{G}_{12}} \mathbb{E}X_{i,b,g}$. This analysis uses the same trajectories as the primary grid. Each of 20,000 bootstrap replicates resamples 32 complete seed vectors, maximizes each method's resampled gap means separately, and subtracts. A shared index matrix (seed 982527) preserves pairing across gaps, methods and configurations. The 2.5th and 97.5th percentiles give the nominal 95% intervals in Table 3. The maximum is non-smooth at tied maximizers, so we also project simultaneous coordinate-mean bands: if every mean lies in $[L_{a,g}, U_{a,g}]$, the difference lies in

$$\left[\max_g L_{a,g} - \max_g U_{b,g},\ \max_g U_{a,g} - \max_g L_{b,g}\right].$$

We retain Student-t/Bonferroni bands with 31 degrees of freedom and bounded Hoeffding bands, separately over 48 coordinates per configuration and all 4,800 coordinates jointly. The former uses an approximate mean model; the latter has distribution-free coverage under independent bounded seed clusters. Their projected intervals contain zero for weighted and order-only versus unweighted EO at every configuration. Across the 100 configurations, the nominal intervals exclude zero in 7 weighted and 10 order-only comparisons against unweighted EO; these counts use the nominal, unadjusted intervals.

F.2 Continuous-spectrum information and risk bounds

This subsection derives the information and finite-family bounds evaluated on the spectrum grid in Section 6.1. The 11,568 channel–gap–horizon configurations are deterministic calculations on the common erased-score law.

Write $a_g = 3/8 - 3g/2$. Given contamination z, the disagreement probability is $q_z = 2g + a_g z$, and the two off-diagonal probabilities are $(q_z + g)/2$ and $(q_z - g)/2$. Consequently the complete tagged-score KL is

$$D_\alpha(g) = g\mathbb{E}\log\frac{3g + a_g Z}{g + a_g Z} = \frac{g^2}{a_g}\int_1^3 \mathsf{H}(gu/a_g)\,du. \tag{109}$$

Table 2: Maximum tested-gap mean regret at $T = 2^{18}$.

α	B	Weighted	EO	Order	MOSS	Selector	W+MOSS	EO+MOSS	O+MOSS
0.25	0	203.86	203.86	203.86	203.86	203.86	203.86	203.86	203.86
0.25	512	1472.69	1544.72	1538.32	203.86	203.86	211.18	211.21	211.21
0.25	4096	288.38	428.14	429.95	203.86	288.38	208.12	208.80	208.36
0.25	32768	70.35	196.23	100.22	203.86	70.35	129.86	148.47	134.88
0.25	262144	13.06	66.28	27.28	203.86	13.06	13.06	66.28	27.28
0.5	0	203.86	203.86	203.86	203.86	203.86	203.86	203.86	203.86
0.5	512	1500.06	1544.72	1541.78	203.86	203.86	211.12	211.21	211.21
0.5	4096	512.43	428.14	430.25	203.86	203.86	207.40	208.80	208.81
0.5	32768	156.24	196.23	104.21	203.86	156.24	134.24	148.47	135.38
0.5	262144	35.88	66.28	37.62	203.86	35.88	35.88	66.28	37.62
0.75	0	203.86	203.86	203.86	203.86	203.86	203.86	203.86	203.86
0.75	512	1740.61	1544.72	1535.60	203.86	203.86	211.09	211.21	211.20
0.75	4096	518.96	428.14	429.25	203.86	203.86	208.29	208.80	208.81
0.75	32768	152.69	196.23	141.11	203.86	152.69	138.06	148.47	140.00
0.75	262144	33.08	66.28	59.99	203.86	33.08	33.08	66.28	59.99
1.5	0	203.86	203.86	203.86	203.86	203.86	203.86	203.86	203.86
1.5	512	1606.69	1544.72	1534.84	203.86	203.86	211.30	211.21	211.20
1.5	4096	365.88	428.14	429.14	203.86	203.86	208.14	208.80	208.81
1.5	32768	115.93	196.23	195.46	203.86	203.86	136.85	148.47	148.38
1.5	262144	46.05	66.28	61.16	203.86	46.05	46.05	66.28	61.16

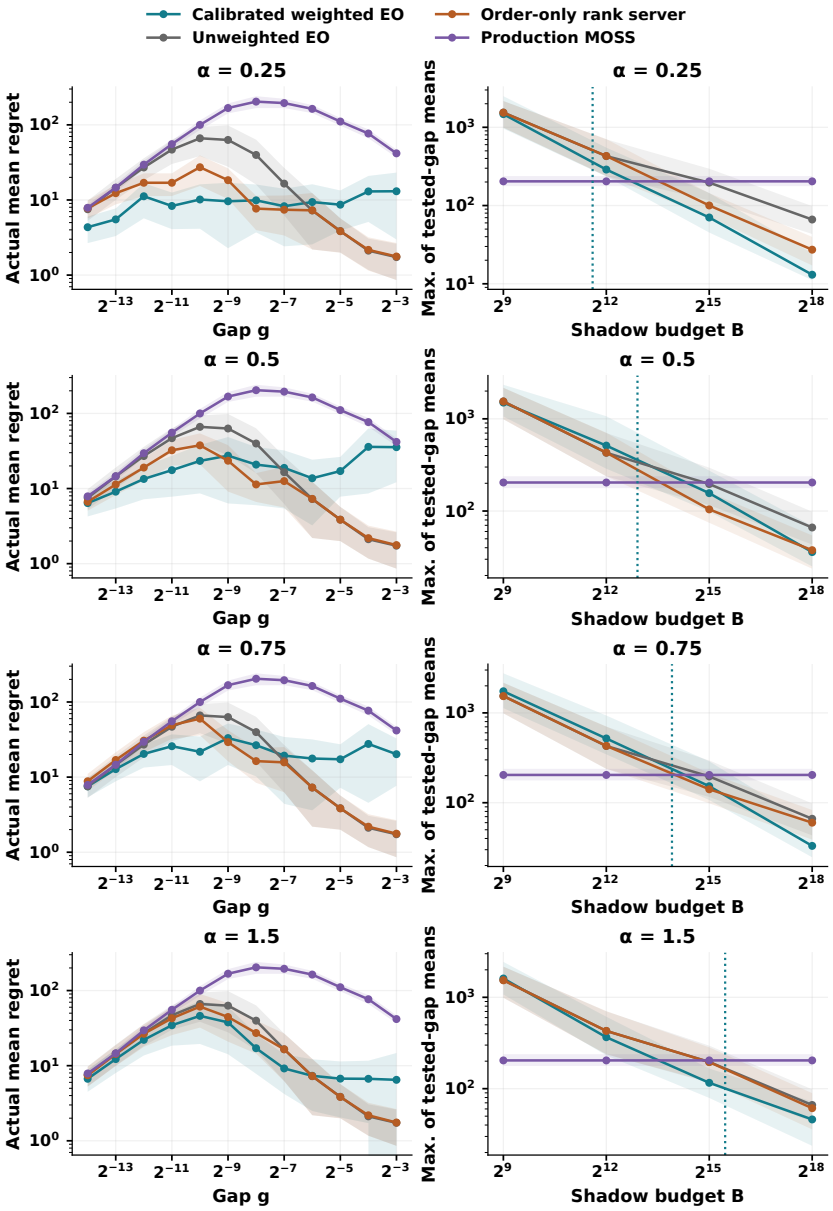

Figure 2: Service regret across provenance channels.

Table 3: Tested-gap risk differences at $T = 2^{18}$.

α	B	Weighted − unweighted	Order-only − unweighted
0.25	2^9	−72.04 [−726.63, 1000.22]	−6.40 [−25.07, 29.04]
0.25	2^{12}	−139.76 [−365.42, 182.11]	1.81 [−2.44, 7.07]
0.25	2^{15}	−125.89 [−221.56, −8.89]	−96.01 [−176.62, −18.55]
0.25	2^{18}	−53.21 [−85.27, −23.19]	−39.00 [−66.43, −18.12]
0.75	2^9	195.89 [−575.03, 1260.83]	−9.12 [−25.62, 35.90]
0.75	2^{12}	90.82 [−216.99, 543.70]	1.12 [−2.83, 6.05]
0.75	2^{15}	−43.54 [−145.70, 118.50]	−55.13 [−108.26, 14.80]
0.75	2^{18}	−33.20 [−63.02, 1.78]	−6.29 [−36.76, 20.42]

The quadrature specified in Section 6.1 evaluates the latter integral using the analytic channel spectrum.

A shared finite-family minimax bracket. The lower bound is $L_{\alpha,g}(T) = (g/4)\sum_{n=0}^{T-1} e^{-nD_\alpha(g)}$, using the full recorded score/tag law. The upper policy updates EO only on tags satisfying the

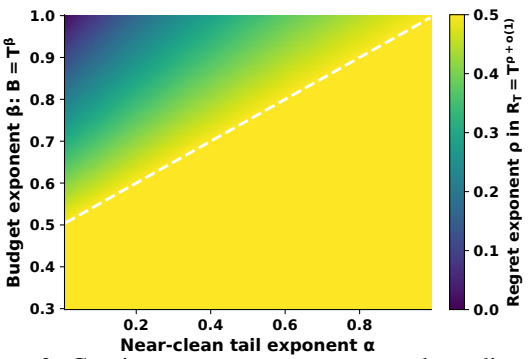

Figure 3: Continuous-spectrum exponent phase diagram.

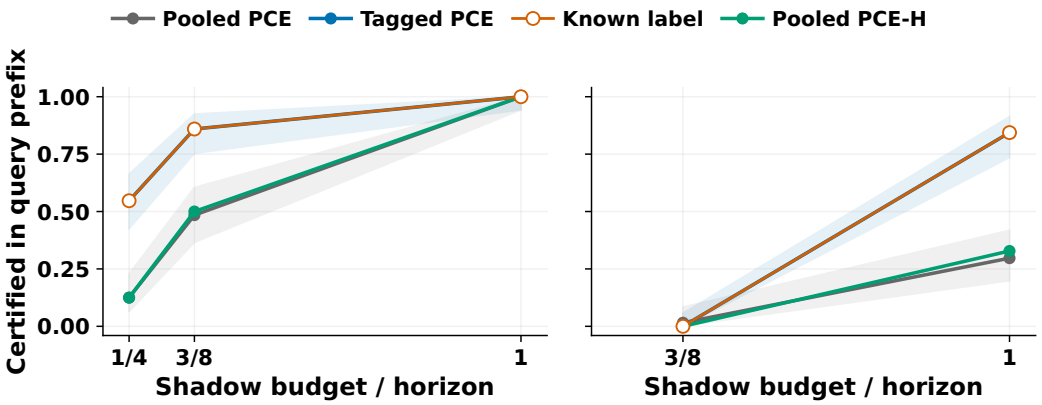

Figure 4: Conservative-channel certification before handoff or horizon end.

horizon-only cutoff $Z \le z_T$ specified in Section 6.1. One channel- and horizon-dependent rule handles every gap and ordering. Its acceptance mass is $r_T = (1/2)(4z_T)^\alpha$, and its accepted mean contamination is $\bar{z}_T = \alpha z_T/(\alpha+1)$. Put

$$q_T(g) = 2g + a_g \bar{z}_T, \qquad \kappa_T(g) = -\log\left(1 - \frac{r_T g^2}{q_T(g) + \sqrt{q_T(g)^2 - g^2}}\right).$$

The ternary accepted-difference walk has Chernoff coefficient $e^{-\kappa_T(g)}$. Its first decision has error $1/2$, giving the analytic policy bound

$$U_{\alpha,g}(T) = \min\left\{Tg,\ g\left[\sum_{n=0}^{T-1} e^{-n\kappa_T(g)} - \tfrac{1}{2}\right]\right\}. \tag{110}$$

Maximizing both bounds over the same gap family brackets its minimax regret. In Figure 1, solid curves are analytic policy upper bounds, dashed curves are universal lower bounds, and shaded regions are their brackets. Every comparison is charged, including rejected tags. The risk panel uses $\alpha < 1$; $\alpha = 3/2$ is an information control.

Figure 3 evaluates the exponent in Corollary 1 at $\gamma = 1, \nu = 0$. Its boundary $\beta = (1+\alpha)/2$ separates the square-root regime from the budget-dependent exponent $1 - \beta/(1+\alpha)$.

F.3 Direct Tagged-PCE mechanism experiment

Conditional laws and policy variants. The four binary channels in Section 6.2 induce $(r_0, \zeta_0) = (1/8, 0), (1/4, 0), (13/50, 1/26), (3/4, 1/3)$ in the stated order. They share erased scores, means, and raw-query costs; the conservative and aggressive channels also share accuracy and mutual information. Tags arrive after endpoint selection.

Each paired interval uses (48) with L_{tag} from the main design, retaining phase-local and historical statistics. Pooled PCE intersects marginal and pooled intervals; Tagged-PCE additionally intersects

Table 4: Tagged-PCE comparisons.

N	Channel	B/T	Certified: pooled/tagged	Shadows saved, millions
4	Rare pure	$1/4$	$8/8$	$0.000\ [0.000, 0.000]$
4	Conservative	$1/4$	$8/35$	$0.181\ [0.002, 0.360]$
4	Residual floor	$1/4$	$8/10$	$0.051\ [-0.043, 0.145]$
4	Aggressive	$1/4$	$8/8$	$0.008\ [-0.060, 0.076]$
4	Rare pure	$3/8$	$31/31$	$0.000\ [0.000, 0.000]$
4	Conservative	$3/8$	$31/55$	$0.866\ [0.490, 1.242]$
4	Residual floor	$3/8$	$31/26$	$0.091\ [-0.114, 0.297]$
4	Aggressive	$3/8$	$31/31$	$0.096\ [-0.086, 0.278]$
4	Rare pure	1	$64/64$	$0.000\ [0.000, 0.000]$
4	Conservative	1	$64/64$	$1.727\ [1.142, 2.311]$
4	Residual floor	1	$64/64$	$0.191\ [-0.244, 0.626]$
4	Aggressive	1	$64/64$	$0.222\ [-0.159, 0.603]$
8	Rare pure	$3/8$	$1/1$	$0.000\ [0.000, 0.000]$
8	Conservative	$3/8$	$1/0$	$-0.001\ [-0.003, 0.001]$
8	Residual floor	$3/8$	$1/1$	$0.0001\ [-0.0001, 0.0003]$
8	Aggressive	$3/8$	$1/2$	$0.002\ [-0.001, 0.006]$
8	Rare pure	1	$19/19$	$0.000\ [0.000, 0.000]$
8	Conservative	1	$19/54$	$1.805\ [1.216, 2.394]$
8	Residual floor	1	$19/22$	$-0.047\ [-0.313, 0.219]$
8	Aggressive	1	$19/19$	$0.092\ [-0.242, 0.426]$

Certification counts are out of 64. Positive savings favor Tagged-PCE; brackets give 95% paired intervals. Every paired production-regret difference is zero.

both label intervals; the known-label reference uses the least-contaminated label. Tagged-PCE receives scores and observed tags, while (τ, ϕ) enter only the channel generator. In this experiment, Pooled PCE, Tagged-PCE, and the known-label reference all use the active-set MOSS handoff of Appendix C.3; PCE-H adds geometric within-phase anchor refresh. All four use pooled-history EO.

Seeds and certification endpoints. Seeds 920000–920063 are shared across methods and channels, yielding $20 \times 4 \times 64 = 5,120$ runs. Figure 4 shows the conservative-channel results for four actions on the left and eight on the right.

Certification requires a singleton before handoff or horizon end. Handoff freezes the active set, leaving uncertified paths censored. Certification-time medians condition on successful issuance; restricted uncertified time counts unsuccessful paths through T. Wrong eliminations, wrong certificates, and shadow-cap violations are recorded separately. A tag-decisive completion occurs when stratified intervals complete a phase whose pooled intersection is still too wide.

Inference and results. Certification intervals and exact discordant-pair tests follow Section 6.2; shadow and regret contrasts use paired-t intervals. Holm correction covers all 20 cells separately per endpoint. Three settings improve certification and three save shadows (Table 4; Figure 5); tags cause phase completion on 855 of 1,280 matched paths. Every paired production-regret difference is zero, and unknown-quality and known-label query paths coincide throughout.

Exact enumeration gives the shared erased laws and conditional marginals to numerical precision. All 5,120 runs respect the cap, with zero observed wrong eliminations, wrong certificates, empty intersections, or fallbacks.

F.4 Fixed-grid ordered-cut evaluation

The ordered-cut grid and confidence logarithm are specified in Section 6.2. The experiment evaluates eleven predetermined numerical cuts, while the ordered-confidence guarantee in Appendix D.10 uses the growing rank bank. The learner receives tags and their clean-to-dirty order; calibration and α remain environment parameters. Pooled PCE and the ordered bank share phase-fixed historical-mean anchors, historical confidence reuse, pooled EO, and active-only MOSS handoff. Every raw shadow is charged.

For each cut, retain count n, sum U, and squared sum V, and use radius $(2\sqrt{L_{\mathrm{cut}} V} + 4L_{\mathrm{cut}})/n$. Since $D \in \{-1, 0, 1\}$, squared sums are integers $0, \ldots, T$; a union over the optimizing exponential test rates at these values justifies this radius. The main-design logarithm covers pair streams, phases,

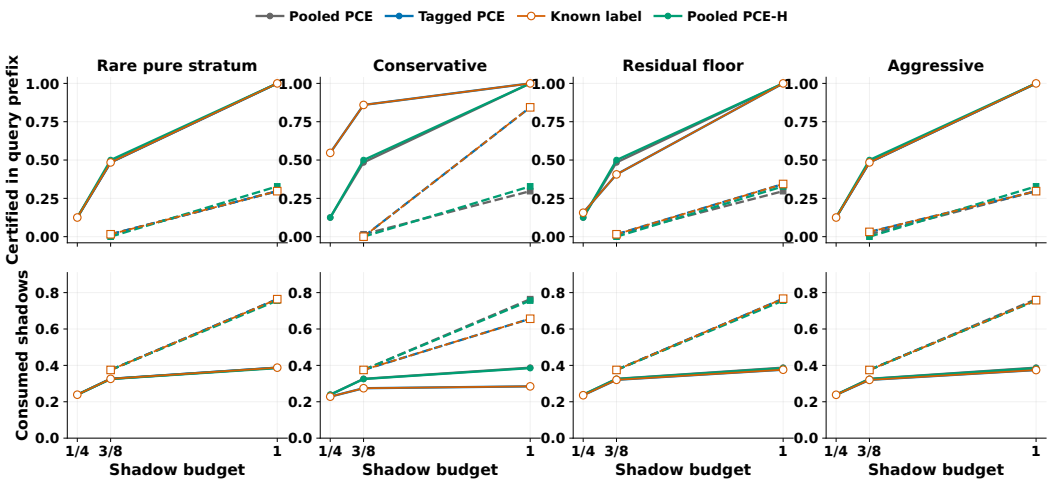

Figure 5: Certification and shadow use.

Table 5: All ordered-cut evaluation cells.

B/T	α	Certificates / 32		Shadows (10^6)		Regret
		Pooled	Ordered	Pooled	Ordered	Both
3/8	0.25	15	32	5.407	3.432	98.74
3/8	0.50	16	29	5.214	4.175	105.28
3/8	0.75	17	28	5.105	4.791	97.32
3/8	1.50	15	13	5.240	5.373	109.88
1	0.25	32	32	6.891	3.432	98.74
1	0.50	32	32	6.474	4.228	105.28
1	0.75	32	32	6.306	4.968	97.32
1	1.50	32	32	6.640	5.994	109.88

cuts, and marginal intervals. Bernoulli thinning yields the best-grid-cut quota while charging rejected tags.

Seeds $930000, \ldots, 930031$ give 512 runs. The 264 conditional strata have identical erased kernels and conditional-gap discrepancy below 3.5×10^{-18} under exact enumeration and quadrature. Every run respects the cap, with zero observed wrong eliminations, wrong certificates, or empty intersections.

Table 5 reports certificates obtained during the paired prefix and mean shadow expenditure and production pseudo-regret. At $B/T = 3/8$, paired exact certification tests have Holm-adjusted p-values $0.000123, 0.0165, 0.0444, 1$ in increasing α order. Full-budget mean shadow reductions are $50.2\%, 34.7\%, 21.2\%, 9.7\%$; their paired two-sided tests have adjusted p-values below 0.0000033, 0.00094, 0.065, and 0.70. Holm correction is applied over all eight cells separately for each endpoint. At $B/T = 3/8, \alpha = 1.5$, ordered cuts yield fewer certificates and greater mean expenditure. Production regrets coincide for every paired seed across the grid. Certification-time summaries use explicit right censoring; successful-run times are conditional on certification.

Figure 6 shows certification fractions with Wilson intervals and mean raw-shadow use with t-based intervals, for $B/T = 3/8$ on the left and $B/T = 1$ on the right.

F.5 Complete multi-action policies with balanced optimal indices

Matched instances. The grid and gap profile are specified in Section 6.2. Extend the shared-noise generator by taking $W_* = 0$ and independent $W_j \sim \text{Bernoulli}(2g_j)$ for nonoptimal actions. Use the same erased-first channel construction as Appendix F.1. The 32 independent clusters place the optimum eight times at each four-action index and four times at each eight-action index, randomly permuting the remaining gaps. All methods, channels, and budgets share these assignments and erased-score primitives, giving 1,280 runs.

Confidence banks and continuation. The three PCE variants share phase-fixed historical-mean anchors, active-target cycles, pooled-history EO, complete-cycle elimination, and all marginal

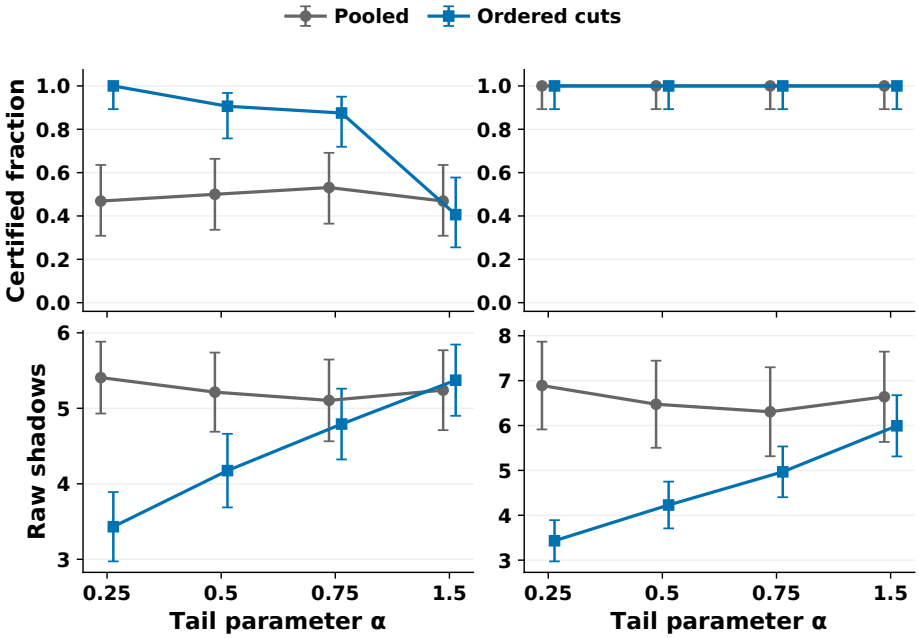

Figure 6: Ordered-tag certification and shadow use.

Table 6: Mean multi-action production regret.

N	α	B/T	Pooled	Calibrated	Rank	MOSS	Handoff
4	0.25	0.25	49472.33	49467.91	49467.91	1339.37	204.43
4	0.25	1	204.43	204.43	204.43	1339.37	204.43
4	0.75	0.25	49472.33	49472.33	49472.33	1339.37	204.43
4	0.75	1	204.43	204.43	204.43	1339.37	204.43
8	0.25	0.25	107521.36	107521.36	107521.36	2466.05	491.94
8	0.25	1	491.94	491.94	491.94	2466.05	491.94
8	0.75	0.25	107521.36	107521.36	107521.36	2466.05	491.94
8	0.75	1	491.94	491.94	491.94	2466.05	491.94

observations. Their phase-local paired banks use pooled contrasts, the dyadic calibrated weights of Section 4.2, or growing dyadic rank prefixes with whole tied groups. Each bank includes the pooled interval. Paired intervals refresh at dyadic counts and use (73) with L_{sp}; marginal intervals update every cycle. The calibrated policy receives $\zeta(c)$ and the rank policy receives the tag order as their respective provenance inputs.

After shadow exhaustion, these PCE variants continue marginal sampling and elimination. The pooled-handoff control has the identical pooled prefix, then freezes the surviving active set and initializes MOSS from all marginal observations. Anchors change only at phase boundaries in every policy. Production-only MOSS uses horizon T, one served-arm observation per request, and no shadows.

Inference and results. The estimand averages uniformly over optimal indices, permutations, and score histories. Within each optimal-index stratum, resample its $32/N$ independent clusters and weight strata equally, using the 4,000 shared resamples specified in the main design. The resulting percentile intervals are approximate and pointwise. Paired tests use stratified variance, Welch–Satterthwaite degrees of freedom, and Holm correction over eight cells separately per method comparison. Figure 7 summarizes mean production regret and normalized shadow use across all eight configurations, while Table 6 reports the corresponding cellwise mean production regret.

No run eliminates the optimum or produces inconsistent intervals. Every run respects the hard shadow cap; each PCE method in every cell issues $0/32$ certificates by the horizon. A missing certificate is right-censored, and its time is excluded from conditional certification-time summaries.

Both provenance policies have identical regret to pooled PCE on every path in seven of eight cells. In the remaining cell ($N = 4, \alpha = 1/4, B = T/4$), four paths change: the paired mean difference

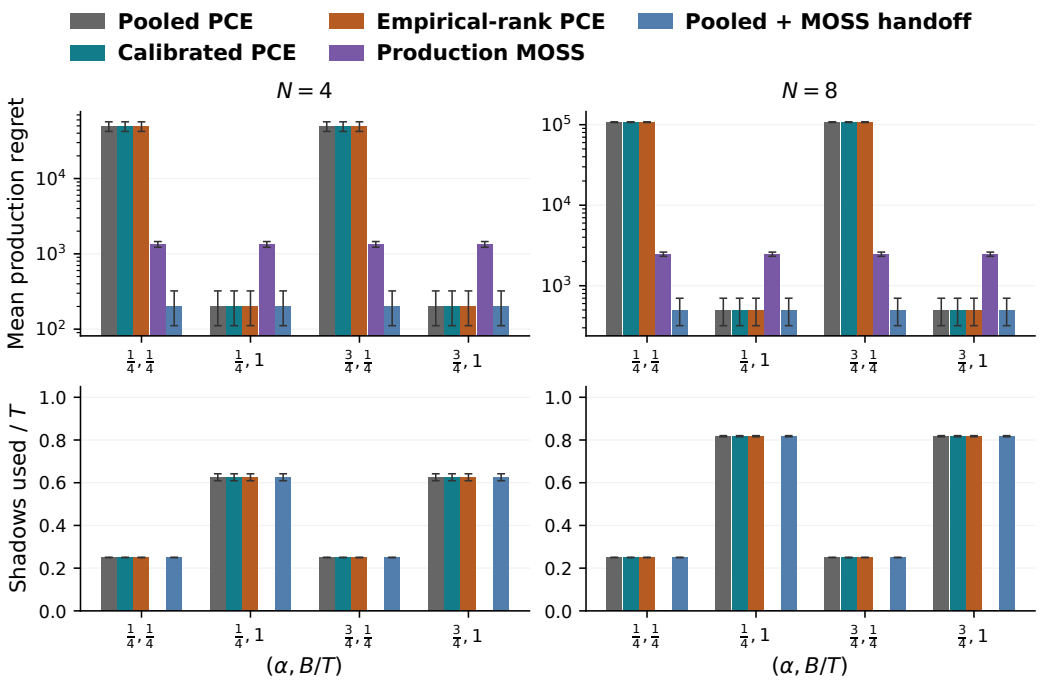

Figure 7: Multi-action regret and shadow use.

Table 7: Full-budget regret by optimal-action index.

N	Best index	Seeds	Each PCE variant	MOSS
4	1	8	166.05	1110.93
4	2	8	288.01	1346.10
4	3	8	231.78	1400.97
4	4	8	131.90	1499.49
8	1	4	259.37	2231.16
8	2	4	433.45	2306.73
8	3	4	724.53	2601.89
8	4	4	258.91	2564.19
8	5	4	370.23	2222.41
8	6	4	1047.75	2729.85
8	7	4	450.98	2421.02
8	8	4	390.29	2651.13

is -4.43, with pointwise bootstrap interval $[-8.99, -0.50]$ and Holm-adjusted $p = 0.577$. At full budget in that channel, each provenance policy uses 124.31 fewer shadows on average out of about 10.50 million, with unchanged production actions. The remaining shadow-count differences are zero. None of the eight provenance-versus-pooled regret comparisons is significant after Holm correction.

Against production-only MOSS, each provenance policy has significantly lower regret in the four full-budget cells and higher regret in the four partial-budget cells; against pooled handoff, it has higher regret in all four partial-budget cells and identical regret at full budget. In this configuration, the largest observed service-regret changes are associated with paired observation and the post-budget continuation rule. The full-budget dependence on the optimal-action index is reported in Table 7.

F.6 Decision value of provenance at a fixed score channel

Matched unlabelled score laws. Mix each two-action law in Proposition 1 equally with its product marginals. Let $a_0 = \tau/2$ and $b_0 = \phi/2$ be accepted clean and independent masses; the complementary masses are $1/2 - a_0$ and $1/2 - b_0$. Tags are independent of scores conditional on the component. Matching accuracy is $a_0 + 1/2 - b_0$.

The score-only, symmetric 75%, conservative 75%, aggressive 75%, and perfect channels respectively use $(a_0, b_0) = (1/2, 1/2), (3/8, 1/8), (1/4, 0), (1/2, 1/4), (1/2, 0)$. Score-only feedback has a constant tag. The three 75% channels share accuracy and erased scores; conservative/aggressive information is $I(F; C) = h_2(1/4) - 1/2 = 0.3112781245$ bits. Their accepted masses and contaminations are $(1/4, 0)$ and $(3/4, 1/3)$; the aggressive complement is independent. The symmetric channel accepts mass $1/2$ at contamination $1/4$. Every round reveals both scores, so $B = T$, and production uses only past observations.

Exact policy risk. The declared EO policy updates its canonical difference sum only on accepted tags. It uses an independent fair coin at a zero sum. Put

$$q_{\mathrm{acc}} = 2a_0 g + b_0(3/8 + g/2), \qquad d_{\mathrm{acc}} = (a_0 + b_0)g, \qquad p_{\mathrm{acc}} = \frac{q_{\mathrm{acc}} + d_{\mathrm{acc}}}{2q_{\mathrm{acc}}}.$$

These are the probability of an accepted nonzero difference, its unconditional drift, and the positive-step probability conditional on a nonzero accepted difference. For $k \geq 0$, define

$$h_k = \mathbb{P}\{\mathrm{Bin}(k, p_{\mathrm{acc}}) < k/2\} + \tfrac{1}{2}\mathbb{P}\{\mathrm{Bin}(k, p_{\mathrm{acc}}) = k/2\}.$$

Then $h_0 = 1/2$ and $h_{2j} = h_{2j-1}$ for $j \geq 1$. The exact paired-prefix policy risk, including commitment after B paired requests, is

$$\begin{aligned} R^{\mathrm{EO}}_{T,B} = \frac{g}{q_{\mathrm{acc}}} \sum_{k=0}^{B} h_k \mathbb{P}\{\mathrm{Bin}(B, q_{\mathrm{acc}}) > k\} \\ + (T - B)g \sum_{k=0}^{B} h_k \mathbb{P}\{\mathrm{Bin}(B, q_{\mathrm{acc}}) = k\}. \end{aligned} \tag{111}$$

To derive the first term, count the nonzero accepted increments before each production decision and use $\sum_{n=0}^{B-1} \mathbb{P}\{\mathrm{Bin}(n, q) = k\} = \mathbb{P}\{\mathrm{Bin}(B, q) > k\}/q$. The second term uses the sign at the end of the paired prefix. Rejected tags consume queries while leaving the sign statistic unchanged.

Bounds for every decision rule. Let $D_{c,g}$ be the KL divergence between the complete recorded (pair scores, tag) laws in the two orderings, for channel c and gap g. Both coordinates can be supplied every round independently of which action serves. At request $n + 1$, any decision rule therefore has average error across the two orderings at least $(1 - \mathrm{TV}(Q^{\otimes n}_{c,g,+}, Q^{\otimes n}_{c,g,-}))/2$. Let a_{Hel} denote their one-request Hellinger affinity. Then

$$1 - \mathrm{TV}(Q^{\otimes n}_{+}, Q^{\otimes n}_{-}) \geq \tfrac{1}{2} a^{2n}_{\mathrm{Hel}} \geq \tfrac{1}{2} e^{-nD_{c,g}}.$$

The last inequality follows from $D_{c,g} \geq -2 \log a_{\mathrm{Hel}}$. Thus every decision rule has worst-ordering cumulative regret at least

$$L_{c,g}(T) = \frac{g}{4} \sum_{n=0}^{T-1} e^{-nD_{c,g}}. \tag{112}$$

The lower bound also holds with the gap supplied to the rule. Let $\mathcal{G}_{\mathrm{ff}}$ denote the common gap grid. Maximizing over this grid gives the finite-family bracket $[\max_{g \in \mathcal{G}_{\mathrm{ff}}} L_{c,g}(T),\ \max_{g \in \mathcal{G}_{\mathrm{ff}}} R^{\mathrm{EO}}_{T,T}]$, with the same EO policy used across gaps and orderings.

For a tag stratum with component masses a, b, its off-diagonal mass and drift are $q = 2ag + b(3/8 + g/2)$ and $d = (a + b)g$. Its KL contribution is $d \log \frac{q+d}{q-d}$. Summing this expression over the two tags gives $D_{c,g}$, with zero-mass strata omitted. For the conservative tag its clean stratum contributes $(g/4) \log 3$. Both aggressive-tag strata have strictly positive disagreement floors and quadratic KL, as do both strata of the symmetric control. Algebraic mixture inversion, including $P = (3Q^0 - Q^1)/2$ for the aggressive tag, is included in the complete recorded-feedback law used by the decision lower bound.

Evaluation grid. The common grid contains 148 gaps: 129 logarithmically spaced values between $1/(64 \cdot 2^{20})$ and $1/8$, augmented by $\min\{1/8, c/T\}$ and $\min\{1/8, c/\sqrt{T}\}$ for every declared horizon and $c \in \{1/4, 1/2, 1, 2, 4\}$, with duplicates removed. The seven horizons are $2^8, 2^{10}, \ldots, 2^{20}$. All 5180 channel–gap–horizon risks are evaluated. An additional fixed 81-gap grid displays recorded-feedback KL. All channel comparisons and both sides of the finite-family bounds use the same gap set.

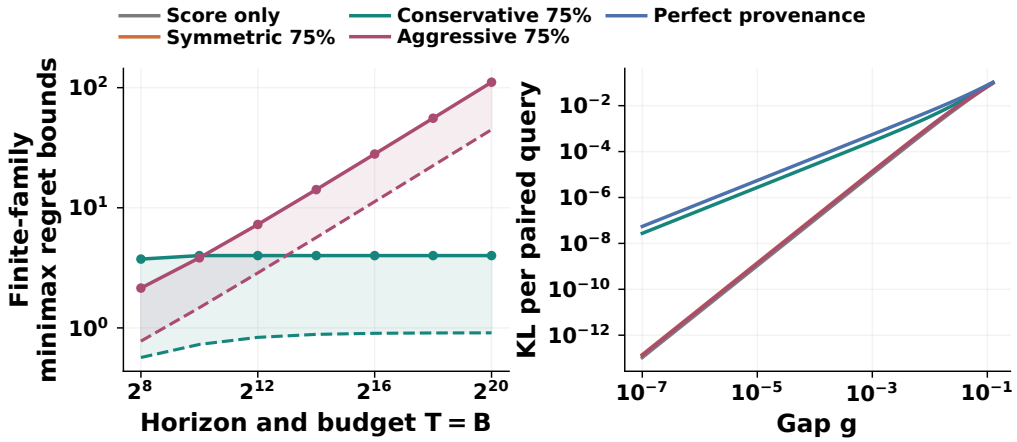

Figure 8: Equal aggregate quality, different decision complexity.

For the risk sum, let $D_{\mathrm{walk}} = -\frac{1}{2}\log(1-(d_{\mathrm{acc}}/q_{\mathrm{acc}})^2)$, so $h_k \le e^{-kD_{\mathrm{walk}}}$. Truncation after index J omits at most

$$Tg\min\{e^{-(J+1)D_{\mathrm{walk}}}, \mathbb{P}(\mathrm{Bin}(B, q_{\mathrm{acc}}) > J)\}.$$

Each of the T decision errors beyond that cutoff obeys either bound, and prefix binomial counts are stochastically dominated by the count at B. We choose the first J for which the displayed remainder is at most 10^{-10} and add this remainder to the policy upper bound. Finite-family lower sums use the corresponding geometric-series expression.

At $T = 2^{20}$ the finite-family bounds are $[0.91, 4.00]$ for conservative 75% provenance, $[44.72, 111.03]$ for aggressive 75% provenance, $[43.35, 117.97]$ for the symmetric 75% control, $[0.46, 2.00]$ for perfect provenance, and $[49.95, 117.65]$ for score-only feedback. These expectation values use the stated truncation control.

Figure 8 separates aggregate tag quality from task-relevant decision information. The left panel is the equal-accuracy-and-$I(F;C)$ comparison: the conservative and aggressive 75% channels have the same tag accuracy and the same mutual information with the contamination indicator, while their finite-family decision complexity differs sharply. Solid curves are the exact EO policy upper bounds, dashed curves are estimator-independent information lower bounds based on the complete recorded score/tag law, and the shaded regions are the resulting finite-family brackets. The right panel reports the information in recorded feedback, measured by KL per paired query, on the common 81-gap grid for all five channels.