ACCESS AND REPRESENTATION BARRIERS TO AUDITING UNDER PROXY-SCORE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models are optimized against cheap proxy scores, but how costly is it to audit the resulting target change? We identify access and representation barriers linked by likelihood-ratio geometry. Baseline archives face a no-hit lower bound, while active auditors trade observable-transcript KL against sample count. An exact L^1 duality characterizes worst-case target change hidden from a fixed feature span. For exact single-prompt one-sided exponential-power tilts, minimax log archive complexity and half-mass likelihood-ratio-tail rarity asymptotically match optimization KL. Using exact conditional reweighting on a finite pool of $207{,}872$ judge-scored Qwen2.5-7B math traces, we find a 2.56-order increase in the tested archive requirement, seven practical samplers above the movement–sample benchmark, and substantial held-out error in fitted score-blind monitors. At 18 score-matchable Group Relative Policy Optimization (GRPO) checkpoints, the plug-in KL residual from a score-matched one-parameter tilt accounts for 96.7–99.8% of measured KL, indicating substantial departure from that tilt path. Reported LR-tail estimates differ substantially from score-tail rarity, although differing likelihood-evaluation paths and a finite reference pool limit their interpretation. This motivates measuring LR-tail overlap under a common likelihood convention before using proxy-score tails as overlap diagnostics.

1 INTRODUCTION

Language models are optimized against cheap *proxy scores* from judges, verifiers, and reward models (Christiano et al., 2017; Stiennon et al., 2020; Bowman et al., 2022; Zheng et al., 2023; Son et al., 2024; Cobbe et al., 2021; Lightman et al., 2024). These signals support reinforcement learning from human feedback, direct preference optimization, GRPO, and best-of-n selection (Ouyang et al., 2022; Rafailov et al., 2023; Shao et al., 2024; Beirami et al., 2025; Huang et al., 2025), but proxy improvement need not imply target improvement (Everitt et al., 2021; Gao et al., 2023; Skalse et al., 2022; Eisenstein et al., 2024; Khalaf et al., 2025; Pan et al., 2024). We ask: *after optimization, how costly is it to determine what changed?* An *audit* here estimates target change statistically, rather than searching for hidden objectives (Marks et al., 2025).

Consider estimating accuracy change after optimization from a labeled baseline archive, baseline-model probes, or unlabeled optimized outputs. Fresh gold labels from both policies would make this routine. Otherwise, two barriers arise (Figure 1). *Access* asks whether data reach behavior made common by optimization (Song et al., 2024; Tajwar et al., 2024). *Representation* asks whether monitor features express the induced shift (Jones et al., 2025; Baker et al., 2025; Chen et al., 2025; Anwar et al., 2026).

Let μ and q denote the baseline and optimized episode laws, and let $W = dq/d\mu$ be their likelihood ratio. A high-ratio region carrying half of q but baseline mass e^{-r} requires roughly e^r baseline draws to encounter. For target classes containing two label worlds that agree off this region but whose target quantities differ by more than 2ϵ, this yields fixed-confidence worst-case lower bounds $\log n \gtrsim r$ for passive access and $K + \log n \gtrsim r$ for active access. Here K measures how far the auditor's observable query–label transcript moves from iid baseline querying. The centered ratio $g = W - 1$ determines worst-case target change invisible to monitor features.

The no-hit bound, active access law, and representation identity require no exponential-power tilt. For a single one-sided exponential-power score law under exponential reweighting, rarity matches

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.

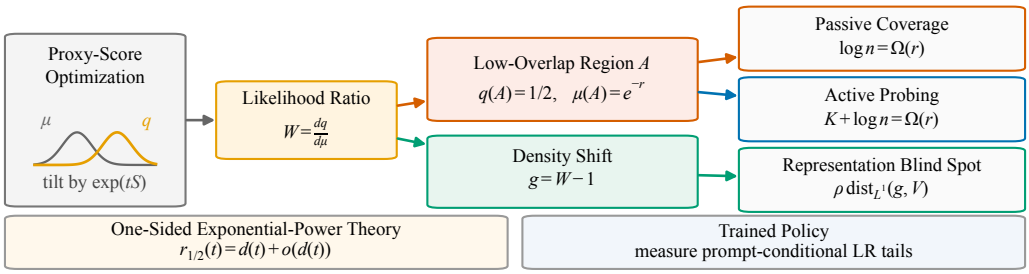


Figure 1: Likelihood-ratio geometry of post-optimization auditing. Under baseline law μ and optimized law q, a high-ratio region A with $q(A) = 1/2$ and $\mu(A) = e^{-r}$ controls coverage. The centered ratio $g = dq/d\mu - 1$ controls representation. Here K is observable-transcript KL, n the query count, V the monitor feature span, and ρ the target-perturbation scale. The schematic suppresses fixed-confidence constants. The displayed calibration holds in the single-prompt one-sided exponential-power model and the extensions in Appendix A, not for arbitrary prompt mixtures or trained policies.

optimization KL at leading order: $r_{1/2}(t) = d(t)+o(d(t))$. We separately use pool-exact conditional reweighting of 207,872 math traces as a controlled finite-pool testbed, not a test of this asymptotic identity, and examine departures from the score-tilt path using GRPO.

Contributions.

1. **Access laws (§4).** For target classes containing separated label worlds that agree off a hard region, we prove a no-hit law and an observable-transcript active trade-off. Exact single-prompt exponential-power tilts yield $\log n^*_{\text{base}} = d+o(d)$, and an idealized likelihood-ratio oracle gives a conditional benchmark.
2. **A representation law (§5).** For any closed monitor subspace of $L^1(\mu)$ containing the constants, an exact duality characterizes the largest bounded hidden target change. We derive a statistic-sufficiency criterion and Gaussian polynomial degree $\Theta(d)$.
3. **Language-model-scale tests (§6).** Pool-exact conditional reweighting tests both barriers in a controlled finite-pool setting. Our GRPO study compares trained policies with score-matched tilts and examines whether proxy-score tails track estimated likelihood-ratio-tail overlap.

2 Related Work

Proxy overoptimization and Goodhart effects. Prior work studies reward tampering, reward gaming, and proxy improvement without target improvement under stronger optimization (Everitt et al., 2021; Skalse et al., 2022; Gao et al., 2023; Huang et al., 2025; Khalaf et al., 2025; Pan et al., 2024; Eisenstein et al., 2024). We study the resources needed to measure target change reliably after optimization.

Coverage, off-policy evaluation, and preference optimization. Coverage limits off-policy evaluation, batch reinforcement learning, and preference optimization (Song et al., 2024; Tajwar et al., 2024; Munos & Szepesvári, 2008; Chen & Jiang, 2019; Metelli et al., 2018). Reweighting (Dudík et al., 2011; Thomas & Brunskill, 2016) and rare-event sampling (Rubino & Tuffin, 2009) address related distribution shifts. No-hit lower bounds (Li et al., 2015) and divergence-dependent importance-sampling requirements (Chatterjee & Diaconis, 2018; Sanz-Alonso, 2018) are established, including exponential dependence on KL under concentration of the log likelihood ratio. Recent work further studies weighted empirical approximations under exponential tilting (Iyer et al., 2026). Within this literature, our passive result studies high-probability minimax estimation over a Lipschitz target class in a fixed label experiment and calibrates fixed-target-mass overlap under exact single-prompt exponential-power tilts. Our general active law charges the auditor's observable transcript through KL (Section 4).

Monitoring, scalable oversight, and audits. Related work studies oversight (Bowman et al., 2022; Burns et al., 2024; Greenblatt et al., 2024), chain-of-thought faithfulness (Lanham et al., 2023; Turpin et al., 2023; Chen et al., 2025), monitorability (Baker et al., 2025; Anwar et al., 2026), and hidden-objective audits or rare-behavior forecasting (Marks et al., 2025; Jones et al., 2025). Balancing-weight methods optimize reweighting estimators under outcome-class assumptions, with duality relating balance to worst-case bias (Bruns-Smith & Feller, 2022). Our representation result instead fixes the density shift and feature span. Through L^1 approximation of the centered density ratio, it characterizes the worst-case bounded target change preserving baseline feature–target moments. A separate corollary gives the corresponding sufficiency criterion for an arbitrary recorded statistic.

3 Post-Optimization Auditing: A Likelihood-Ratio View

To formalize both barriers with one set of objects, we treat optimization as a change from one episode distribution to another.

Distribution shift and audit target. An episode $O \in \mathcal{O}$ includes its prompt, generated trace, answer, and recorded diagnostics. Technical measurability, boundary, and oracle-model details are deferred to Appendix A. Let μ and q be the episode distributions before and after optimization. We assume $q \ll \mu$: every measurable set with zero baseline probability also has zero probability under q. We then define the likelihood ratio $W = dq/d\mu$ (discretely, $W(o) = q(o)/\mu(o)$). The costly gold label is $Y \in [0, 1]$, with target curve $m(O) = \mathbb{E}[Y \mid O]$. We assume that the conditional distribution of Y given O is unchanged by policy optimization, so the same m applies under μ and q. This holds whenever the label is determined by the episode, as for answer correctness. The auditor estimates the optimized mean $R_q(m) := \mathbb{E}_q[m]$ and its shift $\Delta_q(m) = R_q(m) - \mathbb{E}_\mu[m]$ to error ϵ and failure probability δ. Defining the *density shift* $g = W - 1$ and using $R_q(m) = \mathbb{E}_\mu[Wm]$ gives

$$\Delta_q(m) = \mathbb{E}_\mu[g\, m]. \tag{1}$$

Thus g is the shift that interacts with the target. We write $D(\cdot\|\cdot)$ for the Kullback–Leibler (KL) divergence. All logarithms are natural.

The canonical low-overlap region. Among all tests $0 \le \varphi \le 1$ carrying half of q, thresholding the likelihood ratio (LR) W minimizes $\mathbb{E}_\mu[\varphi]$. This is the relaxed Neyman–Pearson construction. When the threshold has no atom, the optimizer is an event A with $q(A) = \frac{1}{2}$. Write $p = \mu(A)$ and $r = -\log p$. Our event-based theorems use this case. Prompt-conditional comparisons average baseline test mass over prompts. Fractional boundaries give exact half-mass tests, whereas inclusive E2 and estimated trained-policy thresholds may not (Section 6 and Appendix F).

Audit resources. An *archive auditor* receives n labeled episodes from μ. An *active auditor* chooses queries adaptively, and K is the KL divergence of its observable query–label transcript from the corresponding transcript generated by n iid baseline queries in the same label experiment. Fresh sampler randomness and intermediate candidates that are discarded and unavailable to the estimator or later query decisions are marginalized. Any state or candidate that is retained, revealed, logged, or used later is part of the transcript and is charged. A *monitor* sees optimized episodes without gold labels through a statistic T or feature span V.

Exact proxy-score tilts. Write an episode as prompt $X \sim \pi$ and output $Z \sim \mu_X$, and let $S(X, Z)$ be the proxy. For a KL-penalty coefficient $\lambda > 0$, the unrestricted conditional optimization problem at fixed π, $\max_q \mathbb{E}_q[S] - \lambda \mathbb{E}_X D(q_X\|\mu_X)$, has the conditional solution below with tilt parameter $t = 1/\lambda$, provided the normalizers and objective are finite:

$$\frac{dq_{t,x}}{d\mu_x}(z) = e^{tS(x,z)-\Lambda_x(t)}, \qquad \Lambda_x(t) = \log \mathbb{E}_{\mu_x} e^{tS}. \tag{2}$$

Here $\Lambda_x(t)$ normalizes the distribution and larger t gives more weight to high scores. We reserve β for GRPO's KL coefficient. Its surrogate need not produce this exact tilt (Section 6). The prompt law stays fixed. Define the optimization KL $d(t) = \mathbb{E}_X D(q_{t,X}\|\mu_X)$. Since $\log W = tS - \Lambda_X(t)$, every exact-tilt LR threshold is a within-prompt high-score tail, although its exact mass in a discrete pool depends on the boundary rule. Moreover, $g = e^{tS-\Lambda_X(t)} - 1$ depends only on (X, S).

Throughout, d denotes optimization KL, K total audit-transcript KL, k_ℓ marginal KL per oracle block, and $\hat{d}$ the estimated GRPO policy KL.

4 The Access Barrier

4.1 Passive Archives

How large must a baseline archive be before it can detect the target shift? If two target worlds differ only on A, their archived data are identical until the archive hits A, which occurs per episode with probability $p = e^{-r}$. Fix either the noiseless or conditionally independent Bernoulli label experiment. The no-hit argument uses no tail model. Part (a) gives this general distribution-shift bound in the fixed experiment. Part (b) gives a matching rate in the exact single-prompt one-sided exponential-power model. There the target is summarized by the score-conditional curve $f(s) = \mathbb{E}_\mu[Y \mid S = s]$. Let $\mathcal{F}_L$ be the $[0,1]$-valued L-Lipschitz curves. Then $n^*_{\text{base}}(t, \epsilon, \delta; \mathcal{F}_L)$ is the smallest archive size for which a possibly randomized estimator estimates the shift to error ϵ with failure probability at most δ, uniformly over $\mathcal{F}_L$, in that fixed experiment (formal definition in Appendix A).

Theorem 1 (Passive access law)**.**
(a) (No-hit bound) Fix $\delta \in (0, \frac{1}{2})$ and one of the two label experiments above. Suppose two target worlds have curves m_0, m_1 that agree outside a set A with $\mu(A) = p$, and $|\Delta_q(m_1) - \Delta_q(m_0)| > 2\epsilon$. Then any archive auditor whose error is at most ϵ with probability at least $1 - \delta$ in both worlds satisfies $(1-p)^n \le 2\delta$. If $p = 0$, no finite n can satisfy both success requirements. If $0 < p \le \frac{1}{2}$, then $n \ge \frac{1}{2p} \log \frac{1}{2\delta}$.
(b) (One-sided exponential-power tilt) Fix $\alpha > 1$, $L \in (0, \infty)$, and $\delta \in (0, \frac{1}{4})$. Let P be the baseline score law with density $p_\alpha(s) = Z_\alpha^{-1} e^{-s^\alpha/\alpha} \mathbf{1}_{\{s \ge 0\}}$, where Z_α is the normalizing constant, and let Q_t be its exponential tilt. Write $d(t) = D(Q_t \| P)$ and $r_{1/2}(t)$ for the half-mass low-overlap rarity of Q_t. There exists $\epsilon_0 = \epsilon_0(L, \alpha) > 0$ such that, for every fixed $\epsilon \in (0, \epsilon_0)$, as $t \to \infty$,

$$\log n^*_{\text{base}}(t, \epsilon, \delta; \mathcal{F}_L) = d(t) + o(d(t)),$$
$$r_{1/2}(t) = d(t) + o(d(t)),$$

Thus each additional nat of optimization KL adds one nat to log archive size at leading order. Here $d(t) = \frac{1}{\alpha} t^{\alpha/(\alpha-1)} + o(t^{\alpha/(\alpha-1)})$. The matching estimator bins the score range carrying most of q_t and estimates the target curve from baseline samples in those bins. The limit keeps ϵ fixed and sufficiently small.

As a control, suppose the auditor can draw fresh labeled samples from both distributions.

Proposition 1 (Direct two-policy control)**.** *For $n_\mu, n_q \in \{1, 2, \ldots\}$, take independent labeled samples of sizes n_μ and n_q from μ and q, respectively. With bounded labels, the plug-in estimator of $\Delta_q(m)$ has error at most ϵ with product-experiment probability at least $1 - \delta$ whenever $n_\mu, n_q \ge 2\epsilon^{-2} \log(4/\delta)$, independent of r.*

Thus estimating two means is easy once both policies can be sampled. The difficulty is baseline coverage. Theorem 1 counts fully labeled archives. If generation and proxy scoring are cheap, selective labeling moves the exponential factor from purchased labels to generated-and-scored episodes, so the invariant claim is a coverage cost.

4.2 Active Probing

Can distributional movement substitute for samples? An active auditor can query A more often, but pays observable-transcript KL K from baseline sampling. For the lower bound, let $q(A) = \eta$ and compare $m_0 \equiv 1$ with $m_1 = 1 - b\mathbf{1}_A$. Their optimized means differ by $b\eta$, yet their full conditional label laws agree until the first query in A.

Theorem 2 (Active access law)**.** *Fix an event A with $q(A) = \eta \in (0,1)$ and $p = \mu(A) \in (0,1)$, and let m_0, m_1 be the two target curves above for some $0 < b \le 1$. Fix a label experiment with conditional means m_0, m_1 whose full conditional laws agree on A^c, and use that same experiment for the active and iid-baseline laws. Let a common nonanticipating audit strategy make a positive*

integer number n of queries. Let $\mathbb{P}_m$ be its observable query–label transcript law in world m, under the convention above, and let $\mathbb{M}_m$ be the corresponding transcript law from n iid μ-queries. Suppose $D(\mathbb{P}_m\|\mathbb{M}_m) \le K$ in both worlds. If an estimator $\widehat{R}$ of $R_q(m)$ has error at most ϵ with probability at least $1-\delta$ in both worlds, for some $\epsilon < b\eta/2$ and $\delta \in (0, \frac{1}{2})$, then

$$\frac{K}{1-2\delta} + \log n \;\ge\; \log\frac{1}{p} - \frac{\log 2}{1-2\delta}. \tag{3}$$

The bound holds even if the strategy knows A.

An accurate estimator must make a hit likely. Data processing yields the KL term in (3). Up to δ-dependent constants, one nat of movement replaces at most one nat of log sample size. For Δ_q, the same pair is separated by $b|\eta - p|$, and the bound holds when $\epsilon < \frac{b}{2}|\eta - p|$. Noiseless and Bernoulli labels are examples of the permitted label experiments. Appendix A separately gives the exact *expected*-risk common-query oracle. It is not the high-probability result above.

When the baseline score is $N(0,1)$, its tilt is $N(t,1)$, $d = t^2/2$, and the median-tail rarity is $d + \frac{1}{2}\log d + O(1)$. The same leading trade-off holds for smooth Lipschitz targets (Theorem A.2), so $K = o(d)$ forces $n \ge e^{d-o(d)}$. Proposition 1 differs because it grants direct sampling from q_t.

Information-theoretic benchmark. Fix an integer block length $\ell \ge 2$ and assume $d = D(q\|\mu) < \infty$. An idealized sampler draws one uniformly chosen coordinate from q and the other $\ell - 1$ from μ, then privately discards the index. Under the policy-invariant conditional label law from Section 3, Theorem A.1 gives $[d - \log \ell]_+ \le k_\ell \le \mathbb{E}_q \log(1 + (W-1)/\ell)$. If $d \to \infty$ and $\mathbb{E}_q|\log W - d| = o(d)$, then $k_\ell = [d - \log \ell]_+ + o(d)$ uniformly over block-size sequences. Revealing the index instead costs exactly d per block. E2 also plots $\ell = 1$ as a direct-q KL/hit reference. The shift estimator below requires $\ell \ge 2$.

For $W_i = W(O_i)$, $Y_i \in [0,1]$, and $\bar{Y}_\ell = \ell^{-1}\sum_i Y_i$, the estimator

$$\widehat{\Delta}_\ell = \frac{\ell}{\ell - 1}\left(\frac{\sum_i W_i Y_i}{\sum_i W_i} - \bar{Y}_\ell\right)$$

is unbiased for $\Delta_q(m)$ and satisfies $|\widehat{\Delta}_\ell| \le 2$. The denominator is positive almost surely under the oracle law. Averaging $B \ge \lceil 8\epsilon^{-2}\log(2/\delta)\rceil$ independent blocks gives error at most ϵ with probability at least $1-\delta$, using $N = B\ell$ labels and $K_{\mathrm{obs}} = Bk_\ell$.

5 The Representation Barrier

Without gold labels, a monitor must represent how probability moved. Let $V \subset L^1(\mu)$ be any closed linear subspace containing the constants. A finite set of recorded features $v_1, \dots, v_k$ induces the linear span $V = \mathrm{span}\{1, v_1, \dots, v_k\}$. For $0 \le \rho \le \frac{1}{2}$, define $\mathcal{M}_{\rho,V} = \{m = \frac{1}{2} + \rho h : \|h\|_\infty \le 1,\ \mathbb{E}_\mu[vh] = 0\ \forall v \in V\}$. The bound on h keeps $m \in [0,1]$. Orthogonality makes its baseline feature–target moments identical to those of $m \equiv \frac{1}{2}$. Such a perturbation is invisible to V. Define the best representation error $\mathrm{dist}_{L^1(\mu)}(g, V) = \inf_{v\in V}\mathbb{E}_\mu|g - v|$.

Theorem 3 (Representation duality)**.** *The extremal shifts satisfy* $\sup_{m\in\mathcal{M}_{\rho,V}}|\Delta_q(m)| = \rho\ \mathrm{dist}_{L^1(\mu)}(g,V)$ *and* $\inf_{m\in\mathcal{M}_{\rho,V}}\Delta_q(m) = -\rho\ \mathrm{dist}_{L^1(\mu)}(g,V)$*. Both are attained.*

By L^1/L^∞ duality, approximation error is exactly worst-case hidden target change. Thus, in this bounded worst-case invisibility model, baseline prediction or proxy correlation is insufficient unless the feature span represents g.

Corollary 1 (Statistic sufficiency)**.** *For a recorded statistic T, let $\mathcal{G} = \sigma(T)$ be the σ-algebra generated by T. For $0 < \rho \le \frac{1}{2}$, the same identity holds with $V = L^1(\mathcal{G})$ and invisibility $\mathbb{E}_\mu[h \mid T] = 0$. At any such scale, T precludes nonzero target shifts from bounded T-invisible perturbations if and only if g is $\sigma(T)$-measurable up to μ-null sets. The case $\rho = 0$ is trivial and is not used for the if-and-only-if statement.*

Under (2), (X, S) determines g and is therefore sufficient in this model, whereas a score-blind monitor may fail. This conclusion need not extend to trained policies, whose density shift may depend on additional episode information.

The duality theorem itself requires no Gaussian, exponential-power, or tail assumption. For such V—indeed, for any linear feature subspace—define its *leakage* into A as the largest fraction of a feature combination's absolute mass concentrated there, $\beta_V(A) = \sup_{v \in V, v \neq 0} \mathbb{E}_\mu[|v|\mathbf{1}_A]/\mathbb{E}_\mu|v|$. Then

$$\mathrm{dist}_{L^1(\mu)}(g, V) \;\geq\; \frac{[\,q(A) - \mu(A) - 2\beta_V(A)\,]_+}{1 + \beta_V(A)}, \tag{4}$$

where $[u]_+ = \max\{u, 0\}$. If $q(A)$ stays constant while $\mu(A)$ and $\beta_V(A)$ vanish, the bound stays large. For Gaussian polynomial features, hypercontractivity (O'Donnell, 2014) bounds leakage and yields:

For a Gaussian tilt with $d = t^2/2$, approximating g by score polynomials to any fixed L^1 tolerance $\tau \in (0, 1/4)$ requires and suffices to use degree $\Theta(d)$. Appendix A.4 states the precise bounds and their qualifications. Thus structured degree, not raw bit count, limits this family of monitors.

6 Empirical Evaluation

We organize the analyses as E1–E5. E1–E3 use known exact-tilt ratios to test access and representation. E4–E5 stress-test overlap diagnostics on GRPO.

6.1 Exact Conditional-Tilt Testbed

For each prompt, assign baseline trace i weight $w_i = e^{tS_i} / \sum_j e^{tS_j}$. Sweeping t gives empirical $d(t)$, W, g, and tail events without reinforcement-learning noise. These quantities are *pool-exact*: exact for the finite sampled pool, not the underlying policy. We report the prompt-level effective sample size $\mathrm{ESS} = 1/\sum_i w_i^2$, bootstrap intervals, and the finite-pool KL ceiling $\log N$.

Pool, proxy, and label. Using vLLM (Kwon et al., 2023), we sample Qwen2.5-7B-Instruct (Yang et al., 2024) at temperature 1 and top-$p = 1$: $N = 1024$ traces for each of 203 level-4/5 numeric-answer problems from MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024), totaling 207,872. The same Qwen2.5-7B-Instruct model serves as the judge, so the setup is self-judging. It sees the problem and trace without the answer key and returns Yes/No log-odds S, clipped to ± 16. Only 0.81% hit the clip. The gold label Y is extracted-answer correctness. Baseline accuracy is 0.571 and the judge's receiver operating characteristic area under the curve (ROC-AUC) is 0.851, yet the top score decile has error 0.062 (0.136 on level 5), leaving an exploitable gap. Hard MATH avoids the near-degenerate gap in the easier mixed pool (Appendix E). Full experimental details are in Appendix F, and synthetic illustrations of both barriers are in Appendix A.

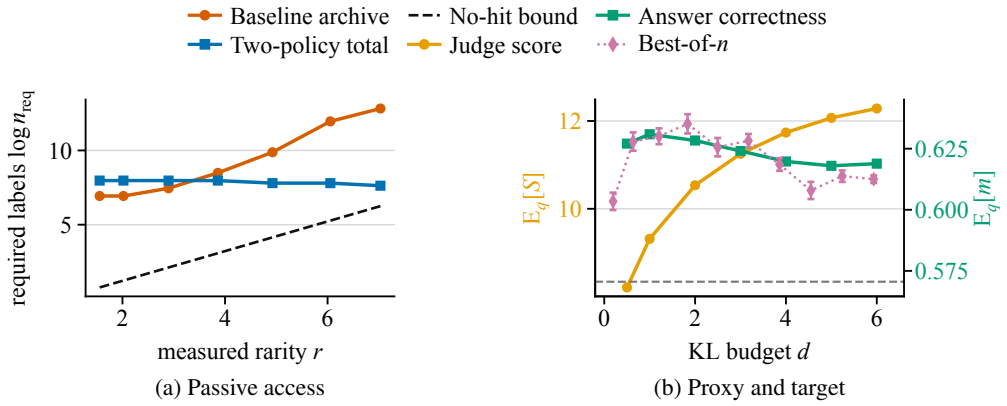
6.2 Access Barriers at Language-Model Scale (E1, E2)

E1: does rarity predict passive archive size? Figure 2 reports passive access and target behavior. For $d \in \{0.5, 1, 2, 3, 4, 5, 6\}$, a prompt-conditional binned estimator reweights baseline labels to q_t. Global bins would confound problem difficulty with within-prompt optimization. We define n_{req} as the smallest tested size with an observed success rate of at least $1 - \delta = 0.8$ at error tolerance $\epsilon = 0.025$ across 500 replications. A direct control labels both μ and q_t.

Plotted against measured r, the archive requirement rises from 1,024 episodes at $d = 0.5$ to 370,728 at $d = 6$ (2.56 orders), while direct access uses 2,048–2,896 total labels. At small d, the archive can be cheaper because q_t remains well covered.

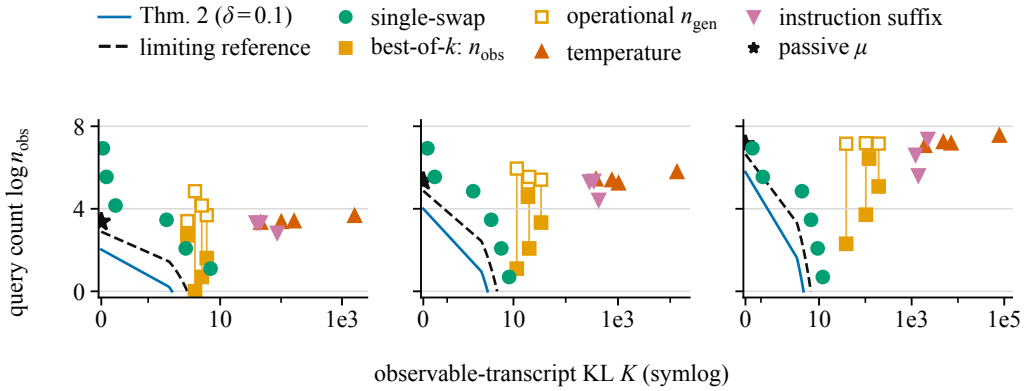
Beyond the pool KL ceiling we report targets as unreachable. The same growth appears against d over this finite experimental range. This empirical alignment is not the single-prompt asymptotic identity. The secant slope 1.34 over $d = 2$ to $d = 6$ is a property of this estimator, not an estimate of the asymptotic coefficient. Writing $\Delta_t := \Delta_{q_t}$, correctness shifts remain nontrivial, $\Delta_t(m) \in [0.047, 0.060]$, but correctness is one realized target whereas the theorem is worst-case over a class.

Judge overoptimization. As d grows, $\mathbb{E}_{q_t}[S]$ rises from 8.2 to 12.3, while accuracy peaks at 0.631 near $d = 1$ and falls to 0.619 at $d = 6$. A weaker judge yields a negative shift (Appendix B). Other tested score families show a similar finite-pool rarity–KL pattern. The asymptotic theorem applies



(a) Passive access (b) Proxy and target

Figure 2: E1 on MATH level 4–5. (a) The natural logarithm of the smallest tested archive size meeting the target error and confidence grows with measured rarity, while direct two-policy labeling uses a nearly constant total label count. The dashed no-hit line is a worst-case lower bound, not a prediction for realized correctness. (b) Proxy score rises with KL, whereas accuracy peaks and then declines. Best-of-n is placed using its analytic induced-policy KL bound. Bars show ± 1 standard error (SE).



(a) $d = 2$; $r = 2.9$ (b) $d = 4$; $r = 4.9$ (c) $d = 6$; $r = 6.6$

Figure 3: E2: samplers against the observable-transcript movement–count benchmark. Panels (a)–(c) use $d = 2, 4, 6$, with rarity r below each panel. Solid blue: Theorem 2 at E2's $\delta = 0.1$, $\log n_{obs} = r - (K + \log 2)/(1 - 2\delta)$. Dashed: the limiting reference $K + \log n_{obs} = r$. The horizontal axis is symlog. Ordinary samplers count generated-and-scored draws. Filled best-of-k squares use retained count n_{obs}, whereas open squares use operational count $n_{gen} = kn_{obs}$. Their horizontal coordinate is the analytic KL-budget upper bound $n_{obs}[\log k - (k-1)/k]$, not a measured divergence. Private-index single-swap points likewise use a conservative block-KL upper bound and are oracle references, not practical guarantees.

only under its displayed tail regularity assumptions (Appendices B–D). Geometry determines how hard a change is to find. Alignment determines its direction.

E2: how do practical samplers compare with the active benchmark? Figure 3 evaluates four temperatures and three instruction suffixes in the $(K, \log n_{obs})$ plane, where n_{obs} is the retained observable count. A suffix belongs to the sampling policy, not the episode: for each problem X, ν_X and μ_X share the trace space, and $\log \mu$ is obtained by rescoring the generated tokens under the original prompt. The empirical mean $\widehat{\kappa}$ of $\log \nu_X(z) - \log \mu_X(z)$ estimates per-draw cost. E2's inclusive upper-score events have realized q_t-masses 0.510, 0.586, and 0.783 at $d = 2, 4, 6$, not $1/2$. Their hit-to-confidence counts give $K = n_{obs}\widehat{\kappa}$. Best-of-$k$ uses retained count n_{obs} but operational count $n_{gen} = kn_{obs}$. Its plotted movement $K = n_{obs}[\log k - (k-1)/k]$ is an analytic KL-budget upper bound, not an exact measured divergence. A single-swap block of length ℓ charges all ℓ queried

and labeled objects. We plot the looser conservative upper bound $\mathbb{E}_q \log(1 + W/\ell)$ on its marginal block KL. Its discount additionally assumes direct q-sampling, likelihood-ratio access, and a private discarded index. At these plotted budget coordinates, the practical points remain above the limiting reference. The oracle points report hit-confidence counts, not empirical sample requirements for the uniformly accurate shift estimator above.

6.3 Representation in Trace Monitors (E3)

Can score-blind monitors represent g? We compare fitted reconstruction errors with certified feature-span gaps. We measure fitted reconstruction by normalized L^1 error, $\mathbb{E}_\mu|g - \widehat{g}|/\mathbb{E}_\mu|g|$. Although (X, S) determines g, we hide S and fit surface-feature polynomials, probes and multilayer perceptrons (MLPs) on frozen Qwen2.5-1.5B embeddings, and a Low-Rank Adaptation (LoRA)-fine-tuned (Hu et al., 2022) 1.5B model. A score control additionally receives S, while prompt-visible variants additionally receive X. The full fitted ladder is supplementary. Across the fitted score-blind ladder, including the constant baseline, relative errors range from 0.466 to 0.882. At $d = 1$, the fine-tuned language model has 0.656 error versus 0.254 for the score control. At high d, even the control degrades as the pool effective sample size falls.

The nonlinear values describe fitted models, not whole classes. For the degree-1 and degree-2 surface-feature spans, the empirical L^1 LP primal and dual agree to numerical tolerance. The degree-3 feasible dual gives a nearly unchanged lower bound (Table 2). The certified projection ratios are 0.818, 0.706, 0.578, and 0.520 at $d = 1, 2, 4, 6$. Because they restrict full-pool g to the monitor subset without renormalizing its likelihood ratios, they certify approximation but not exact target shifts for a normalized subset law. They do not imply that actual correctness is invisible or that all nonlinear monitors fail.

At low d, the score-control gap is consistent with missing score information and finite-pool estimation error. Giving the monitor the prompt produces statistically detectable but small reductions. The tested prompt-visible fits do not close the empirical reconstruction gap (Appendix B). For a trained policy, the density shift may depend on information beyond (X, S), motivating direct likelihood-ratio measurement.

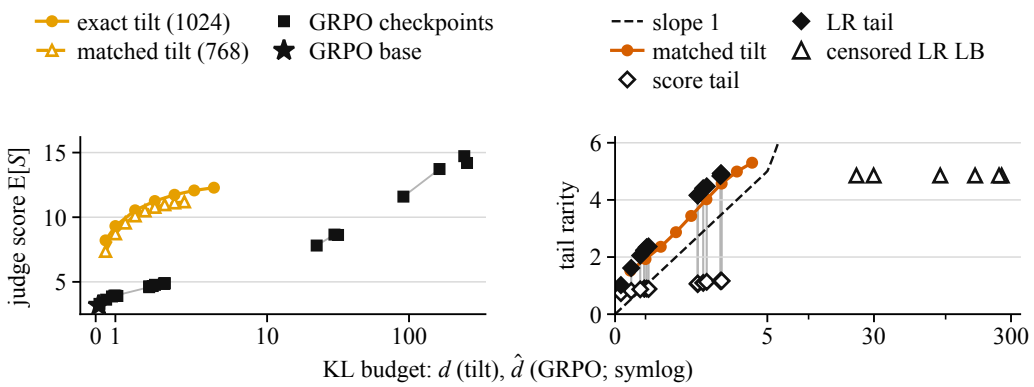
6.4 Direct Overlap Diagnostics for Trained Policies (E4, E5)

Does the tilt calibration transfer to trained policies? Figure 4 compares trained-policy diagnostics with score-matched exact tilts. We evaluate 22 checkpoints from five GRPO (Shao et al., 2024) runs with LoRA, one per configuration (Appendix F), with measured KL $\hat{d} \in [0.19, 256]$, plus a base-policy anchor at 0.16. Here $\hat{d}$ averages sequence log-ratios $\log \pi_\theta(z \mid x) - \log \mu_x(z)$ and has an approximately 0.16-nat measurement floor. Exact tilts are recomputed on the same 768-token baseline pool.

We compare prompt-conditional upper-half score tests with estimated LR-tail tests. Under exact conditional tilts, the population score and likelihood-ratio rankings coincide. For trained policies they need not. Appendix C specifies the fractional-boundary rule, two-fold threshold averaging, and likelihood-evaluation conventions.

The exact tilt reaches $\mathbb{E}[S] \approx 8.7$ at $d = 1$, whereas the closest GRPO checkpoint needs $\hat{d} \approx 30$. Across runs, pooled accuracy peaks at 0.579 ($+0.086$ over the matched baseline 0.493). Some higher-score checkpoints have lower accuracy, although this cross-run comparison does not establish a within-run decline. Because π_θ, q_t, and μ share one prompt law, the exponential-family identity separates measured KL into a matched-tilt term and a residual when mean scores match. At the 18 of 22 checkpoints with a finite-pool score match, the plug-in residual share is 96.7–99.8%, indicating substantial departure from the one-parameter tilt path. The other four lie outside the finite-pool tilt range (Table 3).

At $\hat{d} = 3.48$, fractional score-tail rarity is 1.17, compared with matched-tilt rarity 4.57 at $d = 3.5$ and estimated LR-tail rarity 4.92. The latter is a diagnostic rather than an exact common-sequence comparison because the policy and baseline sides use different tokenization and likelihood-evaluation paths (Appendix C).



(a) Judge score (b) Tail rarity

Figure 4: Trained-policy diagnostics (legend numbers: completion-token caps). Under the matched 768-token data protocol, baseline accuracy is 0.493 and peak GRPO accuracy across runs is 0.579 ($+0.086$). Panel (a) uses a symmetric-logarithmic horizontal axis and shows that GRPO needs substantially more measured KL than the exact tilt for the same mean score. The GRPO-base star appears at the approximately 0.16-nat measurement floor. Panel (b) connects each checkpoint's prompt-conditional fractional-boundary score-tail and estimated likelihood-ratio (LR)-tail rarity. At comparable KL, the reported LR estimates can be comparable to matched exact-tilt rarity. Segments connect uncensored checkpoints. Triangles are right-censored lower bounds ("LR LB") for checkpoints with no baseline-tail hits. Values and likelihood-evaluation caveats are in the supplement.

The matched-tilt residual diagnoses departure from the one-parameter path, but the present LR estimate neither fixes the exact rarity nor shows that LR-tail sampling improves correctness estimation. Nor does the residual prove (X, S) insufficient. That requires measuring the shift after conditioning on (X, S). All measured checkpoints with $\hat{d} \geq 22$ have zero baseline-tail hits and hence only right-censored LR lower bounds, and one adapter was unavailable.

Table 4 reports prompt-clustered paired randomization checks. They are unadjusted descriptive tests, and the GRPO peak comparison is post-selection.

7 Discussion and Limitations

Practical implications. Fixed-policy-mass LR tails diagnose worst-case coverage, while g determines monitor representation. Total KL and proxy-score tails need validated tilt geometry to indicate overlap. Score both likelihoods identically on common observations and report LR-tail baseline mass with uncertainty. For active probes, report K and n. Assess g-representation, not proxy correlation.

Limitations. Our results imply neither universal target harm nor monitor blindness. Under stated assumptions, $r_{1/2}(t) = d(t) + o(d(t))$ holds for Theorem 1(b), the Gaussian example, and Appendix A's regular-tail models, not arbitrary prompt mixtures or trained policies. Experiments use one self-judging model family, finite clipped-score traces, one E1 estimator, and fitted nonlinear models. Linear-span LPs certify empirical monitor-subset approximation only. Full-pool-normalized ratios preclude normalized-subset target-shift bounds. Under a common prompt law, population prompt-conditional rarity lower-bounds joint rarity at fixed target mass. Estimates are not confidence bounds. Zero-hit LR tails are censored. The GRPO check mixes generation/original-token policy thresholds with retokenized teacher-forced baseline scoring, so it does not estimate a common LR quantity. We use public benchmarks and no human subjects.

AI USE STATEMENT

We used OpenAI GPT-based tools, including Codex, and Anthropic Claude during the research and manuscript preparation. AI assistance supported research ideation and execution, including developing theoretical models and conceptual frameworks, formulating mathematical claims and hypotheses, developing and writing proofs, designing methodology and experiments, implementing methods and code, and preparing analysis data. It also aided literature retrieval and discovery, citation and manuscript organization, scientific figures and LaTeX formatting, and drafting, translating, and polishing sections of the manuscript, including the figure-layout work for this submission.

The authors reviewed all AI-assisted work. They checked mathematical arguments against the stated assumptions and complete derivations, tested the generated code and data pipelines, regenerated or audited tables and figures against frozen artifacts, and cross-checked reported numerical values and citations against their underlying sources. The authors take responsibility for the final content of this work, including text, mathematical claims, code, and results produced with AI assistance.

REPRODUCIBILITY STATEMENT

The formal assumptions, theorem statements, and complete proofs are provided in Appendices A and G and the supporting proof sections. Appendix F documents data selection and preprocessing, model identifiers, generation and scoring protocols, training configurations, hardware and software environments, random seeds, and resampling procedures. Appendices B–E report robustness analyses, trained-policy diagnostics, and finite-pool limitations. The anonymized supplementary code archive contains experiment drivers, environment locks, tests, frozen aggregate and prompt-level analysis artifacts, and manifests for regenerating the reported tables and figures. Large raw completion pools, model checkpoints, and adapters are excluded and can be regenerated from the documented public inputs. Deterministic analysis stages are artifact-reproducible, whereas GPU generation and training are intended to be statistically rather than bitwise reproducible across hardware stacks.

REFERENCES

Usman Anwar, Tim Bakker, Dana Kianfar, Cristina Pinneri, and Christos Louizos. Analyzing and improving chain-of-thought monitorability through information theory. arXiv preprint arXiv:2602.18297, 2026. URL https://arxiv.org/abs/2602.18297.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. arXiv preprint arXiv:2503.11926, 2025. URL https://arxiv.org/abs/2503.11926.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3580–3602. PMLR, 2025.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. arXiv preprint arXiv:2211.03540, 2022. URL https://arxiv.org/abs/2211.03540.

David A. Bruns-Smith and Avi Feller. Outcome assumptions and duality theory for balancing weights. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 11037–11055. PMLR, 2022. URL https://proceedings.mlr.press/v151/bruns-smith22a.html.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4971–5012. PMLR, 2024.

Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018. doi: 10.1214/17-AAP1326.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1042–1051. PMLR, 2019.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. arXiv preprint arXiv:2505.05410, 2025. URL https://arxiv.org/abs/2505.05410.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30*, pp. 4299–4307, 2017.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alexander D'Amour, D. J. Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. Helping or herding? Reward model ensembles mitigate but do not eliminate reward hacking. In *Proceedings of the First Conference on Language Modeling*, 2024.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198 (Suppl 27):6435–6467, 2021. doi: 10.1007/s11229-021-03141-4.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16295–16336. PMLR, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is Best-of-N the best of them? Coverage, scaling, and optimality in inference-time alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25075–25126. PMLR, 2025.

Sarvesh Ravichandran Iyer, Himadri Mandal, Dhruman Gupta, Rushil Gupta, Agniv Bandyopadhyay, Achal Bassamboo, Sandeep Kumar Juneja, and Varun Gupta. Fundamental limits for weighted empirical approximations of exponentially tilted distributions. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 5239–5247. PMLR, 2026. URL https://proceedings.mlr.press/v300/iyer26a.html.

Erik Jones, Meg Tong, Jesse Mu, Mohammed Mahfoud, Jan Leike, Roger Grosse, Jared Kaplan, William Fithian, Ethan Perez, and Mrinank Sharma. Forecasting rare language model behaviors. arXiv preprint arXiv:2502.16797, 2025. URL https://arxiv.org/abs/2502.16797.

Émilie 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. URL https://jmlr.org/papers/v17/kaufman16a.html.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio du Pin Calmon. Inference-time reward hacking in large language models. In *Advances in Neural Information Processing Systems 38*, 2025.

Dmitriy Kunisky. Hypothesis testing with low-degree polynomials in the Morris class of exponential families. In *Proceedings of the Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2822–2848. PMLR, 2021.

Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In Paula Cerejeiras and Michael Reissig (eds.), *Mathematical Analysis, its Applications and Computation*, volume 385 of *Springer Proceedings in Mathematics & Statistics*, pp. 1–50. Springer, Cham, 2022. doi: 10.1007/978-3-030-97127-4_1.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. Association for Computing Machinery, 2023. doi: 10.1145/3600006.3613165.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. arXiv preprint arXiv:2307.13702, 2023. URL https://arxiv.org/abs/2307.13702.

Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 608–616. PMLR, 2015. URL https://proceedings.mlr.press/v38/li15b.html.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, et al. Auditing language models for hidden objectives. arXiv preprint arXiv:2503.10965, 2025. URL https://arxiv.org/abs/2503.10965.

Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems 31*, 2018.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. doi: 10.1017/CBO9781139814782.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*, pp. 27730–27744, 2022. doi: 10.52202/068431-2011.

Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. Feedback loops with language models drive in-context reward hacking. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39154–39200. PMLR, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36*, pp. 53728–53741, 2023. doi: 10.52202/075280-2338.

Gerardo Rubino and Bruno Tuffin (eds.). *Rare Event Simulation Using Monte Carlo Methods*. John Wiley & Sons, 2009. doi: 10.1002/9780470745403.

Daniel Sanz-Alonso. Importance sampling and necessary sample size: An information theory approach. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):867–879, 2018. doi: 10.1137/16M1093549.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024. URL https://arxiv.org/abs/2402.03300.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems 35*, pp. 9460–9471, 2022. doi: 10.52202/068431-0687.

Guijin Son, Hyunwoo Ko, Hoyoung Lee, Yewon Kim, and Seunghyeok Hong. LLM-as-a-judge & reward model: What they can and cannot do. arXiv preprint arXiv:2409.11239, 2024. URL https://arxiv.org/abs/2409.11239.

Yuda Song, Gokul Swamy, Aarti Singh, J. Andrew Bagnell, and Wen Sun. The importance of online data: Understanding preference fine-tuning via coverage. In *Advances in Neural Information Processing Systems 37*, pp. 12243–12270, 2024. doi: 10.52202/079017-0392.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems 33*, pp. 3008–3021, 2020.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47441–47474. PMLR, 2024.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2139–2148. PMLR, 2016.

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

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36*, pp. 74952–74965, 2023. doi: 10.52202/075280-3275.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024. URL https://arxiv.org/abs/2412.15115.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

A TECHNICAL APPENDIX OVERVIEW

This appendix states the deferred conditions and results, gives concise proof roadmaps, and reports the full robustness and trained-policy diagnostic suite, including the optimization-KL E1 view, quantile cross-fitting, same-protocol tilt replication, matched-tilt residual decomposition, empirical projection certificates, likelihood-ratio-tail measurement, and prompt–trace monitors. The final sections contain the publication-locked conservative proof package, including all supporting lemmas, boundary cases, measurable-selection conditions, exact coverage-modulus denominators, observable-transcript conventions, and the distinct observation models used by the high-probability, expected-risk, block-oracle, and randomized-monitor results.

Relation to prior importance-sampling sample-complexity results. Importance-sampling theory already establishes closely related exponential-in-divergence sample requirements. Under concentration of the log likelihood ratio, Chatterjee & Diaconis (2018) show that approximately $\exp(D(q\|\mu))$ proposal samples are necessary and sufficient for accurate importance sampling, including in one-parameter exponential families. Sanz-Alonso (2018) gives complementary nonasymptotic information-theoretic necessary conditions in terms of KL and other divergences. More recently, Iyer et al. (2026) study the accuracy of self-normalized weighted empirical approximations of exponentially tilted distributions across bounded and unbounded regimes. Accordingly, we do not claim the generic leading-order importance-sampling threshold $\log n \asymp D(q\|\mu)$ as new. Theorem 1 instead formulates a different statistical experiment for post-optimization auditing: the gold-label curve is unknown but policy-invariant, complexity is minimax over a Lipschitz target class in a fixed label experiment, the lower bound is a no-hit indistinguishability argument, and the exponential-power result identifies archive complexity with the baseline rarity of a fixed-q-mass likelihood-ratio region. Its upper bound uses a score-binning label estimator rather than asserting universal success of an importance-sampling estimator. The contribution is this audit-specific formulation and its connection to active-access and representation barriers, while the leading exponential overlap phenomenon should be read in the context of the cited importance-sampling literature.

A.1 DEFERRED CONDITIONS AND THE COVERAGE MODULUS

One-sided exponential-power scope of Theorem 1(b). The main theorem uses the baseline density $p_\alpha(s) = Z_\alpha^{-1} e^{-s^\alpha/\alpha} \mathbf{1}_{\{s \ge 0\}}$, $\alpha > 1$, where Z_α is the normalizing constant, and its exact exponential tilt. The broader result below is supplementary-only: on a right-tail interval $[s_0, \infty)$, let the baseline have positive density $f_S = e^{-U}$, where U is eventually C^2, increasing, and strictly convex, $U(s) = s^\alpha/\alpha\,(1 + o(1))$, and $U''(s) \asymp s^{\alpha-2}$. No density assumption is imposed to the left of s_0. Under these displayed assumptions, Supplementary Theorem 5.6 gives the regular stretched-exponential extension.

Target and passive-complexity notation. Let $\mathcal{O}$ be the episode space, write $W = dq/d\mu$, $g = W - 1$, $R_q(m) := \mathbb{E}_q m$, and $\Delta_q(m) = R_q(m) - \mathbb{E}_\mu m$. We use $D(P\|M)$ for KL divergence, $\mathrm{kl}(a\|b)$ for Bernoulli KL, $h_2(a)$ for binary entropy, and $[u]_+ = \max\{u, 0\}$. For the passive no-hit law and minimax phase, fix either the noiseless experiment $Y = f(S)$ or the conditionally independent Bernoulli experiment $Y \mid S = s \sim \mathrm{Bernoulli}(f(s))$. The quantity $n^*_{\mathrm{base}}(P, Q, \epsilon, \delta; \mathcal{F})$ is the infimum archive size for which a measurable estimator, allowed an independent random seed, estimates $\mathbb{E}_Q f - \mathbb{E}_P f$ uniformly over $f \in \mathcal{F}$ with error probability at most δ in that fixed experiment. We use $n^*_{\mathrm{base}}(t)$ for the corresponding Q_t shorthand and $d(t) = D(Q_t\|P)$ for the single-score-law tilt KL; the full proof package states the formal product law and boundary conventions.

Prompt mixtures and general target masses. For a target mass $\eta \in (0, 1)$, $\alpha_\eta(q, \mu) = \inf\{\mathbb{E}_\mu[\varphi] : 0 \le \varphi \le 1,\ \mathbb{E}_q[\varphi] \ge \eta\}$ and $r^\star_\eta = -\log \alpha_\eta(q, \mu)$; Neyman–Pearson implies that an optimal φ thresholds W with a randomized boundary atom. In the atomless single-prompt tilt, write $r_\eta(t)$ for the corresponding event rarity. For standard-Borel prompt and output spaces with measurable conditional kernels, define the operational prompt-conditional value by taking the infimum over jointly measurable tests satisfying $\mathbb{E}_{q_x} \varphi(x, \cdot) \ge \eta$ for almost every prompt. It is no smaller than the unrestricted joint value, so its rarity is no larger. Equality with $\mathbb{E}_X \alpha_\eta(q_X, \mu_X)$ additionally requires measurable per-prompt optimizers or measurable near-optimizers; the usual jointly measurable likelihood-ratio and quantile construction supplies them. In our finite-prompt

experiments this condition is automatic. Constructions explicitly using fractional boundary membership give exact half-mass tests; E2 instead uses inclusive score thresholds, and the trained-policy analyses use the estimated-threshold protocols detailed below. Their reported masses are operational measurements unless explicitly labeled exact. Over prompt mixtures these quantities may lie below the mean KL; the sharp relation $r_\eta(t) = d(t) + o(d(t))$ is a single-prompt statement under the one-sided exponential-power or supplementary regular-tail assumptions. The no-hit, trade-off, and duality theorems extend to fixed prompt mixtures by treating (X, O) jointly. The appendix and Theorem 2 use $m_0 \equiv 1$, $m_1 = 1 - b\mathbf{1}_A$ $(0 < b \le 1)$ with general $q(A) = \eta \in (0, 1)$; the shift-target form uses separation $b|\eta - \mu(A)|$.

Coverage modulus. For a score law Q, write $\theta_Q(f) = \mathbb{E}_Q f(S) - \mathbb{E}_P f(S)$, where P is the baseline score law. For $\delta \in (0, \frac{1}{2})$ and $c \in [0, 1]$, define

$$\Psi_\delta(c) = \begin{cases} +\infty, & c = 0, \\ \dfrac{\log(1/(2\delta))}{-\log(1-c)}, & 0 < c < 1, \\ 1, & c = 1. \end{cases}$$

Definition A.1 (Audit coverage modulus)**.** *For $\gamma \ge 0$, the strict and non-strict moduli are*

$$\mathfrak{C}^{>}_{\mathcal{F},Q}(\gamma) = \inf_{\substack{A,\ f_0,f_1\in\mathcal{F};\\ f_0=f_1 \text{ on } A^c;\\ |\theta_Q(f_1)-\theta_Q(f_0)|>\gamma}} P(A),$$

$$\mathfrak{C}^{\ge}_{\mathcal{F},Q}(\gamma) = \inf_{\substack{A,\ f_0,f_1\in\mathcal{F};\\ f_0=f_1 \text{ on } A^c;\\ |\theta_Q(f_1)-\theta_Q(f_0)|\ge\gamma}} P(A).$$

If the strict feasible set is nonempty, applying Theorem 1(a) to near-minimizing pairs gives the globally valid bound

$$n^*_{\text{base}} \ge \Psi_\delta\big(\mathfrak{C}^{>}_{\mathcal{F},Q}(2\epsilon)\big)\,.$$

When $\mathfrak{C}^{>}_{\mathcal{F},Q}(2\epsilon) \le \frac{1}{2}$, this implies

$$n^*_{\text{base}} \ge \frac{1}{2\,\mathfrak{C}^{>}_{\mathcal{F},Q}(2\epsilon)} \log\frac{1}{2\delta}.$$

For a non-strict modulus and any $\tau > 0$, put $C_\tau = \mathfrak{C}^{\ge}_{\mathcal{F},Q}(2\epsilon + \tau)$. The exact statement is $n^*_{\text{base}} \ge \Psi_\delta(C_\tau)$; the simpler $[2C_\tau]^{-1}\log(1/(2\delta))$ form is asserted only when $C_\tau \le \frac{1}{2}$. Empty feasible sets make the corresponding statement vacuous, and zero modulus is interpreted as an infinite lower bound. The proof is a standard two-point argument (Tsybakov, 2009); the content is *where* the worlds can hide: for positive tilts the likelihood ratio increases in s, so among sets with $Q_t(A) \ge \eta$ those minimizing $P(A)$ are high-score tails (randomized sets handle atoms). For negative tilts the direction is reversed. The modulus localizes the classical overlap obstruction at the optimizer-emphasized score region.

A.2 Proof Roadmap for the Passive Law

Theorem 1(a). Two-point argument: on the no-hit event $E_n = \{S_1, \dots, S_n \notin A\}$ the label laws under m_0 and m_1 coincide, so an estimator succeeding in both worlds must fail on E_n in one of them; summing the two success conditions gives $(1-p)^n = \mathbb{P}(E_n) \le 2\delta$, and $-\log(1-p) \le 2p$ for $p \le \frac{1}{2}$.

Theorem 1(b). Fix $L \in (0, \infty)$ and $\delta \in (0, \frac{1}{4})$. For the one-sided exponential-power density, let $s_t = t^{1/(\alpha-1)}$ and $w_t = ((\alpha-1)s_t^{\alpha-2})^{-1/2}$. Direct Laplace estimates show that the tilt concentrates near s_t, $d(t) = s_t^\alpha/\alpha + o(s_t^\alpha)$, and fixed tilted quantiles are $s_t + O(w_t)$. A Lipschitz ramp supported just below a fixed-mass tilted tail yields the lower bound, whose baseline mass is $e^{-d(t)+o(d(t))}$. For the upper bound, cover a fixed tilted central interval by score bins of diameter $O(\epsilon/L)$; every bin has baseline mass $e^{-d(t)+o(d(t))}$ while the bin count is $e^{o(d(t))}$. Conditional Hoeffding bounds within populated bins and a Chernoff population bound give a uniformly accurate estimator with $n = e^{d(t)+o(d(t))}$ for every fixed $\epsilon \in (0, \epsilon_0(L, \alpha))$. The regular-tail extension repeats this argument only after establishing the additional tail and curvature estimates in Supplementary Theorem 5.6.

A.3 Deferred Statements

Theorem A.1 (Single-swap likelihood-ratio block oracle)**.** *Fix an integer block length $\ell \geq 2$, let $O_1, \ldots, O_\ell$ denote its coordinate episodes, and write $W_i = W(O_i)$, $d = D(q \,\|\, \mu) < \infty$, $\bar{W}_\ell = \ell^{-1} \sum_{i=1}^{\ell} W_i$, and $\bar{Y}_\ell = \ell^{-1} \sum_{i=1}^{\ell} Y_i$. Assume a fixed conditional label kernel shared by the oracle and baseline-reference experiments. The block law $d\mathsf{P}_\ell^\star / d\mu^\ell = \bar{W}_\ell$ is the marginal law obtained by privately choosing one uniform swap index, drawing that coordinate from q, drawing the rest from μ, and then discarding the index. Writing $k_\ell = D(\mathsf{P}_\ell^\star \,\|\, \mu^\ell)$, its observable-transcript KL always satisfies*

$$\begin{aligned} k_\ell &\geq [d - \log \ell]_+, \\ k_\ell &\leq \mathbb{E}_q \log(1 + (W-1)/\ell) \\ &\leq \mathbb{E}_q \log(1 + W/\ell). \end{aligned}$$

Along any sequence with $d \to \infty$ and $\mathbb{E}_q |\log W - d| = o(d)$, uniformly over block-size sequences, $k_\ell = [d - \log \ell]_+ + o(d)$. If the swap index is instead retained, revealed, logged, or otherwise available downstream, the augmented per-block KL is exactly d, not k_ℓ. For $\ell \geq 2$, the self-normalized estimator $\widehat{q}_\ell = \sum_i W_i Y_i / \sum_i W_i$ (with an arbitrary zero-denominator fallback) is unbiased for $\mathbb{E}_q Y$, and $\widehat{\Delta}_\ell = \frac{\ell}{\ell-1}(\widehat{q}_\ell - \bar{Y}_\ell)$ is unbiased for Δ_q and bounded by 2 in absolute value. For $\epsilon > 0$ and $\delta \in (0,1)$, any integer $B \geq \lceil 8\epsilon^{-2} \log(2/\delta) \rceil$ private-index blocks suffice, giving $N = B\ell$ labels and observable KL $K_{\mathrm{obs}} = Bk_\ell$; under the same concentration assumption, when $\log \ell \leq d + o(d)$, $K_{\mathrm{obs}}/B + \log(N/B) = d + o(d)$. The construction requires likelihood-ratio access and direct q-sampling access and is an information-theoretic oracle benchmark.

Corollary A.1 (Exponential-power observable-KL active phase)**.** *Fix $L \in (0, \infty)$, $\epsilon \in (0, \frac{1}{4})$, and $\delta \in (0, \frac{1}{2})$. Fix either noiseless labels or conditionally independent Bernoulli labels and use that same experiment for every active law and iid-baseline reference. For an audit $\mathcal{A}$, let $N(\mathcal{A})$ be its query count and $K(\mathcal{A})$ its observable-transcript KL. In the exact single-prompt exponential-power model of Theorem 1(b), log-ratio concentration holds. The private-index block oracle is uniform over bounded targets with $B = O_{\epsilon,\delta}(1)$. Minimax over the L-Lipschitz class, the observable-transcript lower law gives $K(\mathcal{A})/(1-2\delta) + \log N(\mathcal{A}) \geq d(t) - o(d(t))$. Consequently,*

$$\inf_{\mathcal{A}\ \text{successful}} \{K(\mathcal{A}) + \log N(\mathcal{A})\} = \Theta_{\epsilon,\delta,\alpha,L}(d(t)). \tag{5}$$

Here success means uniform accuracy over the Lipschitz class in the fixed experiment, discarded sampler randomness is marginalized, and retained or exposed state is charged. The constants are not claimed to match, and the equality concerns the least combined resource in this oracle model rather than every feasible resource pair or any practical sampler.

Proposition A.1 (Multi-stage path-KL composition)**.** *Let $\mathbb{P}_0$ and $\mathbb{P}_R$ be baseline and optimized laws on a finite-stage standard-Borel trajectory $(O_1, \ldots, O_R)$, assume $\mathbb{P}_R \ll \mathbb{P}_0$ and $D(\mathbb{P}_R \,\|\, \mathbb{P}_0) < \infty$, and let μ, q be the induced laws of the final observable O_R. Then $D(q \,\|\, \mu) \leq D(\mathbb{P}_R \,\|\, \mathbb{P}_0)$. For $j \in \{0, R\}$ and stages $r = 1, \ldots, R$, write $O_{<r} = (O_1, \ldots, O_{r-1})$, $\mathbb{P}_j^{<r} = \mathrm{Law}_{\mathbb{P}_j}(O_{<r})$, and choose regular conditional stage laws $\mathbb{P}_{j,r}(\cdot \mid o_{<r})$ as Borel kernels. Assume $\mathbb{P}_{R,r}(\cdot \mid o_{<r}) \ll \mathbb{P}_{0,r}(\cdot \mid o_{<r})$ for $\mathbb{P}_R^{<r}$-almost every history. After redefining the optimized kernel on the exceptional null set, choose a jointly measurable conditional Radon–Nikodym version $L_r(o_r \mid o_{<r})$ satisfying $\mathbb{P}_{R,r}(B \mid o_{<r}) = \int_B L_r(o_r \mid o_{<r})\, \mathbb{P}_{0,r}(do_r \mid o_{<r})$ for every Borel set B, and define*

$$\begin{aligned} L_r &= L_r(O_r \mid O_{<r}), \\ \kappa_r &:= \mathbb{E}_{O_{<r} \sim \mathbb{P}_R^{<r}} \Big[D\big(\mathbb{P}_{R,r}(\cdot \mid O_{<r}) \,\big\|\, \\ &\qquad\qquad \mathbb{P}_{0,r}(\cdot \mid O_{<r})\big)\Big] \\ &= \mathbb{E}_{\mathbb{P}_R} \log L_r \in [0, \infty]. \end{aligned}$$

The conditional density and entropy chain rules give $d\mathbb{P}_R / d\mathbb{P}_0 = \prod_r L_r$ almost surely and $D(\mathbb{P}_R \,\|\, \mathbb{P}_0) = \sum_r \kappa_r$; because the path has finitely many stages and its KL is finite, every κ_r is finite. Only $D(q \,\|\, \mu) \leq \sum_r \kappa_r$ holds automatically. If $D(q \,\|\, \mu) \to \infty$ and a final tail A satisfies $q(A) \geq \eta > 0$ and $\log(1/\mu(A)) = D(q \,\|\, \mu) + o(D(q \,\|\, \mu))$, consider $m_0 \equiv 1$ and $m_1 = 1 - b\mathbf{1}_A$. For the lower bound, labels are noiseless; the same argument applies to any fixed conditional label model whose full conditional law is identical in the two worlds on A^c. For a common nonanticipating auditor $\mathcal{A}$ with a positive integer query count n, let $\mathbb{P}_j^{\mathcal{A}}$ be its observable query–label transcript

law in world m_j, and let $\mathbb{M}_j^{(n)}$ be the corresponding reference law from n iid μ-queries with the same label experiment. Discarded sampler randomness is marginalized; any retained, revealed, logged, or downstream-used state is included in both laws. Select either $T_j = R_q(m_j)$ with $\epsilon < b\eta/2$, or $T_j = \Delta_q(m_j)$ with $\epsilon < b\eta/4$, and assume that the auditor outputs $\widehat{T}$ satisfying

$$\mathbb{P}_j^{\mathcal{A}}(|\widehat{T} - T_j| \le \epsilon) \ge 1 - \delta, \qquad j = 0, 1,$$

for a fixed $\delta \in (0, \frac{1}{2})$. Also assume

$$D(\mathbb{P}_j^{\mathcal{A}} \,\|\, \mathbb{M}_j^{(n)}) \le K_{\mathrm{audit}}, \qquad j = 0, 1.$$

Then $K_{\mathrm{audit}}/(1 - 2\delta) + \log n \ge D(q \,\|\, \mu) - o(D(q \,\|\, \mu))$. If moreover labels are bounded and generated by the stable shared kernel of Theorem A.1, and the final log-ratio concentrates, $\mathbb{E}_q|\log \frac{dq}{d\mu} - D(q \,\|\, \mu)| = o(D(q \,\|\, \mu))$, that theorem gives the corresponding oracle upper benchmark. The frontier composes through the final density ratio and becomes additive in cumulative stage KL only under a non-cancellation condition such as $D(q \,\|\, \mu) = \sum_r \kappa_r - o(\sum_r \kappa_r)$.

Theorem A.2 (Smooth strict separation)**.** *Fix $L \in (0, \infty)$ and let $\mathcal{F}_L = \{f : \mathbb{R} \to [0, 1] : f \text{ is } L\text{-Lipschitz}\}$. Labels are either noiseless with value $f(X)$ or conditionally Bernoulli with mean $f(X)$, with the same fixed experiment used throughout. For $P = N(0, 1)$, $Q_t = N(t, 1)$, and $d = t^2/2$, there exists $\epsilon_0 = \epsilon_0(L) > 0$ such that, for every fixed $\epsilon \in (0, \epsilon_0)$ and $\delta \in (0, \frac{1}{4})$, direct two-policy auditing needs $O(\epsilon^{-2} \log \frac{1}{\delta})$ labels from each law. Conversely, for all sufficiently large t, any common nonanticipating strategy with a positive integer query count n that estimates the shift $\theta_t(f) = \mathbb{E}_{Q_t} f - \mathbb{E}_P f$ uniformly over $\mathcal{F}_L$ and whose observable query–label transcript law in every world has KL at most K from the corresponding iid-P transcript in that experiment must obey $\frac{K}{1-2\delta} + \log n \ge d + \frac{1}{2} \log d - O_{\delta,L}(1)$. Discarded sampler randomness is marginalized and retained state is charged. In particular $K = o(d)$ forces $n \ge e^{d - o(d)}$.*

Theorem A.3 (Exact risk under a fixed audit sampler)**.** *With $p = \mu(A)$ and a positive integer number n of noiseless labels drawn from audit distribution ν, the exact minimax expected absolute-error risk for $R_q(m) = \mathbb{E}_q[m]$ over $\{m_0, m_1\}$ is $\frac{b\eta}{2}(1 - \nu(A))^n$; for the shift target $\Delta_q(m)$ it is $\frac{b|\eta - p|}{2}(1 - \nu(A))^n$. Under a density-ratio cap $d\nu/d\mu \le R$, one has $\nu(A) \le \min\{1, Rp\}$. Thus, in the nontrivial case $Rp < 1$, achieving risk at most $\epsilon < b\eta/2$ for R_q requires*

$$n \ge \frac{\log(b\eta/(2\epsilon))}{-\log(1 - R\mu(A))};$$

in particular, if $R\mu(A) \le \frac{1}{2}$, then $n \ge \frac{1}{2R\mu(A)} \log \frac{b\eta}{2\epsilon}$. A per-draw KL cap $D(\nu \,\|\, \mu) \le \kappa$ forces, via the data-processing inequality $\mathrm{kl}(\nu(A)\|p) \le \kappa$, that one-draw hit probability $a = \nu(A)$ requires $\kappa \ge a \log(1/p) - h_2(a)$; in particular $\kappa = \Omega(\log(1/p))$ for any constant $a > 0$.

Theorem A.4 (Exact total-KL risk under a common-query oracle)**.** *In the two-world setup $m_0 \equiv 1$, $m_1 = 1 - b\mathbf{1}_A$, fix $p = \mu(A) \in (0, 1)$, $\eta = q(A)$, $0 < b \le 1$, a budget $K \in [0, \infty)$, and a positive integer query count n. Let $p_n = 1 - (1 - p)^n$, and optimize over query laws P^X on $\mathcal{O}^n$ that are common to both target worlds and satisfy $D(\mathsf{P}^X \,\|\, \mu^n) \le K$. Define $a_K = \sup\{a \in [0, 1] : \mathrm{kl}(a\|p_n) \le K\}$. With noiseless labels, the exact minimax expected absolute-error risk for R_q is $\frac{b\eta}{2}(1 - a_K)$, and for Δ_q it is $\frac{b|\eta - p|}{2}(1 - a_K)$. The optimizer tilts the joint query law only according to the hit event H: $d\mathsf{P}^{X,\star}/d\mu^n = (a_K/p_n)\mathbf{1}_H + ((1 - a_K)/(1 - p_n))\mathbf{1}_{H^c}$. This is a common, nonanticipatingly disintegrable information-theoretic oracle law, not a claim of practical sampler access and not the high-probability statement of Theorem 2.*

Proposition A.2 (Bit count is not the operative complexity measure)**.** *For any density shift g, $\inf_{T:\,|\mathrm{range}(T)| \le M} \mathrm{dist}_{L^1(\mu)}(g, L^1(\sigma(T)))$ equals the optimal M-level scalar L^1-quantization distortion of $G = g(O)$. In the Gaussian tilt model with $d = t^2/2$, write g_d for the shift at KL budget d, define $D_M(d) = \inf_{T:\,|\mathrm{range}(T)| \le M} \mathrm{dist}_{L^1(\mu)}(g_d, L^1(\sigma(T)))$, and let $M_\tau(d) = \min\{M : D_M(d) \le \tau\}$. For any fixed $\tau \in (0, \frac{1}{2})$, $M_\tau(d) = \Theta_\tau(\sqrt{d})$, i.e. the required representation uses $\frac{1}{2} \log_2 d + O_\tau(1)$ bits.*

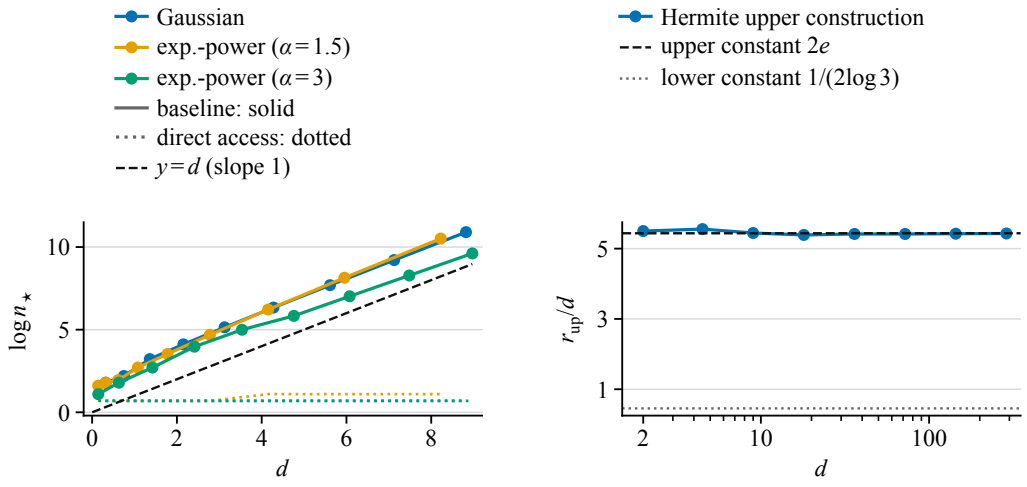


Figure 5: E0 (synthetic): (left) the no-hit construction underlying the baseline lower bound has $\log n = d + o(d)$, while the direct-q/two-policy control stays flat, for Gaussian and stretched-exponential score laws; (right) the smallest degree r_{up} whose Hermite truncation has constructed L^2 residual at most $\tau = \frac{1}{4}$, divided by d, approaches $2e$. The right panel is an upper construction for $k^*(d, \tau)$, not a numerical computation of the minimum L^1 degree; neither panel numerically computes the full minimax optimum.

Proposition A.3 (Structured tail-leakage principle)**.** *Let $q_d \ll \mu$ be a sequence with $g_d = dq_d/d\mu - 1$, and let A_d satisfy $q_d(A_d) \geq \eta > 0$ and $\mu(A_d) = e^{-d+o(d)}$ as $d \to \infty$. For a feature span V, define $\beta_V(A) = \sup_{v \in V, v \neq 0} \mathbb{E}_\mu[|v|\mathbf{1}_A]/\mathbb{E}_\mu|v|$. Suppose $\{V_R\}_{R \geq 0}$ is a nested family of closed linear monitor subspaces containing the constants and $\psi : [0, \infty) \to \mathbb{R}$ satisfies $\beta_{V_R}(A_d) \leq \exp\{\psi(R) - d/2 + o(d)\}$ (this envelope is the defining* structured *property). The remainder must be $o(d)$ along the particular sequence $R = R(d)$ below; no uniformity over all R is otherwise asserted. Fix $\xi > 0$. For any such sequence satisfying $\psi(R(d)) \leq (\frac{1}{2} - \xi)d$, $\mathrm{dist}_{L^1(\mu)}(g_d, V_{R(d)}) \geq \eta - o(1)$ and, for every $\rho \in [0, \frac{1}{2}]$, Theorem 3 yields a bounded invisible residual whose optimized target shift is at most $-\rho(\eta - o(1))$.*

A.4 Polynomial-monitor degree bounds

Corollary A.2 (Polynomial-monitor degree threshold)**.** *In the Gaussian tilt model with $d = t^2/2$, let g_t be the density shift at tilt strength t and let $V_k = \mathrm{span}\{1, S, \ldots, S^k\}$. For an approximation tolerance τ, define $k^*(d, \tau)$ as the smallest polynomial degree for which $\mathrm{dist}_{L^1(\mu)}(g_t, V_k) \leq \tau$. For every fixed $\tau \in (0, \frac{1}{4})$, $k^*(d, \tau) = \Theta(d)$ as $d \to \infty$. More specifically, $\liminf k^*/d \geq 1/(2 \log 3)$ and $\limsup k^*/d \leq 2e$ (Appendix A). These nonmatching constants are not claimed to identify the exact threshold. Consequently, any degree sequence growing strictly slower than d eventually admits a bounded monitor-invisible perturbation that lowers the optimized target by at least $\rho/4$.*

A.5 Synthetic Illustrations

Figures 5 and 6 collect synthetic illustrations of the access and representation barriers. They visualize the proved rates and stated oracle upper bounds; they do not add numerical claims about the full minimax optima.

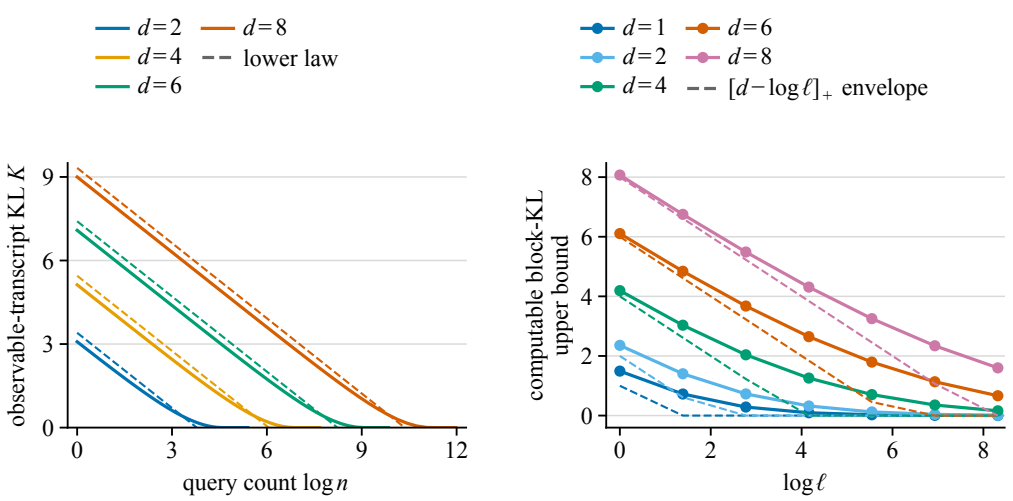


Figure 6: E0, remaining panels: (left) the event-level KL required to force hit probability 0.9 against the budget–sample lower law in the Gaussian model; these are information-theoretic event-tilted points, not practical samplers. Right: the computable block-KL upper bound $\mathbb{E}_q \log(1 + W/\ell)$ for the private-index single-swap oracle versus the $[d - \log \ell]_+$ leading-order envelope, where ℓ is the block size and $[u]_+ = \max\{u, 0\}$. At $\ell = 1$ the sampling law is direct q, whose exact KL is d; the displayed ordinate nevertheless remains the conservative bound $\mathbb{E}_q \log(1+W/\ell)$. The private-index shift-estimator theorem is stated for $\ell \geq 2$.

A.6 Proof Roadmap for the Active Law

Theorem 2 and Theorem A.4. With noiseless labels and a common nonanticipating strategy, before the first hit of A the two worlds induce identical histories, so their hit probabilities coincide; write the common probability as a and let $p_n = 1-(1-p)^n$ be the baseline-reference hit probability. Uniform high-probability success forces $a \geq 1 - 2\delta$, while data processing through the hit indicator gives $K \geq \mathrm{kl}(a\|p_n)$. Using $p_n \leq np$ yields Theorem 2. Separately, for expected absolute risk in the common-query joint-law oracle model, data processing restricts $a \leq a_K$, and the event-tilted law in Theorem A.4 attains hit probability a_K without depending on the unknown world. On a hit, the noiseless label identifies the world; on no hit, the midpoint estimator is minimax. This proves the exact risk formula without conflating it with the high-probability theorem.
The reduction to a binary hit indicator and the resulting change-of-measure step are standard; related adaptive change-of-measure inequalities appear in Kaufmann et al. (2016, Lemma 1). Here KL compares active and iid-baseline transcripts within the same label experiment, rather than transcripts across bandit environments; the audit-specific content is the observable-transcript accounting, including which retained sampler state is charged, and the resulting movement–sample frontier for the hard region.

Theorem A.2. For $Z \sim N(0,1)$, define $c_L = \mathbb{E}\min\{1, LZ_+\}$. The ramp pair $f_0(s) = \frac{1}{2}$, $f_1(s) = \frac{1}{2} - \frac{1}{2}\min\{1, L(s-t)_+\}$ lies in $\mathcal{F}_L$, is separated by $\frac{1}{2}c_L - o(1)$, and coincides below t; the transcript argument of Theorem 2 then applies with $p_t = \bar{\Phi}(t)$, where $\bar{\Phi}$ is the standard normal survival function, and Mills' ratio gives $\log(1/p_t) = d + \frac{1}{2}\log d + O(1)$.

Theorem A.1. $\mathbb{E}_{\mu^\ell}\bar{W}_\ell = 1$ makes $\mathsf{P}^\star_\ell$ a law; it is the marginal mixture "one coordinate from q, the rest from μ" after the uniform swap index I is privately discarded. Conditioning on the swapped coordinate and applying Jensen gives the KL upper bound, while $\bar{W}_\ell \geq W(O_I)/\ell$ gives $k_\ell \geq [d - \log \ell]_+$. Log-ratio concentration yields $k_\ell = [d - \log \ell]_+ + o(d)$. The chain rule shows that exposing the index instead costs exactly d per block. Unbiasedness is an exact cancellation after multiplication by the block density; combining the self-normalized q-mean with the same block's ordinary mean yields $\widehat{\Delta}_\ell$. Corollary A.1 is exponential-power, fixes one noiseless or Bernoulli experiment throughout, and is an oracle phase statement rather than a practical-sampler result.

Proposition A.1. The final observable is a measurable map of the trajectory, so data processing gives $D(q \,\|\, \mu) \le D(\mathbb{P}_R \,\|\, \mathbb{P}_0)$. Choosing jointly measurable conditional Radon–Nikodym versions under the regular conditional laws makes each L_r a transcript function and κ_r a nonnegative conditional KL increment. Under the stated finite-stage and finite-path-KL assumptions, the chain rule gives $D(\mathbb{P}_R \,\|\, \mathbb{P}_0) = \sum_r \kappa_r < \infty$, so every increment is finite. The audit lower bound applies to the final-object law with the target-specific error ranges in the proposition; the upper benchmark additionally requires concentration of the *final* log likelihood ratio. Equality with cumulative stage KL requires the stated non-cancellation condition.

A.7 Proof Roadmap for the Representation Law

Theorem 3 and Corollary 1. Since $\mathbb{E}_\mu g = 0$, $\Delta_q(m) = \rho \mathbb{E}_\mu[gh]$ on $\mathcal{M}_{\rho,V}$. Hahn–Banach duality for the quotient $L^1(\mu)/V$ identifies the supremum over $\{\|h\|_\infty \le 1, h \perp V\}$ with $\mathrm{dist}_{L^1(\mu)}(g, V)$; symmetry gives both signs, and weak-$*$ compactness gives an attaining residual. For a statistic T, put $\mathcal{G} = \sigma(T)$. Then $\mathbb{E}_\mu[h \mid \mathcal{G}] = 0$ is exactly the annihilator condition for the closed subspace $L^1(\mathcal{G})$. The pairwise formulation additionally requires both targets to lie in the symmetric band $[\frac{1}{2} - \rho, \frac{1}{2} + \rho]$; an extremizing pair is $m_\pm = \frac{1}{2} \pm \rho h$. For $0 < \rho \le \frac{1}{2}$, the distance vanishes if and only if g is $\mathcal{G}$-measurable; the endpoint $\rho = 0$ is trivial and is not used for this equivalence.

Proposition A.2. Any approximation $v(T)$ based on a statistic T with M values is an M-level quantizer of G, and conversely an optimal M-level quantizer can be taken as a statistic of G. Under Q_t, $\log W_t \sim N(d, 2d)$, so for any fixed relative-error radius $a \in (0, 1)$, M codepoints cover at most $O(aM/\sqrt{d})$ of the mass. Hence $D_M(d) \ge c \min\{1, \sqrt{d}/M\}$ for a universal $c > 0$, while midpoint coding of a fixed central interval gives $D_M(d) \le \tau$ at $M = O_\tau(\sqrt{d})$.

Proposition A.3 and inequality (4)**.** For any $v \in V$, splitting $\mathbb{E}_\mu|g - v|$ on A and A^c gives $\mathbb{E}_\mu|g - v| \ge (q(A) - \mu(A)) - \beta_V(A)(\mathbb{E}_\mu|g - v| + \mathbb{E}_\mu|g|)$ with $\mathbb{E}_\mu|g| = 2\,\mathrm{TV}(q, \mu) \le 2$, where TV denotes total variation distance; rearranging proves (4). Under the envelope hypothesis, $\psi(R) \le (\frac{1}{2} - \xi)d$ makes $\beta_{V_R}(A_d) \to 0$ while $q_d(A_d) \ge \eta$ and $\mu(A_d) \to 0$, so (4) tends to η.

Corollary A.2. Lower: hypercontractivity $\|p\|_4 \le 3^{k/2}\|p\|_2$ and interpolation give $\|p\|_2 \le 3^k \|p\|_1$, hence $\beta_{V_k}(\{S \ge t\}) \le e^{k \log 3 - d/2 + o(d)}$. Thus every asymptotic coefficient below $1/(2 \log 3)$ leaves L^1 error tending to at least $1/2$. Upper: truncating the Hermite expansion of g_t at degree $\lceil 2ed \rceil$ leaves L^2, hence L^1, residual tending to zero. Therefore $k^*(d, \tau) = \Theta(d)$ with $\liminf k^*/d \ge 1/(2 \log 3)$ and $\limsup k^*/d \le 2e$. The $2e$ value and E0 panel are a Hermite-truncation upper construction, not the exact minimax threshold. This result is adjacent to the low-degree likelihood-ratio literature for hypothesis testing (Kunisky et al., 2022; Kunisky, 2021), which studies L^2 projections of likelihood ratios and their relation to computational testing thresholds. Our criterion instead concerns L^1 approximation of the centered density shift and the induced worst-case bounded hidden target change; it makes no computational-hardness claim.

B Robustness Runs

Optimization-KL axis, tolerance, and seed sensitivity (E1/R1). The main text plots the archive requirement against measured half-mass rarity (Figure 2, left). Figure 7 shows the same data on the optimization-KL axis; the growth pattern is identical, with the finite-pool ceiling $d_{\max} = 6.93$ marked. The direct two-policy comparison uses n labels per arm, so its total label count is $2n = 2{,}048$–$2{,}896$, still more than two orders of magnitude below the 370,728-episode archive requirement at $d = 6$. Re-running E1 with $\epsilon = 0.02$ and 0.03 preserves the shape: at $d = 4$, $n_{\mathrm{req,base}}$ is 23,170, 19,484, and 16,384 for $\epsilon = 0.02, 0.025, 0.03$. The fitted $\log n_{\mathrm{req,base}}$-versus-$d$ slopes across nontrivial points are 1.151, 1.151, and 1.137 for seeds 2026–2028 (seed SD 0.008).

Weaker judge (E1). Replacing the 7B judge with Qwen2.5-1.5B widens the exploitable gap (top-decile error rate 0.134; AUC 0.764) and makes overoptimization strictly worse: $\mathbb{E}_{q_t}[m]$ declines monotonically from 0.609 to 0.564 and the target shift $\Delta_t(m) := \Delta_{q_t}(m)$ turns *negative* beyond

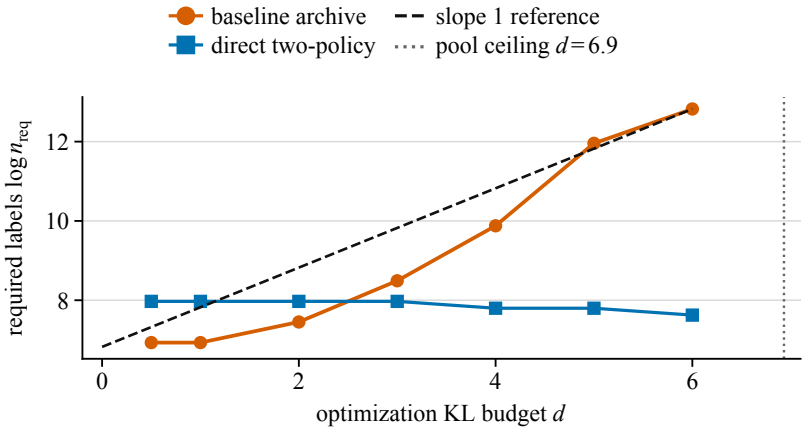


Figure 7: E1 on the optimization-KL axis: the tested baseline-archive requirement $\log n_{\text{req}}$ versus $d(t)$, with the direct two-policy total and the finite-pool ceiling $d_{\max}$ (dotted).

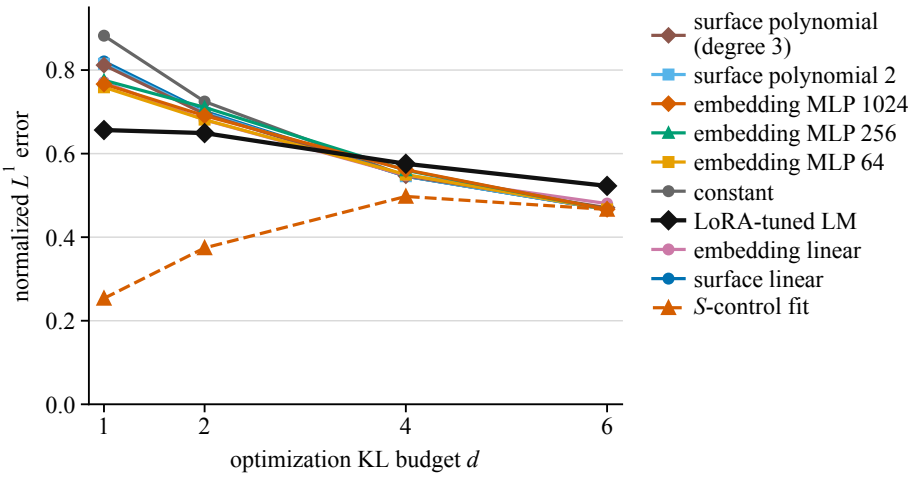


Figure 8: Full monitor ladder: normalized held-out fitted reconstruction error for the constant baseline, all surface/embedding rungs, the LoRA-LM rung, and the learned S-control. These fitted errors are not certified distances for the nonlinear model classes.

$d \approx 3.5$—optimizing a weak judge harms programmatic answer accuracy. Because Δ_t crosses zero, the audit target itself degenerates ($|\Delta_t| < \epsilon$ over most of the grid), which our harness reports via a per-point triviality flag; the coverage scaling claim in the main text is therefore made on the 7B-judge pool where the target is uniformly nontrivial.

Second base model (E1). Repeating the pipeline with Qwen2.5-14B-Instruct (self-judged, N=256 per prompt) yields accuracy 0.624 and a smaller proxy–target gap: the region at or above the empirical 90th score percentile includes ties, contains $9{,}683/51{,}968 = 18.6\%$ of traces, and has error rate 0.025. This run tilts the deterministically jittered score described below. The original clipped score has an empirical finite-pool KL ceiling of 3.58, so the stored $d = 4, 5$ points, the six-point slope 1.156, and $n_{\text{req,base}} = 77{,}936$ at $d = 4$ are jittered-score results, not evidence about further optimization of the original clipped score. Below that ceiling, the measured shift remains near $+0.06$ and nontrivial, while the tested archive requirement rises from 1,218 at $d = 1$ to 8,192 at $d = 3$. We therefore retain this run only as a low-to-moderate-KL robustness check.

Monitor positive control (E3). A learned control given the judge score S reaches held-out relative distortion 0.254 and 0.374 at $d = 1, 2$, compared with 0.656 and 0.649 for the fine-tuned score-blind LM. This supports an informational explanation for the score-blind ladder's low-d gap within the tested model classes. At $d \geq 4$ the fitted control's advantage narrows sharply as the tilt concentrates on a handful of traces per prompt (median ESS 8.3 at $d = 4$ and smaller thereafter) and the empirical L^1 target becomes increasingly spike-dominated. We therefore read the high-d control values as a finite-sample limitation of this fitted diagnostic, not as evidence that S is insufficient or as a contradiction of main-text Corollary 1.

Empirical linear-monitor projection certificates (R5). Let $\widehat{\mu}$ be the uniform law on the same 200×256-trace monitor subset, and let g_{full} be the full-pool density shift evaluated there. We solve the empirical $L^1(\widehat{\mu})$ projection and its LP dual for the linear and degree-2 surface spans; degree 3 uses a feasible dual lower bound. Across all 12 dual witnesses, the maximum feature-orthogonality residual is 7.1×10^{-8}; for the exact primal–dual cells, the relative duality gap is at most 1.2×10^{-9}. The restricted likelihood ratios are not renormalized under $\widehat{\mu}$, so $\mathbb{E}_{\widehat{\mu}}(1 + g_{\mathrm{full}})$ need not equal one. The reported values are therefore exact approximation certificates for the stated empirical function where indicated, not exact target-shift certificates for a normalized subset law. At $d = 1, 2$ the projection ratios differ from the fitted ladder by at most 0.011 (at most 0.005 for the linear and degree-2 spans); at $d = 4, 6$ they are 0.030–0.054 larger.

Table 1: R5 empirical projection: LP certificate / *held-out fitted-model error* under the monitor-subset convention defined in Appendix B. The second number is an estimate on a different empirical split, not a certified upper bound on the LP measure. Linear/quadratic entries are exact LP optima; cubic is a dual-only *lower bound*. The d=6 cubic dual completed in 2.82 h under a 4 h time limit.

rung	d=1	d=2	d=4	d=6
surf_lin	0.818 / 0.821	0.706 / 0.701	0.578 / 0.546	0.520 / 0.466
surf_poly2	0.811 / 0.815	0.701 / 0.698	0.578 / 0.546	0.520 / 0.467
surf_poly3	0.802 / 0.812	0.696 / 0.695	0.577 / 0.547	0.520 / 0.471

Table 2: Low-degree extensions of the R5 empirical projection, using the same $\widehat{\mu}$ and g_{full} convention defined in Appendix B. Linear and quadratic entries are exact primal–dual LP optima; cubic entries are feasible-dual lower bounds. The certified projection changes little through degree three, but this is not an impossibility claim for all nonlinear monitors.

d	Linear exact	Quadratic exact	Cubic dual LB
1	0.818	0.811	0.802
2	0.706	0.701	0.696
4	0.578	0.578	0.577
6	0.520	0.520	0.520

Prompt–trace monitors (R7). A deployment monitor may observe the prompt as well as the generated trace. We therefore repeat the frozen-embedding ladder on concatenated prompt–trace inputs with the identical split and loss. Descriptive minima across the displayed prompt–trace fits are 0.739, 0.669, 0.545, and 0.466 at $d = 1, 2, 4, 6$, compared with 0.759, 0.681, 0.546, and 0.466 for the displayed trace-only surface/embedding fits. The prespecified paired embedding-MLP comparison detects an error reduction at every tested d, but the gains are small and shrink to 0.0001 at $d = 6$, where they are practically negligible. Thus prompt access does not close the low-d gap to the learned S-control in the tested embedding monitors; the minima are descriptive held-out summaries, and this fitted-model comparison is not a class-level impossibility result.

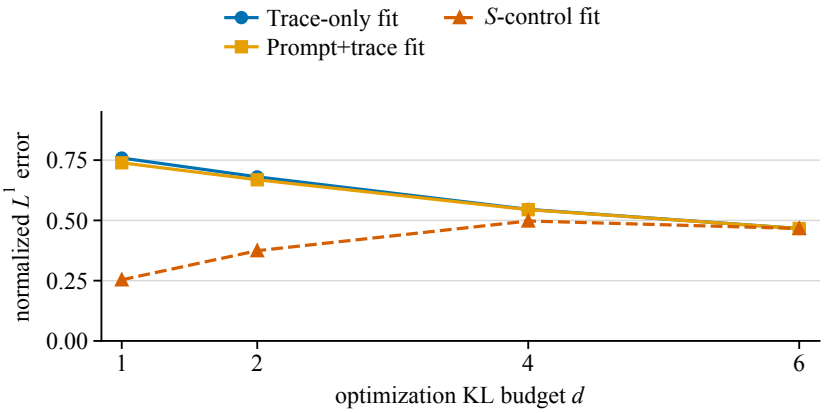


Figure 9: Prompt–trace monitor robustness. Descriptive minima across the displayed prompt–trace surface/embedding fits improve only slightly over the corresponding trace-only fits. The learned S-control remains substantially better at low d. These are held-out fitted-model errors, not certified class-level optima.

Alternative model-derived target: stepwise logical validity (E7). To probe dependence on the target definition, a Qwen2.5-14B step checker (a different model from the 7B judge) labels 128 traces per prompt (25,984 labels), setting m'=1 when the checker answers Yes to whether every step follows from the previous ones—the prompt explicitly asks it to ignore final-answer correctness. The code's re-encoding guard reports $\mathrm{corr}(m', S) = 0.65$ and $\mathrm{corr}(m', m) = 0.59$ and marks the diagnostic inconclusive only when $|\mathrm{corr}(m', S)| > 0.9$; the observed value does not trigger that guard, but correlation alone does not establish independence. The label behaves very differently under the same judge tilt: the programmatic correctness shift peaks at $+0.0553$ at $d = 1$ and erodes to $+0.0301$ at $d = 4$, where it is statistically indistinguishable from zero (prompt-bootstrap 95% CI $[-0.0106, +0.0703]$, two-sided $p = 0.1456$). Correctness shifts are distinguishable from zero only through $d = 3$, while the model-derived stepwise-validity shift remains positive at every tested d: it is $+0.1083$ at $d = 0.5$, peaks at $+0.1657$ at $d = 3$, and is $+0.1654$ at $d = 4$. Our reading is cautionary rather than reassuring: the model-derived target can co-inflate with the optimized judge even as programmatic correctness gains erode. It therefore cannot be trusted as ground truth for auditing that pressure—the reason the main text insists on a programmatic m.

The target effect is label-dependent exactly as it is score-dependent (Figure 11). For fixed baseline and optimized laws, LR-tail overlap geometry is independent of the target label; the estimation lower bounds remain worst-case statements for target classes and observation experiments satisfying their separation assumptions. E7 does not establish equal audit sample complexity for the two labels.

C Testing Whether Tilt Calibration Transfers to GRPO (E4/E5)

The main-text exact tilt provides a calibration regime in which the proxy score orders likelihood ratios and policy KL predicts fixed-mass LR-tail rarity at leading order. The trained-policy question is whether this shortcut survives reinforcement learning. We test it directly by fine-tuning the base policy with GRPO (Shao et al., 2024) against the same judge score S (LoRA adapters (Hu et al., 2022), KL coefficient β; configuration in the Experimental Details section), and evaluating intermediate checkpoints along each training trajectory. For a trained policy π_θ, let (X_i, Z_i) be the problem–trace pair in evaluated sample i, for $i = 1, \ldots, n$. We estimate its optimization budget from sequence log-ratios as $\hat{d} = \frac{1}{n}\sum_{i=1}^{n}(\log \pi_\theta(Z_i \mid X_i) - \log \mu_{X_i}(Z_i))$ (generation-time $\log \pi_\theta$ vs. base-policy teacher-forced $\log \mu$; a Monte-Carlo estimate, not an exact evaluation), and each policy pool is judged and graded by the same pipeline as the tilt pool. We sweep β and learning rate to place checkpoints across $\hat{d} \in [0.16, 256]$; the base-policy calibration row itself has an empirical 0.16-nat generation-time/teacher-forced offset. Besides the original 1024-token tilt, we recompute the exact tilt on the GRPO study's own 768-token baseline pool, yielding matched-protocol score, correctness, and rarity curves.

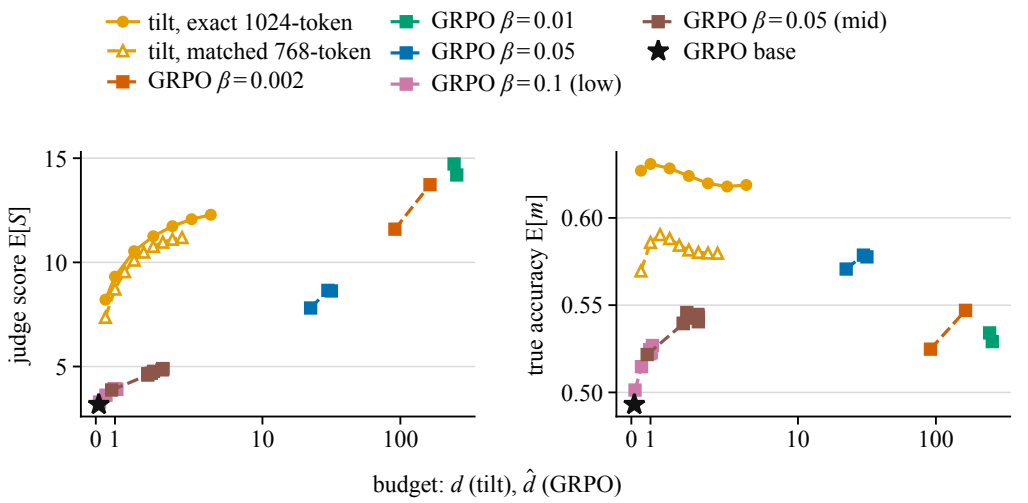


Figure 10: E4: checkpoints from five GRPO training runs (measured $\hat{d}$) against exact tilts (target d). Filled orange circles use the original 1024-token pool; hollow triangles use the same 768-token baseline protocol as GRPO. Left: on the matched protocol, the tilt reaches $\mathbb{E}[S] \approx 8.7$ at $d = 1$, whereas the closest GRPO checkpoint needs $\hat{d} \approx 30$ to reach the same score level. Right: checkpoints pooled across runs show a rise-then-decline accuracy pattern over increasing measured KL. Lines connect checkpoints only within the same run; the pooled Goodhart-shaped pattern is not a single training trajectory.

Figure 10 shows both halves. First, GRPO buys judge-score gains at far higher measured policy-KL budget than the exact optimizer (left panel): on the identical 768-token baseline pool, the tilt reaches $\mathbb{E}[S] \approx 8.7$ at $d = 1$, while the closest GRPO checkpoint reaches $\mathbb{E}[S] = 8.66$ only at $\hat{d} = 29.82$. This matched-protocol comparison documents a large empirical optimization-efficiency gap. Second, the checkpoints pooled across five runs show an empirical rise-then-decline accuracy pattern over measured policy KL (right panel): the pooled maximum is 0.579 at $\hat{d} = 29.82$, a different higher-KL run gives 0.525 at $\hat{d} = 91.30$, and another gives 0.534 at $\hat{d} = 245.12$ while the judge score reaches 14.72. This descriptive cross-run pattern resembles the finite-pool Goodhart-shaped curve in Figure 2 (right), but it is neither a single training trajectory nor a consequence of the analytic-tilt theorem. A reproducibility scan of the 9.6×10^5 saved policy-evaluation traces found no completion above the new character cap and no reconstructed judge prompt above the old token limit, ruling out the truncation discrepancy on those saved evaluation traces. Training-time completions were not saved, so reward-log mean lengths provide only an indirect check for the training samples.

Score-tail rarity and quantile cross-fitting (E5/R2/R3). For each checkpoint and prompt, E5 fits $\phi_p(s) = \mathbf{1}\{s > \theta_p\} + \gamma_p \mathbf{1}\{s = \theta_p\}$ so that its empirical policy mass is exactly $1/2$, including when clipped scores tie at the boundary. The same (θ_p, γ_p) test is then evaluated on an independent 768-token base-policy pool. Prompt-cluster bootstrap intervals are computed in baseline-mass space with zero-mass resamples retained before transforming to rarity; consequently, the two sparse $\beta = .01$ rows have infinite 95% upper endpoints rather than finite positive-only intervals. Splitting policy traces into even/odd folds to fit the test on one half and evaluate it on the other changes rarity by at most 0.022 nats overall (at most 0.011 nats for checkpoints with no prompt-level zero baseline mass). On the same 768-token base pool, the exact fractional-boundary tilt tail has rarity 4.57 at $d = 3.5$, close to the original 1024-token curve, whereas the uncensored GRPO score tail is 1.17 at $\hat{d} = 3.48$. The score-tail gap is therefore not a token-cap artifact.

Likelihood-ratio tails (R6). Within each prompt, we split policy traces into even and odd folds, fit one half-mass threshold per fold from stored sequence log-ratios, evaluate both thresholds on the same independent baseline pool, and average the two baseline masses. This is two-fold threshold averaging, not cross-fitting that applies each threshold to the opposite policy fold; no held-out policy-

fold mass was recorded. The policy thresholds use generation-time $\log \pi_\theta$ and teacher-forced $\log \mu$ on the original generated token IDs. For baseline-mass estimation, saved text is retokenized and both $\log \pi_\theta$ and $\log \mu$ are teacher-forced on that re-encoded sequence. The two sides therefore do not constitute a fully common-token, common-likelihood estimand. A 4,096-trace diagnostic found correlation 0.999976 between generation-time and teacher-forced policy log probabilities, while the absolute re-encoding change in $\log W$ had median 0.487 nat, 90th percentile 1.285 nats, and 99th percentile 2.244 nats; only 15.3% of token counts matched. High correlation does not itself establish tail-membership invariance or calibrate rarity error. Subject to these limitations, the procedure reports prompt-conditional LR-tail rarity for 21 of 22 checkpoints; one saved adapter was unavailable. At $\hat{d} = 3.48$, the estimated LR-tail rarity is 4.92, versus score-tail rarity 1.17 and matched-768-token tilt rarity 4.57. The $\hat{d} = 3.50$ final adapter similarly gives 4.93 versus 1.17. For every measured aggressive checkpoint ($\hat{d} \geq 22$), no base trace enters the estimated policy half-mass LR tail, so the result is right-censored rather than evidence of low rarity. These estimates suggest that the score-tail view can overstate overlap, but they do not establish the exact common-sequence LR-tail rarity.

Matched-tilt residual decomposition (R4). Because π_θ, q_t, and μ share one prompt law, for every t the exponential-family identity is

$$D(\pi_\theta \,\|\, \mu) = D(\pi_\theta \,\|\, q_t) + D(q_t \,\|\, \mu) + t\big(\mathbb{E}_{\pi_\theta}[S] - \mathbb{E}_{q_t}[S]\big).$$

The decomposition is additive only when the population mean scores match exactly. We choose t^* by finite-sample mean matching and report plug-in estimates of all three terms. The residual $D(\pi_\theta \,\|\, q_{t^*})$ measures departure from the matched one-parameter global exponential-tilt path. Because it also absorbs nonlinear and prompt-specific score dependence, we interpret it only as departure from that path. For the 18 of 22 trained checkpoints whose mean scores lie inside the finite-pool tilt range, its estimated share is 96.7–99.8%; the other four have no finite-pool score match, and the smallest-$\hat{d}$ shares are additionally sensitive to the empirical approximately 0.16-nat base-anchor offset. Table 3 reports this matched-tilt residual together with score-tail and estimated LR-tail rarity. Accordingly, the exact-tilt map from total KL to overlap should not be transferred to a trained policy without checking its likelihood-ratio geometry under a consistent evaluation convention.

Prompt-cluster statistical checks. Table 4 reports paired tests for the reported changes that admit a prompt-level comparison. For each test, we save every prompt effect and use $20{,}000$ random sign flips together with a $20{,}000$-resample prompt bootstrap. The GRPO peak exceeds the base by 0.0856 accuracy and the highest-KL final adapter by 0.0446 (two-sided $p < 10^{-4}$), but these are descriptive post-selection checks because the peak checkpoint was chosen using the same accuracy data. E7 answer-correctness shifts are distinguishable from zero through $d = 3$ (two-sided $p \leq 0.0421$) but not at $d = 4$ ($p = 0.1456$); step-validity shifts are positive at every tested d ($p < 10^{-4}$). For the paired R7 embedding MLP, adding the prompt reduces prompt-mean absolute error at every tested d (two-sided $p \leq 0.0030$), although the $d = 6$ effect is only 0.0001 and is practically negligible. These p-values are unadjusted for multiple comparisons, and the paper's geometric and LP-certified claims do not depend on them. The complete audit artifact includes the summary, all prompt-level effects, the readable R7 prompt-error file, and run metadata.

D When Optimization KL Predicts Score-Tail Rarity

The finite-pool comparisons in Appendices B–D do not extend the asymptotic theorem. Its formal scope is the one-sided exponential-power model in the main text and the regular-tail assumptions stated in Appendix A.1 and Supplementary Theorem 5.6.

Table 3: E4/E5/R4/R6 GRPO summary. Score-tail entries report fractional-boundary rarity with a 95% prompt-cluster bootstrap CI; zero-mass resamples are retained, so the upper endpoint can be inf. LR-tail ∞ means no base trace entered the estimated half-mass tail, while $--$ means the adapter was unavailable. The matched-tilt residual share is the plug-in fraction of measured KL left after the score-matched one-parameter tilt; $--$ marks an undefined match or the base anchor. The censored flag marks a point outside the finite exact-tilt window or a zero-hit score-tail point estimate. The low/mid tags distinguish the corresponding learning-rate configurations listed in Appendix F.

run	checkpoint	step	$\hat{d} \pm$ SE	$\mathbb{E}[S]$	$\mathbb{E}[m]$	score tail [CI]	LR tail	resid. (%)	cens.
$\beta = .002$	ckpt 200	200	91.30 ± 0.13	11.59	0.525	4.55 [4.15, 5.05]	inf	–	yes
$\beta = .002$	final	400	163.99 ± 0.12	13.73	0.547	6.73 [6.18, 7.53]	inf	–	yes
$\beta = .01$	ckpt 200	200	256.31 ± 0.48	14.19	0.529	7.60 [6.34, inf]	inf	–	yes
$\beta = .01$	final	400	245.12 ± 0.30	14.72	0.534	9.47 [8.21, inf]	inf	–	yes
$\beta = .05$	ckpt 200	200	22.35 ± 0.07	7.81	0.571	2.43 [2.30, 2.57]	inf	97.2	yes
$\beta = .05$	ckpt 300	300	31.62 ± 0.09	8.63	0.578	2.96 [2.77, 3.14]	–	97.0	yes
$\beta = .05$	final	400	29.82 ± 0.09	8.66	0.579	2.98 [2.82, 3.18]	inf	96.7	yes
$\beta = .1$ (low)	ckpt 10	10	0.19 ± 0.00	3.3	0.501	0.74 [0.72, 0.75]	1.02	99.8	no
$\beta = .1$ (low)	ckpt 20	20	0.53 ± 0.01	3.62	0.515	0.81 [0.80, 0.84]	1.62	99.1	no
$\beta = .1$ (low)	ckpt 30	30	1.10 ± 0.01	3.91	0.527	0.88 [0.85, 0.91]	2.37	98.9	no
$\beta = .1$ (low)	ckpt 40	40	1.04 ± 0.01	3.92	0.523	0.90 [0.87, 0.93]	2.34	98.8	no
$\beta = .1$ (low)	ckpt 50	50	0.96 ± 0.01	3.93	0.524	0.90 [0.87, 0.93]	2.22	98.6	no
$\beta = .1$ (low)	final	60	0.96 ± 0.01	3.93	0.522	0.89 [0.87, 0.92]	2.25	98.6	no
$\beta = .05$ (mid)	ckpt 10	10	0.83 ± 0.01	3.87	0.522	0.87 [0.85, 0.89]	2.05	98.7	no
$\beta = .05$ (mid)	ckpt 20	20	2.71 ± 0.02	4.59	0.539	1.05 [1.02, 1.09]	4.15	98.2	no
$\beta = .05$ (mid)	ckpt 30	30	2.89 ± 0.02	4.68	0.546	1.10 [1.06, 1.14]	4.41	98.1	no
$\beta = .05$ (mid)	ckpt 40	40	2.71 ± 0.02	4.66	0.54	1.08 [1.03, 1.12]	4.17	98.1	no
$\beta = .05$ (mid)	ckpt 50	50	3.01 ± 0.02	4.77	0.543	1.13 [1.08, 1.18]	4.47	98.0	no
$\beta = .05$ (mid)	ckpt 60	60	3.50 ± 0.02	4.9	0.543	1.16 [1.12, 1.21]	4.89	97.9	no
$\beta = .05$ (mid)	ckpt 70	70	3.48 ± 0.02	4.84	0.544	1.17 [1.12, 1.21]	4.92	98.1	no
$\beta = .05$ (mid)	ckpt 80	80	3.46 ± 0.02	4.86	0.545	1.16 [1.11, 1.21]	4.84	98.0	no
$\beta = .05$ (mid)	final	80	3.50 ± 0.01	4.89	0.54	1.17 [1.12, 1.22]	4.93	97.9	no
base	base	0	0.16 ± 0.00	3.18	0.493	–	–	–	–

Figure 11 separates three claims. First, among the non-saturating jittered-score conditional tilts we measured, score-tail rarity follows optimization KL with near-unit slope, consistent with log-ratio concentration. The token-capped length score departs as its original-score ceiling is approached; because $t\sigma$ is 1.26 and 3.26 at $d = 5, 6$, those two points describe optimization of the jittered continuation rather than further optimization of original length. Second, GRPO checkpoints lie below the exact-tilt *score-tail* curves, and the matched 768-token replication confirms this score-tail gap. The reported LR-tail estimates do not show the same gap, although the likelihood-evaluation mismatch prevents a definitive common-sequence comparison. Third, the sign and magnitude of $\Delta_t(m)$ are proxy-dependent: at matched $d = 2$ the shift is $+0.058$ for the judge score, approximately zero for random scores, and -0.187 for token-capped length. Thus KL can price access to a score-emphasized region in the tilt regime without determining whether the gold-standard target improves.

E Mixed-Pool Robustness Run

Repeating E1 on the mixed GSM8K+MATH pool (786 problems, 8.05×10^5 traces, accuracy 0.811) with identical settings gives $\Delta_t(m) \approx +0.03$, a nontrivial-point log-slope of 0.944, and $n_{\mathrm{req,base}} = 1{,}246{,}974$ already at $d = 4$. The proxy–target gap on this easy mixture is much smaller (top-decile error rate 0.009), so we use it only as a robustness check: the coverage scaling is visible even when the exploitable gap is thin, but main-text claims are made on the hard-MATH pool where the audit problem is non-degenerate ($|\Delta_t|$ comfortably above ϵ at every grid point, with a sizable proxy–target gap).

Table 4: Prompt-cluster significance checks for reported improvements/decreases. Intervals are 95% prompt-cluster bootstrap CIs; p-values are paired sign-flip randomization tests over prompts. Directional tests use the direction stated by the row name (accuracy/shift/error-reduction greater than zero); two-sided p-values are also reported. P-values are unadjusted for multiple comparisons. The GRPO peak contrasts are post-selected, and their intervals and p-values are not selection-adjusted.

family	contrast	d	estimate [95% CI]	p_{dir}	$p_{2\mathrm{s}}$	clusters
GRPO acc.	peak > base	–	+0.0856 [+0.0639, +0.1082]	$< 10^{-4}$	$< 10^{-4}$	203
GRPO acc.	peak > high-KL final	–	+0.0446 [+0.0221, +0.0687]	$< 10^{-4}$	$< 10^{-4}$	203
E7 shift	$\Delta m > 0$	0.5	+0.0543 [+0.0363, +0.0730]	$< 10^{-4}$	$< 10^{-4}$	203
E7 shift	$\Delta m_{\mathrm{faith}} > 0$	0.5	+0.1083 [+0.0899, +0.1277]	$< 10^{-4}$	$< 10^{-4}$	203
E7 shift	$\Delta m > 0$	1	+0.0553 [+0.0323, +0.0795]	$< 10^{-4}$	$< 10^{-4}$	203
E7 shift	$\Delta m_{\mathrm{faith}} > 0$	1	+0.1376 [+0.1139, +0.1630]	$< 10^{-4}$	$< 10^{-4}$	203
E7 shift	$\Delta m > 0$	2	+0.0473 [+0.0175, +0.0786]	0.002	0.002	203
E7 shift	$\Delta m_{\mathrm{faith}} > 0$	2	+0.1599 [+0.1295, +0.1920]	$< 10^{-4}$	$< 10^{-4}$	203
E7 shift	$\Delta m > 0$	3	+0.0379 [+0.0025, +0.0745]	0.020	0.042	203
E7 shift	$\Delta m_{\mathrm{faith}} > 0$	3	+0.1657 [+0.1300, +0.2023]	$< 10^{-4}$	$< 10^{-4}$	203
E7 shift	$\Delta m > 0$	4	+0.0301 [-0.0106, +0.0703]	0.074	0.146	203
E7 shift	$\Delta m_{\mathrm{faith}} > 0$	4	+0.1654 [+0.1236, +0.2076]	$< 10^{-4}$	$< 10^{-4}$	203
R7 monitor	prompt+trace reduces L^1	1	+0.0216 [+0.0142, +0.0291]	$< 10^{-4}$	$< 10^{-4}$	200
R7 monitor	prompt+trace reduces L^1	2	+0.0174 [+0.0060, +0.0288]	0.002	0.003	200
R7 monitor	prompt+trace reduces L^1	4	+0.0082 [+0.0061, +0.0104]	$< 10^{-4}$	$< 10^{-4}$	200
R7 monitor	prompt+trace reduces L^1	6	+0.0001 [+0.0001, +0.0001]	$< 10^{-4}$	$< 10^{-4}$	200

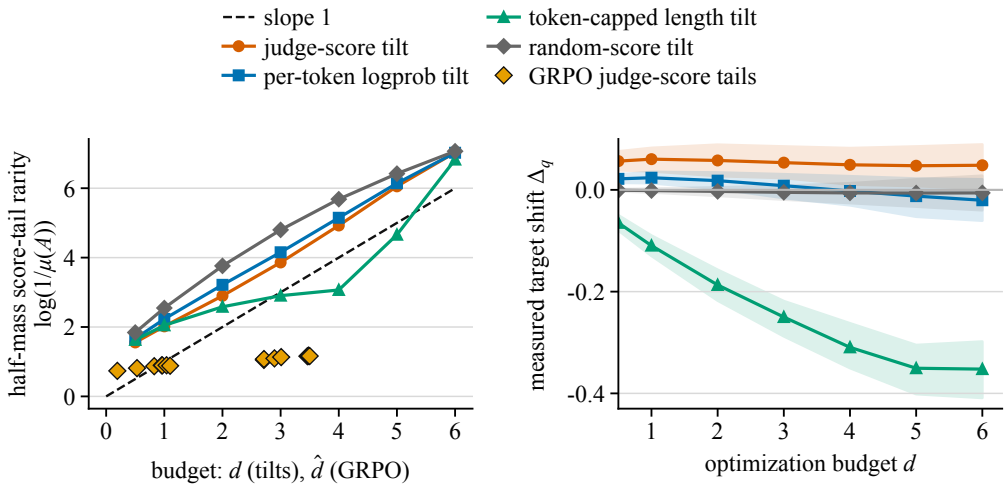


Figure 11: When KL predicts the measured score region. Left: per-prompt upper-score-region rarity versus optimization KL. All tilt series use the pipeline's deterministically jittered scores and fractional-boundary half-mass tests; for token-capped length, $d = 5, 6$ exceed the original-score finite-pool ceiling 4.61 and are only jittered-score continuations. GRPO fits its test on policy samples and evaluates it on an independent 768-token base pool. Judge-score, per-token log-probability, and random-score tilts follow near-unit-slope curves; length departs as its bounded original score saturates. The broad GRPO overlay lies below the tilt curves in score-tail rarity; same-protocol and LR-tail comparisons are in Figure 4. Right: the measured target shift at matched d depends on proxy–target alignment, ranging from positive (judge) through near zero (random) to negative (length).

F EXPERIMENTAL DETAILS

Judging and grading. The judge sees the problem and candidate trace (no answer key) and answers one token; the implementation truncates only traces exceeding the documented character safety cap needed to fit the judge context. The score S aggregates Yes/No variants in the top-20 first-token log-probabilities and is clipped to ± 16, with a defined $S{=}0$ fallback if neither token appears. Before exponential reweighting, the E1 coverage/universality and matched-protocol R3/R4 tilt analyses add fixed-seed Gaussian jitter with $\sigma = 10^{-4}$. This changes the optimized score rather than merely randomizing tail-boundary membership: when $t\sigma$ is not small, originally tied traces can receive materially different weights. For the main 7B E1 curve $t\sigma \leq 6.6 \times 10^{-4}$ through $d = 6$; the high-KL 14B and length exceptions are identified above. The E3/R5/R7 monitor analyses use the recorded clipped scores without jitter. E5 also uses unjittered scores and resolves boundary ties by fractional membership so that fitted empirical policy mass is exactly one half. In the 768-token baseline pool, 1.25% of scores reach the clip. Grading extracts the final answer (boxed or "Final answer" marker, with fraction and unit handling) and compares it with the strict-parsed gold answer.

E1 estimators. Baseline binning uses per-prompt bin edges from the union of P- and Q_t-quantile grids (12 bins/prompt; cell masses known to the auditor per the Theorem 1(b) protocol), prior fallback $\frac{1}{2}$ on unlabeled cells; n_{req} is found by geometric bracketing with 500 Monte-Carlo replications per probe and refinement to a factor of 1.25. All counts are with-replacement draws from the empirical oracle, not distinct pool traces. Global bins create asymptotic estimand bias $+0.11$–0.30, whereas prompt-conditional bias stays below 0.002: the reported asymptotic (infinite-label) value of the binned estimand differs from $\Delta_t(m)$ by less than 0.002 at every d.

E2 measurement. Sampler generation produced a larger pool, but the cost and hit-rate calculations retain only traces whose problems belong to the final 203-prompt hard-MATH pool. On this matched prompt mixture, sampler cost $\widehat{\kappa}$ averages $\log \nu - \log \mu$, where $\log \nu$ is the generation-time (temperature-processed) sequence log-probability and $\log \mu$ is a teacher-forced rescoring of the *same generated token sequence* under the original base prompt at temperature 1. An instruction suffix is thus a property of the sampling policy, not of the episode variable, and $\widehat{\kappa}$ estimates a KL between two distributions on the common space of generated traces for a given problem; no suffixed prompt ever enters the episode, so the comparison is not between mutually singular prompt laws. Standard errors are at most 0.47 nats (at temperature 1.3) and below 0.06 otherwise. Tail cutoffs s_t are per-prompt q_t-weighted score medians, but E2 uses the inclusive events $\{S \geq s_t\}$. Their realized mixture q_t-masses are 0.510, 0.586, and 0.783 at $d = 2, 4, 6$, respectively; hit-to-confidence draw counts use these operational events in $(1 - \nu(A))^n \leq 2\delta$ with $\delta = 0.1$.

E3/R5/R7 training and certification. Monitors regress pool-exact g (from the full-pool tilt) with L^1 loss, Adam, 150 epochs, within-prompt 30% holdout (200 prompts $\times$ 256 traces; MLP widths 64–1024); embeddings are mean-pooled last-hidden states of frozen Qwen2.5-1.5B-Instruct. The fine-tuned-LM rung is Qwen2.5-1.5B with LoRA rank $r{=}16$ (dropout 0.1) and a scalar head on the last token, AdamW (weight decay 0.05). It runs all six epochs and reports the test value from the epoch with minimum validation error; the validation split holds out approximately 12% of training traces separately within each prompt, and the holdout test set is never used for selection. Prompt–trace embeddings concatenate the question and trace and use the identical split. R5 solves the finite-sample L^1 primal and dual with HiGHS after verifying feature/index alignment, using the unrenormalized restriction of the full-pool g described above. The auxiliary saved E3 sign-residual diagnostic is heuristic: it does not impose all feature-moment constraints and is not used for a theorem-linked claim. Monitors never see S or the judge except in the positive control.

E5/R2/R3/R6 measurement. E5 defines a fractional-boundary upper-score test per prompt, fitted to have empirical policy mass exactly $1/2$, and evaluates that same test on an independent base-policy pool generated under the identical 768-token protocol (128 traces per prompt). The E5 value of $\hat{d}$ is recomputed on the same judged subset. Its prompt-cluster bootstrap retains zero-mass resamples and therefore permits an infinite rarity upper endpoint. R2 alternates even/odd trace-id folds for fitting and held-out evaluation of the same fractional test. R3 computes exact fractional-boundary half-mass tails on the same 768-token base pool. R6 instead averages independent-base masses from thresholds fitted separately on even and odd policy folds; it does not evaluate either threshold's

mass on the opposite policy fold. For R6 baseline scoring, each adapter and the base model are teacher-forced on the same retokenized baseline sequence. The policy-side threshold uses stored generation-time policy log probability and stored base-model rescoring of the original generated token IDs; the preceding R6 paragraph reports both limitations. E7 labels are obtained by thresholding at zero the clipped (± 16) first-token Yes/No log-odds produced by a separate Qwen2.5-14B checker.

E4 configuration. GRPO with LoRA adapters (rank r=32, scaling α=64, all attention and MLP projections), group size 8, 8 prompts per step (64 completions), completion cap 768 tokens, generation temperature 1.0/top-p 1.0 (matching μ). Training used a clipped full-softmax Yes/No log-odds reward from the same frozen judge and prompt on a dedicated GPU. All *reported* S values—for tilt pools and policy pools alike—were recomputed uniformly with the top-20 first-token evaluation estimator. The training reward and reported score are closely related but not definitionally identical. Runs: $\beta \in \{0.05, 0.01, 0.002\}$ at learning rate 5×10^{-5} for 400 steps, plus gentler runs at β=0.1, lr 10^{-5} (60 steps) and β=0.05, lr 2×10^{-5} (80 steps) to place checkpoints inside the tilt window; checkpoints saved every 10–100 steps. Policy pools contain 128 samples per problem for checkpoints and 256 for final adapters; each record stores generation-time $\log \pi_\theta$ and a teacher-forced base-policy $\log \mu$ for the same token sequence.

Hyperparameter development and selection. The anonymized code archive, driver scripts, frozen aggregate results, and sanitized reproducibility manifest record the reported sweeps, commands, and selection rules. Simulation accuracy and confidence parameters were fixed before probing archive sizes; E1 used $\epsilon = 0.025$, $\delta = 0.2$, 12 prompt-conditional bins, and 500 replications per probe. E2 reports all seven prespecified practical sampler configurations rather than selecting a winner. E3 reports every stated linear, polynomial, MLP-width, and language-model rung. Only the LoRA-LM training epoch is selected on a within-training validation split; the held-out test traces are not used for that selection. Curves labeled "best" over the surface/embedding ladders are descriptive minima across displayed held-out fits, not a separately selected final model. The GRPO β and learning-rate configurations above were chosen to cover a broad realized-KL range, and all available evaluated checkpoints are reported rather than selecting by test accuracy. The compact submitted bundle preserves paper-used defaults and overrides in the scripts and records package profiles and hashes of the frozen result tree; large raw pools and checkpoints are excluded as documented in the archive.

Independent runs and resampling. Unless stated otherwise, each model-training configuration was run once with its fixed seed; reported checkpoints from the same trajectory are longitudinal measurements, not independent training restarts. E0 panel (a) uses 400 Monte-Carlo replications per operating point, while its single-swap block-KL panel uses 400,000 Gaussian draws per displayed budget. E1 uses 500 replications per archive-size probe and 1,000 prompt-cluster bootstrap resamples per target-KL point; the seed-sensitivity check repeats the E1 estimator simulation for seeds 2026–2028. E2 generates a separate sample pool for each of the seven audit samplers, using sampler- and prompt-specific deterministic seeds; analysis filters each pool to the final hard-MATH prompt mixture and uses 200 Monte-Carlo replications for each best-of-k hit-rate point. E3 fits each monitor architecture once at each of the four target values $d \in \{1, 2, 4, 6\}$ on one fixed train/holdout split; the LoRA-LM and prompt–trace rungs likewise use one fit per target value. E4 consists of five GRPO training runs, one for each configuration listed above; the reported 22 checkpoints are evaluations along those runs. E1 universality, E5, R2, and E7 use 500 prompt-cluster bootstrap resamples. T8 uses 20,000 prompt-level sign flips and 20,000 prompt-bootstrap resamples per comparison. The remaining deterministic reanalyses and certificates (R1 and R3–R6) make one pass over each fixed artifact or checkpoint rather than averaging independent reruns.

Code availability. An anonymized code archive accompanies the submission. Upon publication, we will release the complete preprocessing, generation, training, evaluation, analysis, and figure-generation code, together with the exact experiment configurations, under the MIT License.

Infrastructure. The submitted environment profiles document two Linux GPU machines. Machine A had eight NVIDIA B200 GPUs (183,359 MiB each; driver 580.95.05) and two Intel Xeon 6960P sockets (72 physical cores per socket, 288 logical CPUs). Machine B had four NVIDIA H100 80GB GPUs; its CPU model and driver were not retained in the frozen run metadata and are

therefore not inferred here. Original R6 likelihood-ratio teacher forcing and R7 prompt–trace embedding/monitor training used Machine B. The T8 statistical-test run, including its R7 refit, used one B200; its frozen metadata records Linux 6.8.0, Python 3.11.11, PyTorch 2.9.1+cu130, and Transformers 4.57.6. The submitted result-producing B200 profile pins CUDA 13.0.2, Python 3.11.15, vLLM 0.22.1, PyTorch 2.11.0, and Transformers 5.13.0; the historical H100 profiles pin vLLM 0.22.1, PyTorch 2.11.0, and Transformers 4.57.6. E0/E1/E2 analyses, reanalyses, the KL decomposition, and LP certificates are CPU-only. Submitted environment profiles, driver scripts, compact statistical-test metadata, and reproducibility manifests record package versions, seeds, command structure, artifact hashes, and GPU snapshots where available. Random seeds are fixed and recorded for prompt sampling, generation, judging, Monte-Carlo probes, bootstrap, and train/holdout splits.

G Complete Proofs and Supporting Lemmas

This section contains the complete, self-contained proofs underlying the main and deferred theoretical claims. The shorter roadmaps above summarize the same arguments; this section supplies all assumptions, boundary cases, auxiliary lemmas, and oracle-model qualifications.

G.1 Common Setup and Elementary Tools

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space. Let $q \ll \mu$, let

$$W = \frac{dq}{d\mu}, \qquad g = W - 1,$$

and note

$$E_\mu W = 1, \qquad E_\mu g = 0, \qquad E_\mu |g| \le 2.$$

A target mean is a measurable function $m : \Omega \to [0, 1]$. We write its optimized mean as

$$R_q(m) := E_q m.$$

The optimized-distribution shift is

$$\Delta_q(m) = R_q(m) - E_\mu m = E_\mu[gm].$$

We use

$$\mathrm{kl}(a\|b) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}$$

for binary relative entropy, and

$$h_2(a) = -a \log a - (1-a) \log(1-a)$$

for binary entropy.

Lemma 3.1 (Binary KL lower bound).

For $a \in [0, 1]$ and $b \in (0, 1)$, with the usual boundary conventions,

$$\mathrm{kl}(a\|b) \ge a \log \frac{1}{b} - h_2(a).$$

Proof. For $a \in (0,1)$, expand:

$$\mathrm{kl}(a\|b) = a\log\frac{1}{b} - h_2(a) + (1-a)\log\frac{1}{1-b}.$$

The final term is nonnegative. The endpoint cases follow by continuity.

□

Lemma 3.2 (Hit probability under baseline sampling).

If $O_1, ..., O_n$ are iid from μ and $p = \mu(A)$, then for

$$H = \{\textit{some } O_i \in A\}$$

we have

$$p_n := \mu^n(H) = 1 - (1-p)^n \le np.$$

Proof. The identity follows from independence; the inequality is the union bound.

□

Lemma 3.3 (Data processing to a hit event).

If $P \ll M$ are probability laws on a transcript space and H is measurable, then

$$D(P\|M) \ge \mathrm{kl}(P(H)\|M(H)).$$

Proof. Apply the data-processing inequality for relative entropy to the measurable map $z \to \mathbf{1}_H(z)$.

□

Lemma 3.4 (Two separated targets on a common event).

Let P_0, P_1 be two laws, let E be an event such that the restricted finite measures agree,

$$P_0(\cdot \cap E) = P_1(\cdot \cap E),$$

and let the two target values be θ_0, θ_1 with $|\theta_1 - \theta_0| > 2\epsilon$. If an estimator T satisfies

$$P_j(|T - \theta_j| \le \epsilon) \ge 1 - \delta, \qquad j = 0, 1,$$

then

$$P_0(E) = P_1(E) \le 2\delta.$$

Proof. Let $B_j = |T - \theta_j| \le \epsilon$. The sets B_0 and B_1 are disjoint because the two closed epsilon-balls are disjoint. Write $r = P_0(E) = P_1(E)$. Since $P_j(B_j^c) \le \delta$,

$$P_j(E \cap B_j) \ge r - \delta.$$

On E, the two laws agree, and B_0, B_1 are disjoint, so

$$P_0(E \cap B_0) + P_1(E \cap B_1) = P_0(E \cap B_0) + P_0(E \cap B_1) \le r.$$

Thus $2r - 2\delta \le r$, hence $r \le 2\delta$. Internal estimator randomness can be included in the transcript, so the argument also covers randomized estimators.

□

Lemma 3.5 (Expected-risk version).

Under the assumptions of Lemma 3.4, for every estimator T,

$$\max_{j=0,1} E_j|T - \theta_j| \ge \frac{|\theta_1 - \theta_0|}{2} P_0(E).$$

Proof. On the common event E, the triangle inequality gives

$$|T - \theta_0| + |T - \theta_1| \ge |\theta_1 - \theta_0|.$$

Integrate with respect to the common restricted measure and divide by two.

□

Lemma 3.6 (Binary-KL cost of forcing a minimum event probability).

For $b \in (0,1)$ and $\rho \in [0,1]$, define

$$J_b(\rho) = \inf_{a \in [\rho,1]} \mathrm{kl}(a\|b).$$

Then

$$J_b(\rho) = \begin{cases} 0, & \rho \le b, \\ \mathrm{kl}(\rho\|b), & \rho > b. \end{cases}$$

Proof. For $a \in (0,1)$,

$$\frac{\partial}{\partial a} \mathrm{kl}(a\|b) = \log \frac{a(1-b)}{b(1-a)}.$$

The derivative is negative on $(0,b)$, zero at b, and positive on $(b,1)$. Hence the unconstrained minimum is attained at $a = b$. If $\rho \le b$, this point is feasible and the minimum is zero; if $\rho > b$, the objective is increasing on $[\rho,1]$ and the minimum is attained at ρ. Boundary cases follow by continuity.

□

H Complete Proofs: Passive access

H.1 No-hit lower bound and canonical hard regions

Passive minimax sample-complexity notation

Fix one of the two labeled baseline experiments used in this part:

1. Noiseless: $Y = f(S)$ almost surely.

2. Bernoulli: conditionally on $S = s$, Y is Bernoulli with mean $f(s)$, independently across observations.

For $f \in \mathcal{F}$, let $\mathbb{P}^{(n)}_{P,f,\mathsf{E}}$ denote the joint law of n iid labeled observations $(S_i, Y_i)_{i=1}^n$ from the fixed experiment E, with $S_i \sim P$, together with an independent $U \sim \mathrm{Unif}[0,1]$ available as a random seed. Put

$$\theta_{P,Q}(f) := E_Q f - E_P f.$$

With $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$, define

$$\begin{aligned} &n^*_{\mathrm{base}}(P, Q, \epsilon, \delta; \mathcal{F}, \mathsf{E}) \\ &:= \inf\Big\{ n \in \mathbb{N}_0 : \ \exists\, \widehat{\theta}_n \text{ measurable}, \\ &\qquad \widehat{\theta}_n = \widehat{\theta}_n((S_i, Y_i)_{i=1}^n, U), \\ &\qquad \sup_{f \in \mathcal{F}} \mathbb{P}^{(n)}_{P,f,\mathsf{E}}\big(|\widehat{\theta}_n - \theta_{P,Q}(f)| > \epsilon\big) \le \delta \Big\}. \end{aligned}$$

Use the convention $\inf \varnothing = \infty$. For no-hit events with zero observations, use the empty-product convention

$$(1-p)^0 := 1 \qquad \text{for every } p \in [0,1],$$

including the formal boundary $p = 1$. When P, Q, and the fixed label experiment are clear, abbreviate this as

$$n^*_{\mathrm{base}}(\epsilon, \delta; \mathcal{F}).$$

For an exponential-tilt family Q_t, write

$$n^*_{\mathrm{base}}(t, \epsilon, \delta; \mathcal{F}) := n^*_{\mathrm{base}}(P, Q_t, \epsilon, \delta; \mathcal{F}, \mathsf{E}).$$

Theorem 4.1 (Passive no-hit minimax lower bound).

Let $0 < \delta < 1/2$. Let P be a baseline score law and Q a target score law. The auditor observes iid scores $S_i \sim P$ and labels generated in one of two worlds f_0, f_1. Labels are either noiseless, $Y_i = f_j(S_i)$, or conditionally independent Bernoulli with mean $f_j(S_i)$. Suppose there is a measurable set A such that

$$f_0 = f_1 \quad \text{on } A^c, \qquad p = P(A),$$

and

$$|\theta(f_1) - \theta(f_0)| > 2\epsilon, \qquad \theta(f) = E_Q f - E_P f.$$

Any estimator that succeeds with probability at least $1 - \delta$ in both worlds obeys

$$(1-p)^n \le 2\delta.$$

If $p = 0$, no finite-sample estimator can satisfy the two success requirements. If $0 < p \le 1/2$, then

$$n \ge \frac{1}{2p} \log \frac{1}{2\delta}.$$

Proof. Let

$$E_n = \{S_1, ..., S_n \notin A\}.$$

Under either world, $P(E_n) = (1-p)^n$. On E_n, every conditional label law is the same in the two worlds because $f_0 = f_1$ outside A. Thus the restricted transcript laws agree. Lemma 3.4 yields $(1-p)^n \leq 2\delta$.

If $p = 0$, the inequality $(1-p)^n \leq 2\delta$ contradicts $\delta < 1/2$, so no finite n works. For $p > 0$, taking logarithms gives

$$n[-\log(1-p)] \geq \log \frac{1}{2\delta}.$$

For $0 < p \leq 1/2$, $-\log(1-p) \leq p/(1-p) \leq 2p$. Therefore

$$n \geq \frac{1}{2p} \log \frac{1}{2\delta}.$$

□

Corollary 4.2 (Audit coverage modulus; exact and small-mass forms).

Fix $0 < \delta < 1/2$. Define the strict and non-strict moduli

$$\mathfrak{C}^{>}_{\mathcal{F},Q}(\gamma) = \inf_{\substack{A,\ f_0,f_1 \in \mathcal{F};\\ f_0 = f_1 \text{ on } A^c;\\ |\theta(f_1) - \theta(f_0)| > \gamma}} P(A),$$

and

$$\mathfrak{C}^{\geq}_{\mathcal{F},Q}(\gamma) = \inf_{\substack{A,\ f_0,f_1 \in \mathcal{F};\\ f_0 = f_1 \text{ on } A^c;\\ |\theta(f_1) - \theta(f_0)| \geq \gamma}} P(A).$$

If a feasible set is empty, its infimum is ∞ and the corresponding statement below is vacuous. For $c \in [0,1]$, define

$$\Psi_\delta(c) = \begin{cases} +\infty, & c = 0, \\ \dfrac{\log(1/(2\delta))}{-\log(1-c)}, & 0 < c < 1, \\ 1, & c = 1. \end{cases}$$

If the strict-modulus feasible set is nonempty, then

$$n^*_{\text{base}}(\epsilon, \delta; \mathcal{F}) \geq \Psi_\delta\left(\mathfrak{C}^{>}_{\mathcal{F},Q}(2\epsilon)\right).$$

In particular, if

$$\mathfrak{C}^{>}_{\mathcal{F},Q}(2\epsilon) \leq \frac{1}{2},$$

then

$$n^*_{\text{base}}(\epsilon, \delta; \mathcal{F}) \geq \frac{1}{2\mathfrak{C}^{>}_{\mathcal{F},Q}(2\epsilon)} \log \frac{1}{2\delta},$$

with $+\infty$ understood when the modulus is zero.

For the non-strict modulus, fix any $\tau > 0$ and put

$$C_\tau := \mathfrak{C}^{\geq}_{\mathcal{F},Q}(2\epsilon + \tau).$$

If its feasible set is nonempty, then the globally valid statement is

$$n^*_{\mathrm{base}}(\epsilon, \delta; \mathcal{F}) \geq \Psi_\delta(C_\tau).$$

Consequently, the simpler bound

$$n^*_{\mathrm{base}}(\epsilon, \delta; \mathcal{F}) \geq \frac{1}{2C_\tau} \log \frac{1}{2\delta}$$

is asserted only when $C_\tau \leq 1/2$, again with the infinite interpretation at $C_\tau = 0$.

Proof. Consider first either a strict modulus at threshold 2ϵ or a non-strict modulus at threshold $2\epsilon + \tau$. In both cases every feasible pair has target separation strictly larger than 2ϵ. Let C be the corresponding modulus and choose feasible triples $(A_k, f_{0,k}, f_{1,k})$ with

$$p_k := P(A_k) \downarrow C.$$

For every k, Theorem 4.1 gives

$$(1 - p_k)^n \leq 2\delta.$$

If some $p_k = 0$, no finite n can succeed. For $0 < p_k < 1$, this is equivalent to

$$n \geq \frac{\log(1/(2\delta))}{-\log(1 - p_k)}.$$

Passing to the limit proves $n \geq \Psi_\delta(C)$ for $0 < C < 1$. If $C = 0$, either a feasible triple already has $p = 0$, or feasible masses decrease to zero; in either case no finite n can satisfy all the no-hit inequalities, so $\Psi_\delta(0) = +\infty$. If $C = 1$, every feasible triple has $P(A) = 1$. By the empty-product convention, $n = 0$ gives $(1 - P(A))^0 = 1 > 2\delta$, so Theorem 4.1 forces $n \geq 1 = \Psi_\delta(1)$. Finally, when $0 < C \leq 1/2$,

$$-\log(1 - C) \leq \frac{C}{1 - C} \leq 2C,$$

so the exact lower bound implies the displayed $1/(2C)$ simplification.

□

Proposition 4.3 (Neyman-Pearson characterization of the low-overlap region).

Let $q \ll \mu$, $W = dq/d\mu$, and $\eta \in (0, 1)$. Define the relaxed baseline mass

$$\alpha_\eta(q, \mu) = \inf\{E_\mu \phi : 0 \leq \phi \leq 1,\ E_q \phi \geq \eta\}.$$

There exist a threshold $c \in (0, \infty)$ and a boundary randomization $\gamma \in [0, 1]$ such that

$$\phi^* = 1_{\{W > c\}} + \gamma 1_{\{W = c\}}, \qquad E_q \phi^* = \eta,$$

and ϕ^* attains the infimum. If W has no atom at the threshold, the optimizer is the event $W \geq c$.

For an exponential score tilt with finite normalizer,

$$\frac{dQ_t}{dP}(s) = e^{ts-\Lambda(t)},$$

W is increasing in s when $t > 0$ and decreasing when $t < 0$. Thus the optimizer is an upper score tail for $t > 0$ and a lower score tail for $t < 0$, with boundary randomization when the score law has an atom. At $t = 0$, W is constant and the relaxed optimizer is realized entirely by boundary randomization.

Proof. Because $q(W > 0) = 1$, $q(W > c)$ increases to one as c decreases to zero; because $W < \infty$ q-almost surely, it decreases to zero as c tends to infinity. Since $\eta \in (0, 1)$, a finite positive threshold c and a boundary randomization γ can be chosen so that $E_q\phi^* = \eta$. For any feasible ϕ, pointwise

$$(\phi - \phi^*)(W - c) \leq 0 :$$

on $W > c$, $\phi^* = 1$; on $W < c$, $\phi^* = 0$; and on the boundary the product is zero. Integrating under μ,

$$E_q(\phi - \phi^*) - cE_\mu(\phi - \phi^*) \leq 0.$$

The first term is nonnegative because $E_q\phi \geq \eta = E_q\phi^*$. Since $c > 0$, it follows that $E_\mu\phi \geq E_\mu\phi^*$. The sign of t determines whether the likelihood ratio is increasing or decreasing in the score, giving the stated upper- or lower-tail form; at $t = 0$ the likelihood ratio is constant.

□

Proposition 4.4 (Prompt-conditional versus unrestricted rarity).

Fix $\eta \in (0, 1)$. Suppose the prompt and output spaces are standard Borel,

$$\mu(dx, dz) = \pi(dx)\mu_x(dz), \qquad q(dx, dz) = \pi(dx)q_x(dz),$$

and the conditional laws are measurable kernels. Define the measurable prompt-conditional value by

$$\alpha_\eta^{\mathrm{cond}} = \inf_\phi \big\{ E_\mu\phi : 0 \leq \phi(x, z) \leq 1, \\ E_{q_x}\phi(x, \cdot) \geq \eta \text{ for } \pi\text{-a.e. } x \big\}.$$

Then

$$\alpha_\eta(q, \mu) \leq \alpha_\eta^{\mathrm{cond}}, \qquad -\log \alpha_\eta^{\mathrm{cond}} \leq -\log \alpha_\eta(q, \mu).$$

If measurable per-prompt Neyman-Pearson optimizers exist (as they do under the usual jointly measurable likelihood-ratio/quantile assumptions), then

$$\alpha_\eta^{\mathrm{cond}} = E_{X\sim\pi}\alpha_\eta(q_X, \mu_X).$$

Proof. Every jointly measurable promptwise-feasible test is feasible for the unrestricted joint problem after averaging over prompts, because

$$E_q\phi = E_\pi E_{q_X}\phi(X, \cdot) \geq \eta.$$

The unrestricted infimum is therefore no larger. Moreover, for every promptwise-feasible ϕ, the section $\phi_x = \phi(x, \cdot)$ is feasible in the pointwise problem for π-almost every x, so

$$E_{\mu_x}\phi_x \geq \alpha_\eta(q_x, \mu_x) \qquad \text{for } \pi\text{-a.e. } x.$$

Integrating gives

$$E_\mu \phi \geq E_\pi \alpha_\eta(q_X, \mu_X),$$

and hence $\alpha_\eta^{\mathrm{cond}}$ is at least the displayed average. Under the stated measurable-selection assumption, measurable exact optimizers, or measurable ϵ-optimizers followed by $\epsilon \downarrow 0$, give the reverse inequality. This proves the integral identity.

□

H.2 One-sided exponential-power analysis

Throughout this section,

$$p_\alpha(s) = Z_\alpha^{-1} e^{-s^\alpha/\alpha} 1_{\{s \geq 0\}}, \qquad \alpha > 1,$$

and P denotes the corresponding law. Its exponential tilt is

$$\frac{dQ_t}{dP}(s) = e^{ts - \Lambda(t)}, \qquad \Lambda(t) = \log E_P e^{tS}.$$

Write

$$s_t = t^{1/(\alpha-1)}, \qquad \lambda_t = s_t^\alpha, \qquad w_t = ((\alpha-1)s_t^{\alpha-2})^{-1/2}.$$

Thus $V(s) = s^\alpha/\alpha$, $V'(s_t) = t$, and $w_t = V''(s_t)^{-1/2}$.

Lemma 5.1 (Exponential-power survival asymptotic).
As $x \to \infty$,

$$P(S \geq x) = \frac{1 + o(1)}{Z_\alpha x^{\alpha-1}} \exp\left(-\frac{x^\alpha}{\alpha}\right).$$

In particular,

$$\log \frac{1}{P(S \geq x)} = \frac{x^\alpha}{\alpha} + O(\log x).$$

Proof. Let $V(s) = s^\alpha/\alpha$. Integration by parts using

$$\frac{d}{ds} e^{-V(s)} = -s^{\alpha-1} e^{-V(s)}$$

gives

$$\int_x^\infty e^{-V(s)} ds = \frac{e^{-V(x)}}{x^{\alpha-1}} - (\alpha-1)\int_x^\infty s^{-\alpha} e^{-V(s)} ds.$$

The remainder is nonnegative and at most

$$(\alpha-1)x^{-\alpha}\int_x^\infty e^{-V(s)}ds.$$

If $I(x)$ denotes the tail integral, then

$$\frac{e^{-V(x)}}{x^{\alpha-1}} - (\alpha-1)x^{-\alpha}I(x) \le I(x) \le \frac{e^{-V(x)}}{x^{\alpha-1}}.$$

Rearranging the lower bound yields

$$I(x) \ge \frac{e^{-V(x)}x^{1-\alpha}}{1+(\alpha-1)x^{-\alpha}},$$

which matches the upper bound to relative error $1+o(1)$. Divide by Z_α.

□

Lemma 5.2 (Laplace scale, concentration, and KL).

As $t \to \infty$:

1. *The log-partition function satisfies*

$$\Lambda(t) = \frac{\alpha-1}{\alpha}s_t^\alpha + O(\log s_t).$$

2. *Under Q_t,*

$$E_{Q_t}|S-s_t| = O(w_t), \qquad \mathrm{Var}_{Q_t}(S) = O(w_t^2).$$

3. *For every $\eta \in (0,1)$, any upper-η quantile $a_{t,\eta}$ of Q_t satisfies*

$$a_{t,\eta} = s_t + O_\eta(w_t).$$

4. *The optimization KL divergence satisfies*

$$\begin{aligned} d(t) &:= D(Q_t\|P) \\ &= \frac{s_t^\alpha}{\alpha} + o(s_t^\alpha) \\ &= \frac{1}{\alpha}t^{\alpha/(\alpha-1)} + o\left(t^{\alpha/(\alpha-1)}\right). \end{aligned}$$

5. *The log-likelihood ratio concentrates at KL scale:*

$$E_{Q_t}\left|\log\frac{dQ_t}{dP} - d(t)\right| = O(\sqrt{d(t)}) = o(d(t)).$$

Proof. Put

$$h(u) = u - \frac{u^\alpha}{\alpha}, \qquad h_* = h(1) = \frac{\alpha-1}{\alpha}, \qquad \lambda_t = s_t^\alpha.$$

For $k = 0, 1, 2$, define the shifted Laplace integrals

$$I_{k,t} = \int_0^\infty |u-1|^k \exp\{\lambda_t(h(u)-h_*)\}\, du.$$

Because h has a unique nondegenerate maximum at 1, there are constants $r, c, C > 0$ such that

$$c(u-1)^2 \le h_* - h(u) \le C(u-1)^2 \qquad (|u-1| \le r).$$

On every compact subset of $[0,\infty)$ outside this neighborhood, $h_* - h$ has a strictly positive minimum. For sufficiently large u,

$$h(u) - h_* \le -c_1 u^\alpha$$

for a constant $c_1 > 0$. Therefore the local Gaussian integral gives

$$I_{0,t} \asymp \lambda_t^{-1/2}, \qquad I_{k,t} = O(\lambda_t^{-(k+1)/2}), \quad k = 1, 2.$$

For completeness, the lower bound on $I_{0,t}$ follows by integrating over $|u-1| \le \lambda_t^{-1/2}$. The local upper bounds follow from Gaussian integrals. The compact complement contributes $O(e^{-c_2\lambda_t})$, and the far tail contributes at most $\int_R^\infty u^k e^{-c_3\lambda_t u^\alpha} du$, which is exponentially smaller than the displayed powers.

Since

$$E_P e^{tS} = Z_\alpha^{-1} s_t e^{\lambda_t h_*} I_{0,t},$$

taking logarithms proves item 1.

Under Q_t, the scaled variable $U = S/s_t$ has density proportional to $e^{\lambda_t h(u)}$. Hence, for $k = 1, 2$,

$$E_{Q_t}|S - s_t|^k = s_t^k \frac{I_{k,t}}{I_{0,t}} = O(s_t^k \lambda_t^{-k/2}) = O(w_t^k),$$

because w_t is a fixed constant multiple of $s_t \lambda_t^{-1/2}$. This proves the first-moment bound and the second-moment bound; the latter implies the variance bound in item 2.

Chebyshev's inequality now gives

$$Q_t(|S - s_t| > M w_t) \le C/M^2.$$

For fixed η, choose M so that the two tails have total mass below $\min \eta, 1-\eta$. Every upper-η quantile must then lie in $[s_t - M w_t, s_t + M w_t]$, proving item 3.

Next,

$$d(t) = E_{Q_t}[tS - \Lambda(t)].$$

The first-moment bound gives

$$t E_{Q_t} S = t s_t + O(t w_t) = s_t^\alpha + O(s_t^{\alpha/2}),$$

while item 1 gives

$$\Lambda(t) = \frac{\alpha - 1}{\alpha} s_t^\alpha + O(\log s_t).$$

Both error terms are $o(s_t^\alpha)$, so item 4 follows.

Finally,

$$\log \frac{dQ_t}{dP} = tS - \Lambda(t)$$

has mean $d(t)$ under Q_t. By Cauchy-Schwarz and item 2,

$$\begin{aligned} E_{Q_t}\left|\log \frac{dQ_t}{dP} - d(t)\right| &\le t\sqrt{\mathrm{Var}_{Q_t}(S)} \\ &= O(tw_t) = O(s_t^{\alpha/2}) \\ &= O(\sqrt{d(t)}). \end{aligned}$$

□

Lemma 5.3 (Baseline mass near the tilted mode).

Fix $M \in (0, \infty)$. If

$$I_t = [s_t - Mw_t, s_t + Mw_t],$$

then uniformly for $s \in I_t$,

$$V(s) = V(s_t) + o(d(t)).$$

Moreover, if $J_t \subset I_t$ is an interval whose length ℓ_t obeys

$$|\log \ell_t| = o(d(t)),$$

then

$$P(J_t) = \exp[-d(t) + o(d(t))].$$

The same conclusion holds for a tail start

$$\begin{aligned} u_t &= s_t + O(w_t + 1), \\ P(S \ge u_t) &= \exp[-d(t) + o(d(t))]. \end{aligned}$$

Proof. For $s = s_t + u$ with $|u| \le Mw_t$, Taylor's theorem and local comparability of V'' give

$$\begin{aligned} |V(s) - V(s_t)| &\le t|u| + CV''(s_t)u^2 \\ &= O(tw_t) + O(1) \\ &= O(s_t^{\alpha/2}) = o(s_t^{\alpha}). \end{aligned}$$

By Lemma 5.2, $V(s_t) = s_t^{\alpha}/\alpha = d(t) + o(d(t))$. Therefore the density on J_t is between $\exp[-d(t) + o(d(t))]$ uniformly, and multiplying by a length whose logarithm is $o(d(t))$ gives the interval-mass claim.

For $u_t = s_t + O(w_t + 1)$, the same Taylor estimate has additional terms $O(t) + O(V''(s_t))$, both $o(s_t^{\alpha})$. Lemma 5.1 then gives

$$-\log P(S \ge u_t) = V(u_t) + O(\log u_t) = d(t) + o(d(t)).$$

□

Theorem 5.4 (Exponential-power passive minimax auditability phase).

Fix $\alpha > 1$, $L \in (0, \infty)$, and $\delta \in (0, 1/4)$. Assume baseline labels are conditionally Bernoulli with mean $f(S)$; the noiseless case is no harder. Let

$$\mathcal{F}_L = \{f : [0, \infty) \to [0, 1] : |f(s) - f(s')| \le L|s - s'|\}.$$

There is $\epsilon_0 = \epsilon_0(L, \alpha) > 0$ such that for every fixed $\epsilon \in (0, \epsilon_0)$,

$$\log n^*_{\text{base}}(t, \epsilon, \delta; \mathcal{F}_L) = d(t) + o(d(t))$$

as $t \to \infty$. For every fixed $\eta \in (0, 1)$, define the log-rarity

$$r_\eta(t) := -\log \alpha_\eta(Q_t, P),$$

where α_η is the Neyman–Pearson baseline mass from Proposition 4.3. Since these laws are atomless, if $a_{t,\eta}$ is chosen so that $Q_t(S \ge a_{t,\eta}) = \eta$, then

$$\alpha_\eta(Q_t, P) = P(S \ge a_{t,\eta}).$$

With this definition,

$$r_\eta(t) = d(t) + o(d(t)).$$

Proof: lower bound. Fix, for example, $\eta = 1/2$, and let $a_t = a_{t,\eta}$ be an upper-η quantile of Q_t. Define

$$u_t = a_t - \frac{1}{L}$$

and the ramp

$$\psi_t(s) = \begin{cases} 0, & s \le u_t, \\ L(s - u_t), & u_t < s < a_t, \\ 1, & s \ge a_t. \end{cases}$$

Then $\psi_t \in \mathcal{F}_L$. Take $f_0 = 0$ and $f_1 = \psi_t$. They agree outside

$$A_t = \{s > u_t\}.$$

Their target separation is

$$\begin{aligned} \theta_t(f_1) - \theta_t(f_0) &= E_{Q_t}\psi_t - E_P\psi_t \\ &\ge Q_t(S \ge a_t) - P(S > u_t) \\ &= \eta - o(1), \end{aligned}$$

because Lemma 5.3 gives $P(S > u_t) = \exp[-d(t) + o(d(t))]$. Hence, for any fixed $\epsilon < \eta/4$ and all large t, the separation exceeds 2ϵ. Theorem 4.1 gives

$$n^*_{\text{base}} \ge \frac{1}{2P(A_t)} \log \frac{1}{2\delta} = \exp[d(t) + o(d(t))].$$

Taking logarithms proves

$$\log n^*_{\text{base}} \geq d(t) + o(d(t)).$$

By Proposition 4.3 and atomlessness, $\alpha_\eta(Q_t, P) = P(S \geq a_{t,\eta})$; Lemma 5.3 therefore gives $r_\eta(t) = -\log \alpha_\eta(Q_t, P) = d(t) + o(d(t))$.

□

Proof: upper bound. We construct an estimator uniformly over $f \in \mathcal{F}_L$. Choose $M = M_\epsilon$ so that

$$Q_t(I_t^c) \leq \epsilon/8, \qquad I_t = [s_t - Mw_t, s_t + Mw_t],$$

which is possible by Lemma 5.2. Put

$$h = \frac{\epsilon}{8L}.$$

Partition I_t into B_t equal intervals $J_1, ..., J_{B_t}$ of diameter at most h and, when $|I_t| \geq h$, at least $h/2$. If $|I_t| < h$, use one bin. Then

$$B_t \leq 1 + \frac{2Mw_t}{h}, \qquad \log B_t = o(d(t)).$$

Let

$$p_{j,t} = P(J_j), \qquad q_{j,t} = Q_t(J_j), \qquad p_{\min,t} = \min_j p_{j,t}.$$

Every bin length has logarithm $o(d(t))$, so Lemma 5.3 gives, uniformly in j,

$$p_{j,t} = \exp[-d(t) + o(d(t))],$$
$$p_{\min,t} = \exp[-d(t) + o(d(t))].$$

Set

$$a = \epsilon/8, \qquad m_t = \left\lceil \frac{1}{2a^2} \log \frac{8B_t}{\delta} \right\rceil,$$

and choose

$$n_t = \left\lceil \frac{8}{p_{\min,t}} \left(m_t + \log \frac{8B_t}{\delta} \right) \right\rceil.$$

Because $\log B_t = o(d(t))$, $\log m_t = o(d(t))$, and $-\log p_{\min,t} = d(t) + o(d(t))$,

$$\log n_t = d(t) + o(d(t)).$$

From the n_t baseline samples, let N_j be the number whose score lies in J_j, and let $\bar{Y}_j$ be their average label when $N_j > 0$, with the arbitrary fallback $\bar{Y}_j = 1/2$ when $N_j = 0$. A multiplicative Chernoff bound and the definition of n_t give

$$P(N_j < m_t) \leq \frac{\delta}{4B_t}$$

for all large t. For every integer $k \geq m_t$, conditional on $N_j = k$ and after relabeling the k observations whose scores fall in J_j, their labels are iid bounded variables with mean

$$\mu_j = E_P[f(S) \mid S \in J_j].$$

Hoeffding's inequality therefore gives, for every $k \geq m_t$,

$$\begin{aligned} P(|\bar{Y}_j - \mu_j| > a \mid N_j = k) &\leq 2e^{-2ka^2} \\ &\leq 2e^{-2m_t a^2} \leq \frac{\delta}{4B_t}. \end{aligned}$$

Averaging this bound over the conditional law of N_j on $N_j \geq m_t$ yields

$$P(|\bar{Y}_j - \mu_j| > a \mid N_j \geq m_t) \leq \frac{\delta}{4B_t}.$$

A union bound shows that with probability at least $1 - \delta/2$, every bin is populated and

$$|\bar{Y}_j - \mu_j| \leq a \qquad \text{for all } j.$$

Define

$$\widehat{Q}_t f = \sum_{j=1}^{B_t} q_{j,t} \bar{Y}_j + \frac{1}{2} Q_t(I_t^c).$$

Within a bin, the oscillation of f is at most $Lh = \epsilon/8$. Hence

$$|E_P[f \mid J_j] - E_{Q_t}[f \mid J_j]| \leq \epsilon/8.$$

On the good event,

$$\begin{aligned} |\widehat{Q}_t f - E_{Q_t} f| &\leq \sum_j q_{j,t} |\bar{Y}_j - \mu_j| \\ &\quad + \sum_j q_{j,t} |\mu_j - E_{Q_t}[f \mid J_j]| \\ &\quad + \frac{1}{2} Q_t(I_t^c) \\ &\leq \epsilon/8 + \epsilon/8 + \epsilon/16 = 5\epsilon/16. \end{aligned}$$

Estimate the baseline mean by the global sample average

$$\widehat{P} f = \frac{1}{n_t} \sum_{i=1}^{n_t} Y_i.$$

Hoeffding gives, for all large t,

$$P(|\widehat{P} f - E_P f| > \epsilon/4) \leq 2e^{-n_t \epsilon^2/8} \leq \delta/2.$$

Therefore

$$\widehat{\theta}_t = \widehat{Q}_t f - \widehat{P} f$$

has error at most

$$5\epsilon/16 + \epsilon/4 = 9\epsilon/16 < \epsilon$$

with probability at least $1 - \delta$, uniformly over $f \in \mathcal{F}_L$. Thus

$$n^*_{\text{base}} \le n_t = \exp[d(t) + o(d(t))].$$

Combining the bounds proves the theorem.

□

Assumption 5.5 (Regular stretched-exponential tail).

Let P be a probability law on $\mathbb{R}$. Assume that, on some right-tail interval $[s_0, \infty)$, its restriction has a positive Lebesgue density p, and write $U(s) = -\log p(s)$ there. Assume, as $s \to \infty$:

1. *U is eventually C^2, strictly convex, and increasing;*

2. *$U(s) = s^\alpha/\alpha(1 + o(1))$ for some $\alpha > 1$;*

3. *there are constants $0 < c < C < \infty$ such that*

$$cs^{\alpha-2} \le U''(s) \le Cs^{\alpha-2}$$

eventually.

These are the regularity conditions stated in the supplementary material, written in a form sufficient for the estimates below.

Theorem 5.6 (Regular-tail extension).

Under Assumption 5.5, for $t > 0$ define the exponential tilt

$$\frac{dQ_t}{dP}(s) = e^{ts-\Lambda(t)}, \qquad \Lambda(t) = \log E_P e^{tS}.$$

The right-tail assumption makes this normalizer finite for every positive t (the contribution from $S < s_0$ is bounded by e^{ts_0}, while the stretched-exponential right tail dominates e^{ts}). Assume labels are conditionally Bernoulli with mean $f(S)$, or noiseless. Fix $L \in (0, \infty)$ and $\delta \in (0, 1/4)$, and let

$$\mathcal{F}_L = \{f : \mathbb{R} \to [0, 1] : f \text{ is } L\text{-Lipschitz}\}.$$

Let s_t be the eventual solution of $U'(s_t) = t$ and put $w_t = U''(s_t)^{-1/2}$. Then

$$s_t = t^{1/(\alpha-1)}(1 + o(1)), \qquad w_t \asymp s_t^{(2-\alpha)/2},$$

and

$$D(Q_t \| P) = U(s_t) + o(U(s_t)).$$

Fixed tilted quantiles are $s_t + O(w_t)$. Every interval contained in a fixed $O(w_t)$ neighborhood of s_t whose positive length ℓ_t satisfies $|\log \ell_t| = o(U(s_t))$ has baseline mass $\exp[-U(s_t) + o(U(s_t))]$, and

$$-\log P(S \ge s_t + O(w_t + 1)) = U(s_t) + o(U(s_t)).$$

Consequently, there exists $\epsilon_0 = \epsilon_0(L, \alpha) > 0$ such that, for every fixed $\epsilon \in (0, \epsilon_0)$, the proof of Theorem 5.4 applies to $\mathcal{F}_L$ and

$$\log n^*_{\mathrm{base}}(t,\epsilon,\delta;\mathcal{F}_L) = D(Q_t\|P) + o(D(Q_t\|P)).$$

Proof. **Step 1: derivative asymptotics.** Convexity and the tail asymptotic imply

$$U'(s) \sim s^{\alpha-1}.$$

Indeed, for each fixed $\eta \in (0,1)$, convexity gives

$$\frac{U(s)-U((1-\eta)s)}{\eta s} \le U'(s) \le \frac{U((1+\eta)s)-U(s)}{\eta s}.$$

After division by $s^{\alpha-1}$, first let $s \to \infty$ and then $\eta \to 0$. The limiting lower and upper secant coefficients both tend to one. Hence $U'(s_t) = t$ gives the displayed asymptotic for s_t, and the curvature assumption gives the one for w_t.

Step 2: normalization and moments. Choose s_0 large enough that the restriction of P to $[s_0,\infty)$ has the positive Lebesgue density $p = e^{-U}$ and all parts of Assumption 5.5 hold there. On that right-tail domain put

$$\Phi_t(s) = U(s) - ts.$$

For all large t, this is minimized at s_t. For $|u| \le s_t/2$, the integral remainder formula and the curvature assumption give

$$c_1\frac{u^2}{w_t^2} \le \Phi_t(s_t+u) - \Phi_t(s_t) \le C_1\frac{u^2}{w_t^2}.$$

At $|u| = s_t/2$, the exponent loss is of order s_t^α. More explicitly, convexity and the unique minimum at s_t give, for $k = 0,1,2$,

$$\int_{s_0}^{s_t/2} |s-s_t|^k \exp\{-[\Phi_t(s)-\Phi_t(s_t)]\}\,ds \le C_k s_t^{k+1} e^{-c_2 s_t^\alpha}.$$

For $s \ge 3s_t/2$, derivative asymptotics give $\Phi_t'(s) \ge c_3 t$, while the endpoint exponent loss is at least $c_4 s_t^\alpha$. Consequently,

$$\int_{3s_t/2}^{\infty} |s-s_t|^k \exp\{-[\Phi_t(s)-\Phi_t(s_t)]\}\,ds \le C_k e^{-c_4 s_t^\alpha}\left(\frac{s_t^k}{t} + \frac{1}{t^{k+1}}\right).$$

For $k = 0,1,2$, define the right-tail Laplace integrals

$$J^+_{k,t} = \int_{s_0}^{\infty} |s-s_t|^k \exp\{-[\Phi_t(s)-\Phi_t(s_t)]\}\,ds$$

and the modal-normalized left remainders

$$R^-_{k,t} = e^{-[ts_t - U(s_t)]} \int_{(-\infty,s_0)} |s-s_t|^k e^{ts}\,P(ds).$$

The central quadratic bounds and the preceding right-tail estimates give

$$J_{0,t}^{+} \asymp w_t, \qquad J_{k,t}^{+} = O(w_t^{k+1}), \quad k = 1, 2.$$

For the left remainders,

$$\int_{(-\infty, s_0)} |s - s_t|^k e^{ts}\, P(ds) \le \sup_{s<s_0} |s - s_t|^k e^{ts}.$$

Writing $s_0^{+} = \max s_0, 0$, an explicit bound is

$$\sup_{s<s_0} |s - s_t|^k e^{ts} \le C_{k,s_0}\big(1 + s_t^k + t^{-k}\big) e^{ts_0^{+}}.$$

Moreover, derivative asymptotics and $U(s_t) = s_t^{\alpha}/\alpha(1 + o(1))$ give

$$ts_t - U(s_t) = \left(1 - \frac{1}{\alpha} + o(1)\right) s_t^{\alpha}, \qquad t = o(s_t^{\alpha}).$$

Hence, for $k = 0, 1, 2$,

$$R_{k,t}^{-} = e^{-\Theta(s_t^{\alpha})}\, \mathrm{poly}(s_t, t^{-1}) = o(w_t^{k+1}).$$

The tilted normalizer and centered moment numerators therefore satisfy

$$E_P e^{tS} = e^{ts_t - U(s_t)}\big(J_{0,t}^{+} + R_{0,t}^{-}\big)$$

and, for $k = 1, 2$,

$$E_{Q_t}|S - s_t|^k = \frac{J_{k,t}^{+} + R_{k,t}^{-}}{J_{0,t}^{+} + R_{0,t}^{-}} = O(w_t^k).$$

Thus

$$E_{Q_t}|S - s_t| = O(w_t), \qquad E_{Q_t}(S - s_t)^2 = O(w_t^2),$$

and fixed quantiles lie at $s_t + O(w_t)$.

Step 3: log partition and KL. The normalization estimate gives

$$\begin{aligned} \Lambda(t) &= ts_t - U(s_t) + \log w_t + O(1) \\ &= ts_t - U(s_t) + o(U(s_t)). \end{aligned}$$

because $|\log w_t| = O(\log s_t)$. Also

$$tw_t \asymp s_t^{\alpha/2} = o(U(s_t)).$$

Therefore

$$D(Q_t \| P) = tE_{Q_t} S - \Lambda(t) = U(s_t) + o(U(s_t)).$$

Step 4: baseline survival. Convexity gives

$$U(s) \ge U(x) + U'(x)(s - x), \qquad s \ge x,$$

and hence

$$\int_x^\infty e^{-U(s)}ds \le \frac{e^{-U(x)}}{U'(x)}.$$

For the reverse logarithmic bound, integrate over $[x, x + 1/U'(x)]$. The curvature and derivative asymptotics imply that U' changes by $o(U'(x))$ on this interval, so $U(s) - U(x) = O(1)$ there. Thus

$$\int_x^\infty e^{-U(s)}ds \ge \frac{e^{-U(x)-O(1)}}{U'(x)}.$$

Since $\log U'(x) = o(U(x))$,

$$-\log P(S \ge x) = U(x) + o(U(x)).$$

Step 5: local baseline masses. If $s = s_t + O(w_t + 1)$, Taylor's theorem and the curvature bound give

$$\begin{aligned} U(s) - U(s_t) &= O(t(w_t + 1)) \\ &\quad + O(s_t^{\alpha-2}(w_t + 1)^2) \\ &= o(U(s_t)). \end{aligned}$$

The density is therefore $\exp[-U(s_t) + o(U(s_t))]$ uniformly on a fixed mode window. Multiplication by a positive interval length whose logarithm is $o(U(s_t))$ proves the interval-mass claim. The survival estimate proves the tail-start claim. The lower-ramp and central-bin arguments of Theorem 5.4 now apply unchanged.

□

Proposition 5.7 (Direct two-policy control).

Suppose labels lie in $[0, 1]$*, and let*

$$n_P, n_Q \in \{1, 2, \ldots\}.$$

Let $(Y_i^P)_{i=1}^{n_P}$ *be iid labels from the* P *experiment and* $(Y_j^Q)_{j=1}^{n_Q}$ *be iid labels from the* Q *experiment, with the two samples independent. Define*

$$\theta := E[Y_1^Q] - E[Y_1^P].$$

Let $\mathbb{P}$ *denote the product law of the two samples, together with any independent estimator randomization. The plug-in estimator*

$$\widehat{\theta} = \frac{1}{n_Q}\sum_{j=1}^{n_Q} Y_j^Q - \frac{1}{n_P}\sum_{i=1}^{n_P} Y_i^P$$

satisfies

$$\mathbb{P}(|\widehat{\theta} - \theta| > \epsilon) \le 2e^{-n_P\epsilon^2/2} + 2e^{-n_Q\epsilon^2/2}.$$

In the stable-label setting with common conditional mean f*, this estimand is* $E_Q f - E_P f$*. Thus* $n_P, n_Q = O(\epsilon^{-2}\log(1/\delta))$ *suffice, independently of the overlap rarity or optimization KL.*

Proof. If both sample means are within $\epsilon/2$ of their expectations, the difference is within ϵ. Apply Hoeffding to each mean and union bound.

□

I Complete Proofs: Active access and audit pressure

I.1 Formal active-audit model

Fix an event $A \subset \Omega$ with $0 < \mu(A) < 1$, and let

$$p = \mu(A), \qquad \eta = q(A),$$

and consider the two target worlds

$$m_0 \equiv 1, \qquad m_1 = 1 - b1_A, \qquad 0 < b \leq 1.$$

Throughout the active-access proofs, the query count satisfies

$$n \in \{1, 2, \ldots\}.$$

The two optimized means are

$$R_q(m_0) = 1, \qquad R_q(m_1) = 1 - b\eta,$$

so their separation is $b\eta$. Fix one label experiment throughout the active lower-law argument: either noiseless labels $Y = m_j(O)$, or conditionally independent Bernoulli labels with mean $m_j(O)$. More generally, the same argument applies whenever the full conditional label laws agree in the two worlds on A^c.

Observable-transcript convention

Throughout the active-access proofs and every upper benchmark that cites this convention, a randomized query rule is specified intrinsically as a sequence of stochastic kernels on the **observable** history. Fresh random bits and intermediate candidates used only to realize a kernel, then discarded and unavailable to the estimator and to later query rules, are marginalized out. The transcript contains the retained queried objects or scores, their labels, and every random variable, candidate, or internal state that is retained, explicitly revealed, logged, or otherwise available to the estimator or future decisions. Thus public, retained, or downstream-used state is included; only genuinely discarded sampler randomization and candidates are not.

This convention prevents the KL cost from depending on an arbitrary implementation of the same stochastic kernel. If $\widetilde{P}_j, \widetilde{M}_j$ are augmented laws that additionally expose private state and P_j, M_j are their observable projections, data processing gives

$$D(P_j \| M_j) \leq D(\widetilde{P}_j \| \widetilde{M}_j).$$

Consequently, every lower bound proved for observable-transcript KL also applies when a stronger exposed-state KL cap is imposed. Upper bounds, however, must use the transcript actually charged; Theorem 10.1 records the exact change when its latent swap index is exposed.

A common nonanticipating strategy selects the ith query through a stochastic kernel that depends only on the previously observed queries and labels, not on the unknown world except through those observations. Let P_j be the resulting **observable query-label transcript law** in world j. Let M_j be the reference observable transcript law obtained from n iid baseline queries from μ, with labels generated by the same fixed experiment in world j. Any public, logged, or retained auxiliary state has the same declared reference law and is included in both transcripts; only fresh kernel-realization randomness that is discarded and unavailable downstream is marginalized.

Let

$$H = \{\text{at least one query lies in } A\}, \qquad p_n = 1 - (1-p)^n.$$

Lemma 6.1 (First-hit coupling).

For a common nonanticipating strategy,

$$P_0(H) = P_1(H) =: a,$$

and the restrictions of P_0 and P_1 to H^c agree.

Proof. Before the first query in A, the full conditional label laws agree in the two worlds. For the proof, couple draws from the common conditional kernels using a common auxiliary randomization; this coupling variable is not part of the charged transcript unless it is explicitly exposed. As long as no hit has occurred, the observable histories are identical, so the next conditional query law is identical. Therefore the first-hit time has the same law in both worlds, and on no hit the observable transcripts are identical.

□

I.2 Exact risk under a fixed audit sampler

Theorem 7.1 (Fixed-sampler exact minimax absolute risk).

Let $O_1, ..., O_n$ be iid from a fixed audit distribution ν, with noiseless labels from one of the two worlds above. Then

$$\inf_{\widehat{R}} \sup_{j=0,1} E_j|\widehat{R} - R_q(m_j)| = \frac{b\eta}{2}(1-\nu(A))^n.$$

For the shift target $\Delta_q(m) = E_q m - E_\mu m$, the exact risk is

$$\frac{b|\eta - p|}{2}(1-\nu(A))^n.$$

Proof. Let

$$E_n = \{O_1, ..., O_n \notin A\}.$$

Its probability is $(1-\nu(A))^n$. On E_n, all observed labels are 1 in both worlds, so Lemma 3.5 gives the lower bound with target separation $b\eta$. If a hit occurs, the label is 1 in world 0 and $1-b$ in world 1, so the world is identified exactly. The estimator that returns the correct target after a hit and the midpoint

$$1 - \frac{b\eta}{2}$$

after no hit has error $b\eta/2$ exactly on E_n in both worlds, attaining the bound. The shift-target statement is identical with separation $b|\eta - p|$.

□

Corollary 7.2 (Density-ratio-capped sampler).

If

$$\frac{d\nu}{d\mu} \le R \quad \mu\text{-a.s.},$$

then

$$\nu(A) \le \min\{1, Rp\}.$$

If $Rp < 1$, every such sampler has risk at least

$$\frac{b\eta}{2}(1 - Rp)^n.$$

To make expected absolute risk at most $\epsilon < b\eta/2$, it is necessary that

$$n \ge \frac{\log(b\eta/(2\epsilon))}{-\log(1 - Rp)}.$$

If $Rp \le 1/2$, then

$$n \ge \frac{1}{2Rp} \log \frac{b\eta}{2\epsilon}.$$

Proof. Integrating the density-ratio bound over A gives $\nu(A) \le Rp$, capped at one. Apply Theorem 7.1. Solving the resulting exponential inequality gives the exact displayed bound. For $x \le 1/2$, $-\log(1 - x) \le 2x$.

□

Corollary 7.3 (Per-draw KL cap).

If $D(\nu||\mu) \le \kappa$ and $a = \nu(A)$, then

$$\kappa \ge \mathrm{kl}(a\|p) \ge a \log \frac{1}{p} - h_2(a).$$

Thus, for every fixed $a_0 \in (0, 1)$, achieving $\nu(A) \ge a_0$ as $p \to 0$ requires

$$\kappa \ge a_0 \log \frac{1}{p} - O_{a_0}(1).$$

Proof. Use Lemma 3.3 with the map $\mathbf{1}_A$, then Lemma 3.1.

□

I.3 TRANSCRIPT-KL ACTIVE LAW

Theorem 8.1 (Hit probability required for high-confidence estimation).

Suppose an estimator satisfies

$$P_j(|\widehat{R} - R_q(m_j)| \le \epsilon) \ge 1 - \delta, \qquad j = 0, 1,$$

where $\epsilon < b\eta/2$. Then the common hit probability from Lemma 6.1 satisfies

$$a = P_j(H) \ge 1 - 2\delta.$$

The same conclusion holds for the shift target $\Delta_q(m_j)$ when

$$\epsilon < \frac{b}{2}|\eta - p|.$$

Proof. For the optimized mean, the target separation is $b\eta > 2\epsilon$; for the shift, it is $b|\eta - p| > 2\epsilon$. On H^c, the transcript laws agree by Lemma 6.1. Apply Lemma 3.4 with $E = H^c$, obtaining $P(H^c) \le 2\delta$.

□

Theorem 8.2 (Active access law).

Assume

$$D(P_j \| M_j) \le K, \qquad j = 0, 1,$$

and either the optimized-mean or shift-target high-confidence condition of Theorem 8.1, with $\delta \in (0, 1/2)$. Then

$$\frac{K}{1 - 2\delta} + \log n \ge \log \frac{1}{p} - C_\delta,$$

where one may take

$$C_\delta = \frac{\log 2}{1 - 2\delta}.$$

In particular, the theorem holds even if the auditor knows A.

Proof. Under the reference law M_j, the hit probability is $p_n = 1 - (1 - p)^n$. Under the active law it is the common value a. By Lemma 3.3,

$$K \ge \mathrm{kl}(a \| p_n).$$

Theorem 8.1 gives $a \ge a_0 := 1 - 2\delta$. Lemma 3.1 and $h_2(a) \le log2$ yield

$$K \ge a \log \frac{1}{p_n} - h_2(a) \ge a_0 \log \frac{1}{p_n} - \log 2.$$

Since $p_n \le np$,

$$\log \frac{1}{p_n} \geq \log \frac{1}{p} - \log n.$$

Therefore

$$K \geq a_0 \left(\log \frac{1}{p} - \log n \right) - \log 2,$$

which rearranges to the claim. Knowledge of A was never restricted.

□

Theorem 8.3 (Exact total-KL minimax risk in the common-query oracle model).

Fix $K \in [0, \infty)$. Consider the abstract class of query laws P^X on Ω^n that are common to both worlds and satisfy

$$D(P^X \| \mu^n) \leq K.$$

Labels are then attached noiselessly according to the world. Define

$$a_K = \sup\{a \in [0,1] : \mathrm{kl}(a \| p_n) \leq K\}.$$

Then

$$\inf_{P^X :\, D(P^X \| \mu^n) \leq K} \inf_{\widehat{R}} \sup_{j=0,1} E_j |\widehat{R} - R_q(m_j)| = \frac{b\eta}{2}(1 - a_K).$$

For the shift target, replace $b\eta$ by $b|\eta - p|$.

Proof: lower bound. For any admissible P^X, put $a = P^X(H)$. Data processing gives

$$\mathrm{kl}(a \| p_n) \leq K,$$

so $a \leq a_K$ by definition of the upper feasible endpoint. On H^c, labels are identical in the two worlds. Lemma 3.5 gives risk at least

$$\frac{b\eta}{2}(1 - a) \geq \frac{b\eta}{2}(1 - a_K).$$

□

Proof: attainment. Define the event-tilted query law

$$\frac{dP^{X,*}}{d\mu^n} = \frac{a_K}{p_n} 1_H + \frac{1 - a_K}{1 - p_n} 1_{H^c},$$

with the evident conventions at boundary cases. It is a probability law, its hit probability is a_K, and

$$D(P^{X,*} \| \mu^n) = \mathrm{kl}(a_K \| p_n) \leq K.$$

The law is common to both worlds because its density depends only on the query event H. The exact risk statement is a joint-law oracle statement on the marginal observable query history. If Ω is standard Borel, regular conditional distributions disintegrate this joint law into sequential kernels, giving a nonanticipating implementation; fresh kernel-realization randomness is discarded and

marginalized, and no realized future information is supplied to a kernel. Exposing a particular implementation seed would define a different, generally larger, KL cost.

On H, a noiseless label in A identifies the world exactly. On H^c, output the midpoint. The risk is exactly $(b\eta/2)(1 - a_K)$ in each world.

□

Corollary 8.4 (Budget-sample conservation laws; corrected exact threshold).

For expected absolute risk at most $\epsilon < b\eta/2$, it is necessary and sufficient in the oracle model that

$$a_K \geq \rho_\epsilon := 1 - \frac{2\epsilon}{b\eta}.$$

Because $a_K \geq p_n$ even at $K = 0$, the exact minimum KL budget at fixed n is

$$K_{\min}(n, \epsilon) = \begin{cases} 0, & \rho_\epsilon \leq p_n, \\ \mathrm{kl}(\rho_\epsilon \| p_n), & \rho_\epsilon > p_n. \end{cases}$$

Equivalently,

$$K_{\min}(n, \epsilon) = \mathrm{kl}(\max\{\rho_\epsilon, p_n\} \| p_n).$$

Every successful oracle therefore obeys the globally valid positive-part bound

$$K \geq \left[\rho_\epsilon \log \frac{1}{p_n} - h_2(\rho_\epsilon) \right]_+,$$
$$K \geq \left[\rho_\epsilon \left(\log \frac{1}{p} - \log n \right) - h_2(\rho_\epsilon) \right]_+.$$

In particular,

$$\frac{K}{\rho_\epsilon} + \log n \geq \log \frac{1}{p} - \frac{h_2(\rho_\epsilon)}{\rho_\epsilon}.$$

For a high-probability guarantee with failure probability δ, Theorem 8.2 gives the analogous statement with the coefficient $1 - 2\delta$. Thus, up to constants fixed by the desired risk or confidence,

$$K + \log n \gtrsim \log \frac{1}{\mu(A)}.$$

Proof. The exact risk formula gives $a_K \geq \rho_\epsilon$. The feasible set of $a \to kl(a||p_n)$ contains p_n, and the binary KL is strictly increasing on $[p_n, 1]$. Hence zero budget suffices when $\rho_\epsilon \leq p_n$, while otherwise the exact threshold is $kl(\rho_\epsilon||p_n)$.

When $\rho_\epsilon > p_n$, Lemma 3.1 gives

$$K \geq \rho_\epsilon \log(1/p_n) - h_2(\rho_\epsilon);$$

when $\rho_\epsilon \leq p_n$, the same expression is nonpositive because

$$\rho_\epsilon \log(1/p_n) - h_2(\rho_\epsilon) \leq \rho_\epsilon \log(1/\rho_\epsilon) - h_2(\rho_\epsilon) \leq 0.$$

This proves the positive-part form. Finally use $p_n \le np$. If the bracket is positive, divide the resulting inequality by ρ_ϵ; if it is nonpositive, its nonpositivity itself yields the same displayed conservation law.

□

Corollary 8.5 (Gaussian audit-pressure phase).

Let $t > 0$, $\mu = N(0,1)$, $q_t = N(t,1)$, and $d = t^2/2$. For

$$A_t = \{S \ge t\},$$

as $t \to \infty$ we have

$$q_t(A_t) = \frac{1}{2}, \qquad p_t = \mu(A_t) = \bar{\Phi}(t) \sim \frac{e^{-d}}{\sqrt{4\pi d}},$$

and therefore

$$\log \frac{1}{p_t} = d + \frac{1}{2}\log d + O(1).$$

Any one-draw audit sampler that hits A_t with fixed positive probability has $D(\nu\|\mu) = \Omega(d)$. More generally, for estimating the optimized mean $R_{q_t}(m)$ in the two-world pair of Appendix I.1, fix $0 < \epsilon < b/4$ and $0 < \delta < 1/2$. Then every high-confidence n-query auditor with observable-transcript KL K obeys

$$\frac{K}{1-2\delta} + \log n \ge d + \frac{1}{2}\log d - O_\delta(1).$$

The same conclusion holds for estimating the shift $\Delta_{q_t}(m)$ for every fixed $0 < \epsilon < b/4$ and all sufficiently large t, because its target separation is

$$b\left(\frac{1}{2} - p_t\right),$$

which eventually exceeds 2ϵ.

Proof. The tilt identity and $q_t(A_t) = 1/2$ are immediate. Mills' ratio gives

$$\bar{\Phi}(t) = \frac{\phi(t)}{t}(1 + o(1)) = \frac{e^{-t^2/2}}{t\sqrt{2\pi}}(1 + o(1)).$$

Substitute $t = \sqrt{2d}$. The one-draw statement follows from Corollary 7.3. For $R_{q_t}(m)$, apply Theorem 8.2 with $\eta = 1/2$ and $\epsilon < b/4$. For $\Delta_{q_t}(m)$, Theorem 8.1 requires

$$\epsilon < \frac{b}{2}\left(\frac{1}{2} - p_t\right),$$

which holds for all sufficiently large t for every fixed $\epsilon < b/4$; Theorem 8.2 then gives the same bound.

□

Proposition 8.6 (Adaptive per-round KL caps).

In the common nonanticipating two-world model of Appendix I.1, let $\kappa_i \geq 0$ be deterministic caps. Suppose that, before the first hit, the conditional query law at round i is $\nu_i(\cdot|history)$ and satisfies almost surely

$$D(\nu_i(\cdot|history)\|\mu) \leq \kappa_i.$$

For $\kappa \in [0, \infty)$, let

$$r_\kappa(p) = \sup\{r \in [0,1] : \mathrm{kl}(r\|p) \leq \kappa\}.$$

Then the no-hit probability is at least

$$\prod_{i=1}^{n}(1 - r_{\kappa_i}(p)),$$

and the two-world expected minimax risk is at least

$$\frac{b\eta}{2}\prod_{i=1}^{n}(1 - r_{\kappa_i}(p)).$$

Proof. Conditional on any no-hit history, data processing to $\mathbf{1}_A$ gives

$$\nu_i(A|history) \leq r_{\kappa_i}(p).$$

Iterating conditional no-hit probabilities proves the product bound. Apply Lemma 3.5 on the no-hit event.

□

I.4 Smooth strict separation beyond the indicator pair

Theorem 9.1 (Gaussian Lipschitz separation).

Let

$$P = N(0,1), \qquad Q_t = N(t,1), \qquad d = t^2/2.$$

Fix $L \in (0, \infty)$ and $\delta \in (0, 1/4)$, and let

$$\mathcal{F}_L = \{f : \mathbb{R} \to [0,1] : f \text{ is } L\text{-Lipschitz}\}.$$

The target is

$$\theta_t(f) = E_{Q_t} f - E_P f.$$

Labels are either noiseless with value $f(X)$ or conditionally Bernoulli with mean $f(X)$. There exists $\epsilon_0(L) > 0$ such that for every fixed $\epsilon \in (0, \epsilon_0)$:

***1.** Direct labeled sampling from both P and Q_t needs only*

$$O\left(\epsilon^{-2}\log\frac{1}{\delta}\right)$$

samples from each law, uniformly over $\mathcal{F}_L$.

2. *If a common nonanticipating active auditor makes n score queries and, in every world f, its observable query-label transcript law $\mathbb{P}_f^{\mathcal{A}}$ and the corresponding iid-baseline observable transcript law $\mathbb{M}_f^{(n)}$ satisfy*

$$D\Big(\mathbb{P}_f^{\mathcal{A}}\Big\|\mathbb{M}_f^{(n)}\Big) \le K,$$

under the observable-transcript convention of Appendix I.1, while estimating $\theta_t(f)$ to error ϵ with probability at least $1-\delta$ uniformly over $\mathcal{F}_L$, then for all sufficiently large t,

$$\frac{K}{1-2\delta} + \log n \ge d + \frac{1}{2}\log d - O_{\delta,L}(1).$$

In particular, $K = o(d)$ implies $n \ge \exp(d - o(d))$.

Proof. The direct upper bound is Proposition 5.7.

For the lower bound, let

$$A_t = \{x > t\}, \qquad \phi_t(x) = \min\{1, L(x-t)_+\},$$

and define

$$f_0(x) = \frac{1}{2}, \qquad f_1(x) = \frac{1}{2} - \frac{1}{2}\phi_t(x).$$

Both functions lie in $\mathcal{F}_L$: the Lipschitz constant of f_1 is at most $L/2$, and its range is contained in $[0, 1/2]$. Under Q_t, write $X = t + Z$, $Z \sim N(0,1)$. Then

$$E_{Q_t}\phi_t(X) = E\min\{1, LZ_+\} =: c_L > 0.$$

Under P,

$$0 \le E_P\phi_t(X) \le P(X > t) = \bar{\Phi}(t) = o(1).$$

Therefore

$$|\theta_t(f_1) - \theta_t(f_0)| = \frac{1}{2}(c_L - o(1)).$$

Take $\epsilon_0(L) = c_L/16$. For fixed $\epsilon < \epsilon_0$, the target separation exceeds 2ϵ (indeed 4ϵ) for all sufficiently large t.

Before the first query in A_t, the two label laws coincide because $f_0 = f_1 = 1/2$ on $(-\infty, t]$. The first-hit coupling therefore applies exactly as in Lemma 6.1, including for Bernoulli labels with these means. Uniform success forces a common hit probability $a \ge 1 - 2\delta$ by Lemma 3.4. By data processing from the observable query-label transcript to its query-score marginal, that marginal has KL at most K from P^n. Put $p_{n,t} = 1 - (1-p_t)^n$, where $p_t = \bar{\Phi}(t)$. Data processing to the hit indicator and Lemma 3.1 give

$$\begin{aligned} K &\ge \mathrm{kl}(a\|p_{n,t}) \\ &\ge a\log\frac{1}{p_{n,t}} - h_2(a) \\ &\ge (1-2\delta)\log\frac{1}{p_{n,t}} - \log 2. \end{aligned}$$

Using $p_{n,t} \le np_t$ exactly as in Theorem 8.2 yields

$$\frac{K}{1-2\delta} + \log n \ge \log \frac{1}{p_t} - O_\delta(1).$$

Mills' ratio gives

$$\log \frac{1}{p_t} = \frac{t^2}{2} + \log t + O(1) = d + \frac{1}{2}\log d + O(1).$$

This proves the claim.

□

I.5 SIZE-BIASED LIKELIHOOD-RATIO BLOCK ORACLE

Theorem 10.1 (Single-swap block law, exact KL sandwich, and unbiased shift estimator). *Let $q \ll \mu$,*

$$W = \frac{dq}{d\mu}, \qquad d = D(q\|\mu) = E_q \log W < \infty,$$

and let ℓ be an integer with $\ell \ge 2$. On a block $(O_1, ..., O_\ell)$ define

$$\bar{W}_\ell = \frac{1}{\ell}\sum_{i=1}^{\ell} W(O_i),$$
$$\frac{dP_\ell^*}{d\mu^\ell} = \bar{W}_\ell, \qquad k_\ell = D(P_\ell^*\|\mu^\ell).$$

Then:

1. *P_ℓ^* is a probability law. Equivalently, choose I uniformly from $1, ..., \ell$, draw $O_I \sim q$, and draw every other coordinate independently from μ.*

2. *Its KL cost obeys*

$$k_\ell \le E_q \log\left(1 + \frac{W-1}{\ell}\right) \le E_q \log\left(1 + \frac{W}{\ell}\right).$$

3. *It also obeys the exact lower bound*

$$k_\ell \ge [d - \log \ell]_+.$$

4. *If $d \to \infty$ and*

$$E_q|\log W - d| = o(d),$$

then uniformly over block-size sequences,

$$k_\ell = [d - \log \ell]_+ + o(d).$$

5. *Suppose labels are generated by a fixed conditional kernel shared by the oracle experiment and its baseline reference. Conditional on the queried block,*

$$Y_i \in [0,1], \qquad E[Y_i \mid O_1, \ldots, O_\ell] = m(O_i).$$

On the event $\sum_i W_i > 0$, define

$$\widehat{q}_\ell = \frac{\sum_i W_i Y_i}{\sum_i W_i},$$

and set $\widehat{q}_\ell = 0$ otherwise. Then

$$E_{P_\ell^*} \widehat{q}_\ell = E_q m.$$

6. *With*

$$\bar{Y}_\ell = \frac{1}{\ell} \sum_i Y_i, \qquad \widehat{\Delta}_\ell = \frac{\ell}{\ell - 1} (\widehat{q}_\ell - \bar{Y}_\ell),$$

one has

$$E_{P_\ell^*} \widehat{\Delta}_\ell = \Delta_q(m), \qquad |\widehat{\Delta}_\ell| \le 2.$$

7. *The private-index and exposed-index costs differ exactly. Define the augmented active and baseline laws by*

$$\widetilde{P}_\ell(i, do_{1:\ell}) = \frac{1}{\ell} q(do_i) \prod_{j \neq i} \mu(do_j),$$

$$\widetilde{M}_\ell(i, do_{1:\ell}) = \frac{1}{\ell} \prod_{j=1}^{\ell} \mu(do_j).$$

Then their marginals on $O_{1:\ell}$ are P_ℓ^* and μ^ℓ, while

$$\begin{aligned} D(\widetilde{P}_\ell \| \widetilde{M}_\ell) &= d, \\ d - k_\ell &= E_{P_\ell^*} D\Big(\widetilde{P}_\ell(I \in \cdot \mid O_{1:\ell}) \Big\| \mathrm{Unif}\{1, \dots, \ell\} \Big) \\ &= I_{\widetilde{P}_\ell}(I; O_{1:\ell}). \end{aligned}$$

Thus the refined leading-order discount encoded by $[d - \log \ell]_+$ requires the swap index to remain private and marginalized.

Consequently, for every $\epsilon > 0$ and $\delta \in (0, 1)$, averaging any integer number of blocks B satisfying

$$B \ge \left\lceil \frac{8}{\epsilon^2} \log \frac{2}{\delta} \right\rceil$$

*estimates $\Delta_q(m)$ to error ϵ with probability at least $1 - \delta$. The total labels and **observable-transcript** KL cost are*

$$N = B\ell, \qquad K_{\mathrm{obs}} = Bk_\ell.$$

Here K_{obs} is the KL of the observable query-label transcript relative to B independent baseline blocks. The two experiments use the same conditional label kernel, so relative entropy reduces by the chain rule to the marginal query-law KL. The latent indices are sampled privately and discarded; if they are logged, the augmented cost is $K_{\mathrm{aug}} = Bd$ by item 7.

Under the concentration assumption, every sequence with $\log \ell \le d + o(d)$ satisfies

$$\frac{K_{\mathrm{obs}}}{B} + \log \frac{N}{B} = d + o(d).$$

If K denotes an available observable-transcript budget cap, the construction is feasible whenever $K \geq K_{\mathrm{obs}}$. The asymptotic equality concerns the incurred observable KL K_{obs}, not an arbitrarily reported cap or an index-exposed implementation.

Proof. Since $E_\mu W = 1$,

$$E_{\mu^\ell} \bar{W}_\ell = 1,$$

so the density defines a probability law. The single-swap mixture has exactly the same density relative to μ^ℓ.

For the upper bound, condition on the swapped coordinate, say $O_1 \sim q$, and apply Jensen to the remaining baseline coordinates:

$$\begin{aligned} k_\ell &= E_{q \otimes \mu^{\ell-1}} \log \left(\frac{W(O_1) + \sum_{i=2}^{\ell} W(O_i)}{\ell} \right) \\ &\leq E_q \log \left(\frac{W + \ell - 1}{\ell} \right) \\ &= E_q \log \left(1 + \frac{W - 1}{\ell} \right) \\ &\leq E_q \log \left(1 + \frac{W}{\ell} \right). \end{aligned}$$

For the lower bound, retain the latent swapped index I. Pointwise under the augmented mixture,

$$\bar{W}_\ell \geq \frac{W(O_I)}{\ell}.$$

Therefore

$$k_\ell = E \log \bar{W}_\ell \geq E_q \log W - \log \ell = d - \log \ell.$$

Since KL is nonnegative, $k_\ell \geq [d - \log \ell]_+$.

For the matching upper asymptotic, put $Z = \log W$. The inequalities

$$\log(1 + e^{Z - \log \ell}) \leq [Z - \log \ell]_+ + \log 2$$

and

$$[Z - \log \ell]_+ \leq [d - \log \ell]_+ + |Z - d|$$

give

$$k_\ell \leq [d - \log \ell]_+ + E_q |Z - d| + \log 2.$$

Together with the exact lower bound, this proves item 4.

For item 7, the augmented likelihood ratio is

$$\frac{d\widetilde{P}_\ell}{d\widetilde{M}_\ell}(I, O_{1:\ell}) = W(O_I),$$

so

$$D(\widetilde{P}_\ell \| \widetilde{M}_\ell) = E_{\widetilde{P}_\ell} \log W(O_I) = E_q \log W = d.$$

Projecting away I gives the marginals P_ℓ^* and μ^ℓ. On $\sum_j W(O_j) > 0$, the posterior is explicitly

$$\widetilde{P}_\ell(I = i \mid O_{1:\ell}) = \frac{W(O_i)}{\sum_{j=1}^{\ell} W(O_j)}.$$

The KL chain rule for the projection therefore gives

$$d = k_\ell + E_{P_\ell^*} D\Big(\widetilde{P}_\ell(I \in \cdot \mid O_{1:\ell}) \Big\| \mathrm{Unif}\{1, \dots, \ell\}\Big).$$

Because I is uniform under $\widetilde{P}_\ell$, the final conditional KL is exactly $I_{\widetilde{P}_\ell}(I; O_{1:\ell})$, and is at most $\log \ell$. Attaching labels through the same conditional kernel does not change either KL value.

The event $\sum_i W_i = 0$ has zero P_ℓ^*-probability. Conditioning on the queried block and using the stable label kernel,

$$\begin{aligned} E_{P_\ell^*} \widehat{q}_\ell &= E_{\mu^\ell}\left[\bar{W}_\ell \frac{\sum_i W_i m(O_i)}{\sum_i W_i}\right] \\ &= E_{\mu^\ell}\left[\frac{1}{\ell}\sum_i W_i m(O_i)\right] = E_q m. \end{aligned}$$

Under the single-swap representation,

$$E_{P_\ell^*} \bar{Y}_\ell = \frac{1}{\ell} E_q m + \frac{\ell - 1}{\ell} E_\mu m.$$

The displayed shift estimator is therefore unbiased. Both component averages lie in $[0, 1]$, so its absolute value is at most $\ell/(\ell - 1) \le 2$.

Repeat the fixed query-label kernel independently across blocks. Hoeffding for the resulting independent variables in $[-2, 2]$ yields

$$P\left(\left|B^{-1}\sum_{b=1}^{B} \widehat{\Delta}_\ell^{(b)} - \Delta_q(m)\right| > \epsilon\right) \le 2e^{-B\epsilon^2/8}.$$

Finally, if $\log \ell \le d + o(d)$, then

$$[d - \log \ell]_+ + \log \ell = d + o(d),$$

and item 4 gives the observable-cost frontier.

□

I.6 EXPONENTIAL-POWER ACTIVE PHASE WITH AN ORACLE UPPER BENCHMARK

Theorem 11.1 (Exponential-power observable-KL active phase; oracle upper and Lipschitz minimax lower).

In the one-sided exponential-power score model of Appendix H.2, take

$$\Omega = [0, \infty), \qquad \mu = P, \qquad q = Q_t,$$

and fix

$$L \in (0, \infty), \qquad 0 < \epsilon < \frac{1}{4}, \qquad 0 < \delta < \frac{1}{2}.$$

Fix once and for all one of the following observation experiments throughout the theorem:

1. *noiseless labels $Y = m(S)$; or*

2. *conditionally independent Bernoulli labels $Y|S = s \sim Bernoulli(m(s))$.*

All observable query-label transcript laws, iid baseline references, success probabilities, and KL costs below refer to this same fixed experiment and to the convention in Appendix I.1: discarded sampler randomness is marginalized and public, logged, or retained state is included. The auditor in the lower bound is a common nonanticipating strategy whose observable-transcript KL from iid baseline querying is at most K in each world.

Uniform oracle upper bound. *For every integer $\ell \geq 2$, the same block law and estimator work for every target mean $m : [0, \infty) \to [0, 1]$, with*

$$B = O\left(\epsilon^{-2} \log \frac{1}{\delta}\right), \qquad N = B\ell,$$

and observable-transcript KL cost

$$K_{\text{obs}} = B\left([d(t) - \log \ell]_+ + o(d(t))\right).$$

Whenever $\log \ell \leq d(t) + o(d(t))$,

$$\frac{K_{\text{obs}}}{B} + \log \frac{N}{B} = d(t) + o(d(t)).$$

The same upper construction extends to a standard-Borel episode space whenever $dq_t/d\mu = e^{tS - \Lambda(t)}$, the induced score marginal is the exponential-power tilt above, and the oracle and its baseline reference use the same fixed conditional label kernel.

Lipschitz minimax lower bound. *Let*

$$\mathcal{F}_L = \{f : [0, \infty) \to [0, 1] : f \text{ is } L\text{-Lipschitz}\}.$$

Every auditor that estimates

$$\theta_t(f) = E_{Q_t} f - E_P f$$

to error ϵ with probability at least $1 - \delta$ uniformly over $\mathcal{F}_L$, using N queries and observable-transcript KL at most K, satisfies

$$\frac{K}{1 - 2\delta} + \log N \geq d(t) - o(d(t)).$$

For a fixed-query common nonanticipating auditor $\mathcal{A}$, let $N(\mathcal{A})$ be its total number of queries and define its transcript cost in the selected experiment by

$$K(\mathcal{A}) = \sup_{f \in \mathcal{F}_L} D\Big(\mathbb{P}_f^{\mathcal{A}} \Big\| \mathbb{M}_f^{N(\mathcal{A})}\Big),$$

where $\mathbb{M}_f^{N(\mathcal{A})}$ is the iid-baseline observable query-label transcript law with the same label experiment. Discarded query-generation randomness is marginalized, while any public, logged, or retained state is included. Consequently, in this fixed nontrivial-error regime,

$$\inf_{\substack{\mathcal{A} \text{ successful} \\ \text{in the fixed experiment}}} \big\{K(\mathcal{A}) + \log N(\mathcal{A})\big\} = \Theta_{\epsilon,\delta,\alpha,L}(d(t)).$$

where success means uniform ϵ-accuracy with probability at least $1-\delta$ over $\mathcal{F}_L$ in the one observation experiment selected above. The infimum is over fixed-query auditors in the observable-transcript-KL oracle model just defined. This is a phase-level statement about the least combined resource among successful auditors, not an assertion that every admissible resource pair lies on an equality curve. No matching leading constant for the full Lipschitz class is claimed.

Proof. Lemma 5.2 gives

$$E_{Q_t}\left|\log \frac{dQ_t}{dP} - d(t)\right| = o(d(t)).$$

The uniform upper bound follows from Theorem 10.1; its concentration bound is independent of the particular bounded target mean.

For the lower bound, let a_t be an upper-half quantile of Q_t, put

$$u_t = a_t - \frac{1}{L},$$

and define the L-Lipschitz ramp

$$\psi_t(s) = \begin{cases} 0, & s \le u_t, \\ L(s-u_t), & u_t < s < a_t, \\ 1, & s \ge a_t. \end{cases}$$

Take

$$f_0 \equiv 1, \qquad f_1 = 1 - \psi_t.$$

Both functions lie in $\mathcal{F}_L$ and agree on A_t^c, where

$$A_t = \{s > u_t\}.$$

Their shift-target separation satisfies

$$\begin{aligned} |\theta_t(f_1) - \theta_t(f_0)| &= E_{Q_t}\psi_t - E_P\psi_t \\ &\ge \frac{1}{2} - P(A_t) = \frac{1}{2} - o(1). \end{aligned}$$

Lemma 5.3 gives

$$\log \frac{1}{P(A_t)} = d(t) + o(d(t)).$$

Thus the separation is larger than 2ϵ for all sufficiently large t. Before the first query in A_t, the two complete conditional label laws agree, so the first-hit coupling and the proof of Theorem 8.2 apply verbatim. With $n = N$,

$$\frac{K}{1-2\delta} + \log N \geq \log \frac{1}{P(A_t)} - O_\delta(1) = d(t) - o(d(t)).$$

The upper bound gives $K_{\mathrm{obs}} + \log N = O(d(t))$ (for example with fixed ℓ), while the lower bound gives the reverse fixed-constant order. If the swap index is exposed, the refined upper curve changes, but the coarse $O(d(t))$ upper bound remains.

□

Remark 11.1.

The oracle upper bound is uniform over bounded target means because the query law and Hoeffding bound do not depend on m. The lower bound is already valid on the smaller smooth class $\mathcal{F}_L$; an indicator-only hard pair is not needed.

I.7 Multi-stage path-KL composition

Theorem 12.1 (Composition through the final density ratio).

Let the trajectory spaces and the final observable space be standard Borel. Let P_0 and P_R be baseline and optimized laws on a finite-stage trajectory $Z_{1:R}$, with

$$P_R \ll P_0, \qquad D(P_R \| P_0) < \infty.$$

Let the final observable be $O_R = T(Z_{1:R})$, and let

$$\mu = \mathrm{Law}_{P_0}(O_R), \qquad q = \mathrm{Law}_{P_R}(O_R).$$

Then

$$D(q\|\mu) \leq D(P_R\|P_0).$$

For $j \in \{0, R\}$, write $P_j^{<r} = \mathrm{Law}_{P_j}(Z_{<r})$ and let

$$P_{j,r}(\,\cdot \mid z_{<r})$$

be a regular conditional law of Z_r given $Z_{<r}$ under P_j. Assume, for every stage r,

$$P_{R,r}(\,\cdot \mid z_{<r}) \ll P_{0,r}(\,\cdot \mid z_{<r})$$

for $P_R^{<r}$-almost every history. Choose the regular conditional laws as Borel kernels. On the exceptional $P_R^{<r}$-null set, redefine $P_{R,r}(\,\cdot \mid z_{<r})$ to equal $P_{0,r}(\,\cdot \mid z_{<r})$. The measurable Radon–Nikodym theorem for kernels on standard-Borel spaces then gives a jointly measurable version

$$(z_{<r}, z_r) \longmapsto L_r(z_r \mid z_{<r})$$

such that, for $P_R^{<r}$-almost every $z_{<r}$ and every Borel set B,

$$P_{R,r}(B \mid z_{<r}) = \int_B L_r(z_r \mid z_{<r})\, P_{0,r}(dz_r \mid z_{<r}).$$

Equivalently, on those histories,

$$L_r(z_r \mid z_{<r}) = \frac{dP_{R,r}(\cdot \mid z_{<r})}{dP_{0,r}(\cdot \mid z_{<r})}(z_r).$$

Define the stage conditional-KL increment, initially as an extended nonnegative number, by

$$\begin{aligned}\kappa_r &:= E_{Z_{<r}\sim P_R^{<r}} D(P_{R,r}(\cdot \mid Z_{<r}) \| P_{0,r}(\cdot \mid Z_{<r})) \\ &= E_{P_R} \log L_r(Z_r \mid Z_{<r}) \in [0,\infty].\end{aligned}$$

Then the conditional density chain rule gives, P_R-almost surely,

$$\frac{dP_R}{dP_0}(Z_{1:R}) = \prod_{r=1}^{R} L_r(Z_r \mid Z_{<r}),$$

and the relative-entropy chain rule gives

$$D(P_R \| P_0) = \sum_{r=1}^{R} \kappa_r, \qquad D(q\|\mu) \le \sum_{r=1}^{R} \kappa_r.$$

Because the path has finitely many stages and $D(P_R\|P_0) < \infty$, every κ_r is finite. Without the finite-path-KL assumption, the same chain rule is valid in the extended nonnegative reals, but the later non-cancellation display would require a separate finiteness hypothesis.

Now consider a sequence with $D(q\|\mu) \to \infty$. Suppose there is a final event A such that

$$q(A) \ge \eta > 0, \qquad \log \frac{1}{\mu(A)} = D(q\|\mu) + o(D(q\|\mu)).$$

Fix $b \in (0,1]$, $\delta \in (0,1/2)$, and $\epsilon > 0$. Consider the pair

$$m_0 \equiv 1, \qquad m_1 = 1 - b\mathbf{1}_A.$$

For the lower bound, labels are noiseless. The same argument applies to any conditional label model whose full conditional law is identical in the two worlds whenever the queried object lies in A^c.

- *For the optimized mean, set $T_j = R_q(m_j) = E_q m_j$ and assume $\epsilon < b\eta/2$.*
- *For the shift, set $T_j = \Delta_q(m_j)$ and assume $\epsilon < b\eta/4$.*

Fix one of these targets. For a common nonanticipating final-object auditor $\mathcal{A}$ with n queries, let $\mathbb{P}_j^{\mathcal{A}}$ be its observable query-label transcript law in world m_j, under the convention in Appendix I.1, and let $\mathbb{M}_j^{(n)}$ be the reference observable transcript law obtained from n iid μ-queries with labels generated by the same observation experiment in world m_j. Discarded sampler randomization is marginalized; any public, logged, or retained state is included in both laws. Assume that the auditor outputs an estimator $\widehat{T}$ satisfying

$$\mathbb{P}_j^{\mathcal{A}}(|\widehat{T} - T_j| \le \epsilon) \ge 1 - \delta, \qquad j = 0,1,$$

and that

$$D\Big(\mathbb{P}_j^{\mathcal{A}} \Big\| \mathbb{M}_j^{(n)}\Big) \le K_{\mathrm{audit}}, \qquad j = 0,1.$$

Then, for all sufficiently large indices,

$$\frac{K_{\mathrm{audit}}}{1-2\delta} + \log n \geq D(q\|\mu) - o(D(q\|\mu)).$$

For the upper benchmark, assume in addition that labels are bounded and generated by the stable label kernel of Theorem 10.1, and that

$$E_q \left| \log \frac{dq}{d\mu} - D(q\|\mu) \right| = o(D(q\|\mu)).$$

Then, for every integer $\ell \geq 2$, $B = O(\epsilon^{-2}\log(1/\delta))$ independent single-swap blocks with private, discarded swap indices and $N = B\ell$ labels estimate $R_q(m)$ by averaging $\widehat{q}_\ell$, or estimate $\Delta_q(m)$ by averaging $\widehat{\Delta}_\ell$, and incur observable-transcript cost

$$K_{\mathrm{obs}} = B\left([D(q\|\mu) - \log \ell]_+ + o(D(q\|\mu))\right),$$

and, whenever $\log \ell \leq D(q\|\mu) + o(D(q\|\mu))$,

$$\frac{K_{\mathrm{obs}}}{B} + \log \frac{N}{B} = D(q\|\mu) + o(D(q\|\mu)).$$

If the swap indices are exposed, the corresponding augmented cost is instead $BD(q\|\mu)$; only the coarse $O(D(q\|\mu))$ upper rate survives unchanged.

Thus the audit frontier composes through the final density ratio. It becomes additive in cumulative stage KL only under a non-cancellation condition such as

$$D(q\|\mu) = \sum_r \kappa_r - o\left(\sum_r \kappa_r\right).$$

Proof. The final observable is a measurable map of the full trajectory, so data processing gives the first inequality. By the displayed conditional Radon–Nikodym definition,

$$\kappa_r = E_{Z_{<r}\sim P_R^{<r}} D(P_{R,r}(\,\cdot \mid Z_{<r})\|P_{0,r}(\,\cdot \mid Z_{<r})) \geq 0.$$

The conditional-density and relative-entropy chain rules then give

$$\frac{dP_R}{dP_0} = \prod_r L_r, \qquad D(P_R\|P_0) = \sum_r \kappa_r < \infty.$$

Since the sum has finitely many nonnegative terms, each κ_r is finite. For the optimized mean, the two target values are separated by

$$bq(A) \geq b\eta > 2\epsilon.$$

For the shift target, their separation is

$$b|q(A) - \mu(A)| \geq b(\eta - o(1)) > 2\epsilon$$

for all sufficiently large indices under the safer assumption $\epsilon < b\eta/4$. The two label laws agree outside A; applying Theorem 8.2 to $\mathbb{P}_j^{\mathcal{A}}$ relative to $\mathbb{M}_j^{(n)}$ therefore yields

$$\frac{K_{\text{audit}}}{1-2\delta} + \log n \geq \log \frac{1}{\mu(A)} - O_\delta(1)$$
$$= D(q\|\mu) - o(D(q\|\mu)).$$

Under final log-ratio concentration, Theorem 10.1 applied to $W = dq/d\mu$ gives the upper benchmark and the observable-cost equality. Substitution proves the non-cancellation statement. The lower bound also remains valid under any stronger exposed-state KL cap by data processing from the augmented transcript to the observable one.

□

J Complete Proofs: Representation and monitor complexity

J.1 Functional-analytic preliminaries

Assume throughout this part that μ is a probability measure on a standard-Borel space, so the usual identification $(L^1(\mu))^* = L^\infty(\mu)$ is available.

Lemma 13.1 (Distance to a closed subspace).

Let V be a closed linear subspace of $L^1(\mu)$ and $g \in L^1(\mu)$. Then

$$\operatorname{dist}_{L^1}(g, V) = \sup\big\{E_\mu[gh] : \|h\|_\infty \leq 1, \; E_\mu[vh] = 0 \; \forall v \in V\big\}.$$

The supremum is attained.

Proof. For any feasible h and any $v \in V$,

$$E_\mu[gh] = E_\mu[(g-v)h] \leq \|g-v\|_1 \|h\|_\infty \leq \|g-v\|_1.$$

Taking the infimum over v proves that the supremum is at most the distance.

For the reverse inequality, if $dist(g, V) = 0$, the feasible choice $h = 0$ already attains the value zero. Otherwise consider the quotient Banach space L^1/V. The norm of the nonzero coset $g + V$ is exactly $dist(g, V)$. By Hahn-Banach, there is a norm-one functional on the quotient that attains this norm at $g + V$. Pull it back to a norm-one functional on L^1 that annihilates V. By the identification $(L^1)^* = L^\infty$, this functional has the form $f \to E_\mu[fh]$ for some h with $||h||_\infty \leq 1$, and it satisfies

$$E_\mu[gh] = \operatorname{dist}_{L^1}(g, V).$$

This also proves attainment.

□

Lemma 13.2 (Weak-star compact proof of attainment).

If V is finite-dimensional, the feasible set

$$\{h : \|h\|_\infty \leq 1, \; h \perp V\}$$

is weak-star compact in L^∞, and $h \to E_\mu[gh]$ is weak-star continuous. Hence the extremizer may also be obtained directly by compactness.

Proof. The unit ball is weak-star compact by Banach-Alaoglu. Each constraint $E_\mu[vh] = 0$, $v \in V$, defines a weak-star closed hyperplane. The objective pairs h with the fixed $g \in L^1$, so it is weak-star continuous.

□

J.2 Exact representation duality

Let $V \subset L^1(\mu)$ be a closed linear subspace containing the constants. Define

$$V_\infty^\perp = \{h \in L^\infty(\mu) : E_\mu[vh] = 0 \; \forall v \in V\}$$

and, for $0 \le \rho \le 1/2$,

$$\mathcal{M}_{\rho,V} = \left\{ m = \frac{1}{2} + \rho h : \|h\|_\infty \le 1,\ h \in V_\infty^\perp \right\}.$$

Theorem 14.1 (Representation duality).

For $g = dq/d\mu - 1$,

$$\sup_{m \in \mathcal{M}_{\rho,V}} |\Delta_q(m)| = \rho \operatorname{dist}_{L^1(\mu)}(g, V).$$

Moreover,

$$\inf_{m \in \mathcal{M}_{\rho,V}} \Delta_q(m) = -\rho \operatorname{dist}_{L^1(\mu)}(g, V),$$

and both extrema are attained.

Proof. For $m = 1/2 + \rho h$,

$$\Delta_q(m) = E_\mu[g(1/2 + \rho h)] = \rho E_\mu[gh],$$

because $E_\mu g = 0$. Lemma 13.1 gives

$$\sup_{\|h\|_\infty \le 1,\, h \perp V} E_\mu[gh] = \operatorname{dist}_{L^1}(g, V).$$

The feasible set is symmetric under $h \to -h$, so the infimum is the negative of the supremum and the maximum absolute value equals the same quantity. Attainment follows from Lemma 13.1 or 13.2. The bound $\rho \le 1/2$ ensures $m \in [0, 1]$.

□

Corollary 14.2 (Explicit invisible adverse target).

Let h^* attain Lemma 13.1. Then

$$m^* = \frac{1}{2} - \rho h^*$$

has the same baseline feature-target moments as the constant target $1/2$,

$$E_\mu[vm^*] = \frac{1}{2} E_\mu v \qquad \forall v \in V,$$

but

$$\Delta_q(m^*) = -\rho \operatorname{dist}_{L^1}(g, V).$$

Proof. Because $1 \in V$, $E_\mu h^* = 0$; because h^* annihilates every $v \in V$, the moment identity follows. The shift identity is Theorem 14.1.

□

J.3 STATISTIC SUFFICIENCY

Let $T = T(O)$ be a statistic, let $\mathcal{G} = \sigma(T)$, and fix $0 < \rho \le 1/2$. Define

$$\mathcal{M}_{\rho,T} = \left\{ m = \frac{1}{2} + \rho h : \|h\|_\infty \le 1, E_\mu[h \mid \mathcal{G}] = 0 \right\}.$$

Lemma 15.1 (Annihilator of a statistic).

For $h \in L^\infty(\mu)$, the following are equivalent:

1. *$E_\mu[h|\mathcal{G}] = 0$;*

2. *$E_\mu[hv] = 0$ for every bounded $\mathcal{G}$-measurable v;*

3. *$E_\mu[hv] = 0$ for every $v \in L^1(\mathcal{G})$.*

Proof. The equivalence of 1 and 2 is the defining property of conditional expectation. Condition 3 implies 2. Conversely, for $v \in L^1(\mathcal{G})$, truncate v to bounded v_k; then $E[hv_k] = 0$, and $hv_k \to hv$ in L^1 because h is bounded.

□

Theorem 15.2 (Statistic sufficiency iff the density shift is measurable).

We have

$$\inf_{m \in \mathcal{M}_{\rho,T}} \Delta_q(m) = -\rho \operatorname{dist}_{L^1(\mu)} \left(g, L^1(\mathcal{G})\right),$$

and

$$\sup_{m \in \mathcal{M}_{\rho,T}} |\Delta_q(m)| = \rho \operatorname{dist}_{L^1(\mu)} \left(g, L^1(\mathcal{G})\right).$$

Equivalently, among pairs m_0, m_1 satisfying

$$\frac{1}{2} - \rho \le m_j \le \frac{1}{2} + \rho, \qquad E_\mu[m_0 \mid T] = E_\mu[m_1 \mid T],$$

we have

$$\sup |\Delta_q(m_1) - \Delta_q(m_0)| = 2\rho \operatorname{dist}_{L^1(\mu)} \left(g, L^1(\mathcal{G})\right).$$

Therefore the statistic rules out every nonzero target shift induced by a bounded T-invisible residual if and only if

$$g \in L^1(\mathcal{G}),$$

i.e. g is measurable with respect to T up to μ-null sets.

Proof. The subspace $L^1(\mathcal{G})$ is closed because it is the range of the contractive projection $f \to E[f|\mathcal{G}]$. By Lemma 15.1, its L^∞ annihilator is exactly the set of bounded h with $E[h|\mathcal{G}] = 0$. Apply Lemma 13.1 and repeat the calculation in Theorem 14.1.

For the pairwise formula, write

$$h = \frac{m_1 - m_0}{2\rho}.$$

The band restriction gives $||h||_\infty \leq 1$, and equality of conditional means gives $E[h|\mathcal{G}] = 0$. This proves the upper bound. Conversely, an extremizing h yields the pair

$$m_\pm = \frac{1}{2} \pm \rho h,$$

which attains the bound. Finally, the distance to the closed subspace is zero exactly when g belongs to that subspace.

□

Corollary 15.3 (Exact score tilts).

For a fixed-prompt exponential score tilt satisfying

$$\frac{dq}{d\mu}(o) = e^{tS(o)-\Lambda(t)},$$

the centered likelihood ratio

$$g(o) = e^{tS(o)-\Lambda(t)} - 1$$

is measurable with respect to S.

With prompts, suppose the prompt and output spaces are standard Borel and

$$\mu(dx, do) = \pi(dx)\mu_x(do), \qquad q(dx, do) = \pi(dx)q_x(do),$$

where μ_x and q_x are measurable probability kernels. Assume that, for π-almost every x,

$$\begin{aligned} q_x &\ll \mu_x, \\ \frac{dq_x}{d\mu_x}(o) &= e^{tS(x,o)-\Lambda_x(t)}, \\ \Lambda_x(t) &= \log \int e^{tS(x,o)}\, \mu_x(do) < \infty. \end{aligned}$$

and that a jointly measurable version of the displayed conditional likelihood ratio is used. Then

$$\frac{dq}{d\mu}(x, o) = e^{tS(x,o)-\Lambda_x(t)}$$

and hence

$$g(X, O) = e^{tS(X,O)-\Lambda_X(t)} - 1$$

is measurable with respect to (X, S). Thus (X, S) is sufficient in the sense of Theorem 15.2 at every nonzero scale $0 < \rho \le 1/2$ in this prompt-conditional exact-tilt model, while S alone need not be sufficient when the normalizer depends on the prompt. This prompted conclusion is not asserted for arbitrary prompt mixtures with different prompt marginals.

Proof. The fixed-prompt claim follows directly from the displayed likelihood ratio. In the prompted case, for every measurable set B,

$$\begin{aligned} q(B) &= \int \pi(dx) \int 1_B(x, o) e^{tS(x,o)-\Lambda_x(t)}\, \mu_x(do) \\ &= \int_B e^{tS(x,o)-\Lambda_x(t)}\, \mu(dx, do). \end{aligned}$$

Therefore the displayed function is a version of $dq/d\mu$. Its centered form is (X, S)-measurable, so Theorem 15.2 gives the sufficiency conclusion.

□

J.4 Deterministic and randomized monitor complexity

Let $G = g(O)$ under $O \sim \mu$.

Theorem 16.1 (Deterministic finite-range monitor equals scalar quantization).

For a deterministic standard-Borel-valued statistic $T = T(O)$, define

$$\begin{aligned} \mathcal{E}_{\det}(T) &:= \inf_{v \in L^1(P_T)} E_\mu |G - v(T)| \\ &= \operatorname{dist}_{L^1(\mu)}(g, L^1(\sigma(T))). \end{aligned}$$

where P_T is the law of T under μ.

For every integer $M \ge 1$,

$$\inf_{|\operatorname{range}(T)| \le M} \mathcal{E}_{\det}(T) = D_M(G),$$

where

$$D_M(G) = \inf_{\substack{Q:\mathbb{R}\to\mathbb{R} \text{ Borel measurable} \\ |\operatorname{range}(Q)| \le M}} E|G - Q(G)|$$

is the optimal deterministic scalar M-level L^1 quantization distortion. Consequently, for every fixed $0 < \rho \le 1/2$, Theorem 15.2 implies that the smallest worst-case bounded hidden-target vulnerability among deterministic M-value monitors is $\rho D_M(G)$. The same formula is trivially valid at $\rho = 0$, where the perturbation class has zero scale. Randomized assignment to a fixed codebook cannot improve the reconstruction infimum under absolute loss.

Proof. Because T is standard-Borel-valued, the Doob-Dynkin lemma identifies $L^1(\sigma(T))$, modulo μ-null sets, with the class

$$\{v(T) : v \in L^1(P_T)\}.$$

This proves the first displayed equality. Let the distinct reconstruction values of $v(T)$ be a codebook C of size at most M. Conditional on $G = x$, the reconstruction induces a probability distribution on C; its conditional loss is a convex combination of the numbers $|x-c|$, $c \in C$, and is therefore at least

$\min_{c\in C}|x-c|$. Choosing a measurable minimizing codepoint (for example, the first minimizer in a fixed ordering of C) gives an M-level deterministic scalar quantizer of G with no larger distortion. Hence every finite-range statistic has distortion at least $D_M(G)$. Conversely, an M-level scalar quantizer $Q(G)$ is implemented by the statistic $T(O)=Q(g(O))$. This also proves that randomized assignment to a fixed codebook cannot improve the infimum.

□

Theorem 16.2 (Randomized-monitor vulnerability is an operator-range distance).

Let $\mathsf{K}(dt \mid o)$ be a Markov kernel from the episode space to a standard-Borel monitor-output space. Let $P_{\mathsf{K}}(do,dt)=\mu(do)\mathsf{K}(dt \mid o)$ be the joint law and P_T its output marginal. For bounded output functions define the conditional-expectation operator

$$(\mathsf{K}v)(o)=E[v(T)\mid O=o]=\int v(t)\mathsf{K}(dt \mid o),$$

and the closed linear subspace

$$\mathcal{V}_{\mathsf{K}}=\overline{\{\mathsf{K}v : v\in L^\infty(P_T)\}}^{L^1(\mu)}.$$

For $0<\rho\le 1/2$, let

$$\mathcal{M}_{\rho,\mathsf{K}}=\big\{m=\tfrac{1}{2}+\rho h(O) : \|h\|_\infty\le 1,\ E[h(O)\mid T]=0\big\}.$$

Then

$$\sup_{m\in\mathcal{M}_{\rho,\mathsf{K}}}|\Delta_q(m)|=\rho\,\mathrm{dist}_{L^1(\mu)}(g,\mathcal{V}_{\mathsf{K}}),$$

and the negative extremum is attained. In particular, the randomized monitor rules out every nonzero target shift induced by a bounded episode-only invisible perturbation if and only if

$$g\in\mathcal{V}_{\mathsf{K}}.$$

For a deterministic monitor, $\mathcal{V}_{\mathsf{K}}=L^1(\sigma(T))$, so this theorem reduces to Theorem 15.2.

Proof. For bounded $h=h(O)$,

$$E[h(O)\mid T]=0$$

if and only if

$$E[h(O)v(T)]=0\qquad\text{for every } v\in L^\infty(P_T).$$

By conditioning on O, this is equivalent to

$$E_\mu[h\,\mathsf{K}v]=0\qquad\text{for every bounded } v.$$

Thus the admissible h are exactly the bounded annihilator of $\mathcal{V}_{\mathsf{K}}$. Lemma 13.1 and the calculation in Theorem 14.1 give the displayed distance and attainment. For a deterministic statistic, bounded functions of T are dense in $L^1(\sigma(T))$, proving the reduction.

□

Proposition 16.3 (Mutual-information rate-distortion is a reconstruction benchmark only).

Fix $B \geq 0$. Allow Markov kernels from the episode space to arbitrary standard-Borel output spaces $\mathcal{T}$. For each such channel, let the inner infimum range over Borel measurable reconstructions $v : \mathcal{T} \to \mathbb{R}$. Define the scalar reconstruction objective

$$\mathcal{R}_{\mathrm{rec}}(B) = \inf_{\mathsf{K}:\, I_\mu(O;T) \leq B} \inf_{v:\mathcal{T}\to\mathbb{R} \text{ Borel measurable}} E|G - v(T)|.$$

Then

$$\mathcal{R}_{\mathrm{rec}}(B) = D_G(B),$$

where

$$D_G(B) = \inf_{P_{\widehat{G}|G}:\, I(G;\widehat{G}) \leq B} E|G - \widehat{G}|$$

is the scalar Shannon rate-distortion function under absolute loss.

*This equality does **not** imply that optimized randomized-monitor hidden-target vulnerability equals $\rho D_G(B)$. The vulnerability is the operator-range quantity in Theorem 16.2.*

Proof. For any channel and reconstruction $v(T)$, data processing gives

$$I(G; v(T)) \leq I(O; T) \leq B,$$

so the distortion is at least $D_G(B)$. Conversely, compute $G = g(O)$ and pass it through any channel that attains or approximates the rate-distortion infimum. Since $O \to G \to \widehat{G}$ is a Markov chain and G is a deterministic function of O,

$$I(O; \widehat{G}) = I(G; \widehat{G}) \leq B.$$

This proves the reconstruction equality.

To see why it is not a vulnerability identity, let O be uniform on $-1, +1$, put $g(O) = O$ (so $W = 1+g$ is a valid likelihood ratio), and let T be the output of a binary symmetric channel with crossover $\varepsilon \in (0, 1/2)$. The matrix of K is

$$\begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix},$$

whose determinant is $1 - 2\varepsilon \neq 0$. Hence $\mathcal{V}_{\mathsf{K}}$ is the full two-dimensional function space and the hidden-target vulnerability is zero. Yet

$$I(O; T) = \log 2 - h_2(\varepsilon) < \log 2.$$

It remains to justify strict positivity of the reconstruction distortion at this information level. Let $\widehat{G} = v(T)$ be any reconstruction and put $\widetilde{G} = \mathrm{sign}(\widehat{G})$, with ties broken arbitrarily. Whenever $\widetilde{G} \neq G$, one has $|G - \widehat{G}| \geq 1$, so

$$P(\widetilde{G} \neq G) \leq E|G - \widehat{G}| =: D.$$

For $D < 1/2$, the binary Fano inequality and data processing give

$$I(G; \widehat{G}) \geq I(G; \widetilde{G}) \geq \log 2 - h_2(D).$$

Consequently, for every $B < \log 2$,

$$D_G(B) \geq h_2^{-1}(\log 2 - B) > 0,$$

where the inverse is taken on $[0, 1/2]$; for $D \geq 1/2$ strict positivity is immediate. Taking $B = I(O;T)$ proves that the reconstruction benchmark is positive although the operator-range vulnerability is zero.

□

Theorem 16.4 (Gaussian deterministic finite-range complexity is logarithmic in KL).

Let $X \sim N(0,1)$ and $t \geq 0$ (the case $t < 0$ follows by replacing X with $-X$), and put

$$L_t = e^{tX - t^2/2}, \qquad d = t^2/2.$$

Let $D_M(L_t)$ be the optimal deterministic M-level $L^1(\mu)$ distortion. There is a universal constant $c > 0$ such that, for $d \geq 1$ and every $M \geq 1$ integer,

$$D_M(L_t) \geq c \min\left\{1, \frac{\sqrt{d}}{M}\right\}.$$

For every fixed $\tau \in (0,1)$, there are constants $0 < c_\tau < C_\tau < \infty$ such that, for all large d,

$$c_\tau \sqrt{d} \leq M_\tau(d) := \min\{M : D_M(L_t) \leq \tau\} \leq C_\tau \sqrt{d}.$$

Thus

$$\log_2 M_\tau(d) = \frac{1}{2} \log_2 d + O_\tau(1).$$

The asymptotic assertion concerns $d \to \infty$; at $d = 0$, L_t is identically one and the distortion is zero. The same statements hold for $g_t = L_t - 1$.

Proof: change of measure. Under $Q_t = N(t, 1)$,

$$Z = \log L_t \sim N(d, 2d).$$

For every reconstruction $\widehat{L}$,

$$E_\mu |L_t - \widehat{L}| = E_{Q_t} \left| 1 - \frac{\widehat{L}}{L_t} \right|.$$

□

Proof: universal lower bound. Suppose the reconstruction has codepoints $c_1, ..., c_M$. Fix $a \in (0,1)$. On the event of relative error at most a, the selected codepoint is positive and Z lies in an interval of length

$$\ell(a) = \log \frac{1+a}{1-a}.$$

The density of $N(d, 2d)$ is at most $(4\pi d)^{-1/2}$, so

$$Q_t\left(\left|1-\frac{\widehat{L}}{L_t}\right|\le a\right)\le\frac{M\ell(a)}{\sqrt{4\pi d}}.$$

For $a\le 1/2$, $\ell(a)\le 3a$. Choosing

$$a=c_0\min\left\{1,\frac{\sqrt{d}}{M}\right\}$$

with sufficiently small universal c_0 makes the covered probability at most $1/2$, and hence the expected relative error is at least $a/2$.

For the fixed-distortion lower bound, no attainment of the quantization infimum is needed. If $D_M(L_t)\le\tau$, choose, for every sufficiently small $\epsilon'>0$, an M-level reconstruction with expected relative error at most $\tau+\epsilon'$, and put

$$a_{\tau,\epsilon'}=\frac{1+\tau+\epsilon'}{2}\in(\tau+\epsilon',1).$$

Markov's inequality gives covered mass at least

$$\frac{1-\tau-\epsilon'}{1+\tau+\epsilon'}.$$

The same interval bound yields

$$M\ge\frac{1-\tau-\epsilon'}{1+\tau+\epsilon'}\frac{\sqrt{4\pi d}}{\ell(a_{\tau,\epsilon'})}.$$

Letting $\epsilon'\downarrow 0$ gives

$$M\ge\frac{1-\tau}{1+\tau}\frac{\sqrt{4\pi d}}{\ell((1+\tau)/2)}=:c_\tau\sqrt{d}.$$

□

Proof: upper bound. Choose $R=R_\tau$ so that $2\bar{\Phi}(R)\le\tau/2$. Under Q_t, $X\sim N(t,1)$. Partition $[t-R,t+R]$ into $M-1$ equal intervals, reconstruct L_t by its midpoint value inside, and use codepoint zero outside. Then

$$D_M(L_t)\le 2\bar{\Phi}(R)+\exp\left(\frac{tR}{M-1}\right)-1.$$

Choosing

$$M-1\ge\frac{tR}{\log(1+\tau/2)}$$

makes the distortion at most τ. Since $t=\sqrt{2d}$, this gives the upper bound. Shifting every codepoint by -1 proves the identical result for g_t.

□

J.5 TAIL LEAKAGE AND STRUCTURED MONITOR FAMILIES

For a linear subspace $V \subset L^1(\mu)$ and event A, define

$$\beta_V(A) = \sup_{v \in V,\, v \neq 0} \frac{E_\mu[|v|1_A]}{E_\mu|v|}.$$

If $V = 0$, set $\beta_V(A) = 0$. In all monitor applications below, V contains the constants and is therefore nonzero.

Theorem 17.1 (Tail-leakage lower bound).

For $g = dq/d\mu - 1$,

$$\mathrm{dist}_{L^1(\mu)}(g, V) \geq \frac{[q(A) - \mu(A) - 2\beta_V(A)]_+}{1 + \beta_V(A)}.$$

Proof. Fix $v \in V$ and put

$$D_v = E_\mu|g - v|, \qquad \beta = \beta_V(A).$$

Because

$$q(A) - \mu(A) = E_\mu[g1_A],$$

we have

$$\begin{aligned} q(A) - \mu(A) &\leq E_\mu[|g - v|1_A] + E_\mu[|v|1_A] \\ &\leq D_v + \beta E_\mu|v|. \end{aligned}$$

Also

$$E_\mu|v| \leq E_\mu|g - v| + E_\mu|g| \leq D_v + 2.$$

Therefore

$$q(A) - \mu(A) \leq (1 + \beta)D_v + 2\beta.$$

Rearranging and taking the infimum over v proves the claim, with a positive part because distances are nonnegative.

□

Corollary 17.2 (Structured tail-leakage principle).

As $d \to \infty$, let $q_d \ll \mu$, $g_d = dq_d/d\mu - 1$, and let A_d satisfy

$$q_d(A_d) \geq \eta > 0, \qquad \mu(A_d) = e^{-d+o(d)}.$$

Let V_R be a nested family of closed linear monitor subspaces containing the constants, such that

$$\beta_{V_R}(A_d) \leq \exp\{\psi(R) - d/2 + o(d)\}.$$

Here the remainder is required to be $o(d)$ along the particular sequence $R = R(d)$ considered below (uniformity over all R is not otherwise asserted). If, for some fixed $\xi > 0$,

$$\psi(R(d)) \le (1/2 - \xi)d,$$

then

$$\mathrm{dist}_{L^1(\mu)}(g_d, V_{R(d)}) \ge \eta - o(1).$$

Consequently, for every $0 \le \rho \le 1/2$*, Theorem 14.1 yields a bounded invisible target with optimized shift at most*

$$-\rho(\eta - o(1)).$$

Proof. The leakage bound gives $\beta_{V_R}(A_d) \to 0$; the baseline mass also tends to zero. Substitute into Theorem 17.1.

□

J.6 Gaussian polynomial monitors

Let $X \sim N(0,1)$ and $t > 0$, and put

$$q_t = N(t,1), \qquad g_t = e^{tX - t^2/2} - 1, \qquad d = t^2/2.$$

For each nonnegative integer $r \in \mathbb{N}_0$, let

$$V_r = \mathrm{span}\{1, X, ..., X^r\}.$$

Lemma 18.1 (Polynomial tail leakage).

For every polynomial p *of degree at most* r *and every measurable event* A*,*

$$E[|p(X)|1_A] \le 3^r \sqrt{P(A)}\, E|p(X)|.$$

In particular, for $A_t = X \ge t$*,*

$$\beta_{V_r}(A_t) \le 3^r \sqrt{\bar{\Phi}(t)}.$$

Proof. Gaussian hypercontractivity for degree-r polynomials gives

$$\|p\|_4 \le 3^{r/2}\|p\|_2.$$

Interpolation between L^1 and L^4 gives

$$\|p\|_2 \le \|p\|_1^{1/3}\|p\|_4^{2/3}.$$

Substitute hypercontractivity:

$$\|p\|_2 \le 3^{r/3}\|p\|_1^{1/3}\|p\|_2^{2/3},$$

hence

$$\|p\|_2 \le 3^r\|p\|_1.$$

Cauchy-Schwarz now gives

$$E[|p|1_A] \le \|p\|_2 \sqrt{P(A)} \le 3^r \sqrt{P(A)}\, \|p\|_1.$$

□

Theorem 18.2 (Linear-in-KL lower bound on polynomial degree).

For every constant

$$c_- < \frac{1}{2\log 3},$$

there is d_0 such that, whenever $d \ge d_0$ and $r \le c_- d$,

$$\mathrm{dist}_{L^1(\mu)}(g_t, V_r) \ge \frac{1}{4}.$$

In fact the lower bound tends to $1/2$.

Proof. For $A_t = X \ge t$,

$$q_t(A_t) = \frac{1}{2}, \qquad \mu(A_t) = \bar{\Phi}(t) = e^{-d+o(d)}.$$

By Lemma 18.1,

$$\beta_{V_r}(A_t) \le \exp\left(r\log 3 - \frac{d}{2} + o(d)\right).$$

If $r \le c_- d$ with $c_- < 1/(2\log 3)$, the exponent tends to $-\infty$, so the leakage tends to zero. Theorem 17.1 gives

$$\mathrm{dist}_{L^1}(g_t, V_r) \ge \frac{1/2 - o(1)}{1 + o(1)} = \frac{1}{2} - o(1).$$

□

Lemma 18.3 (Hermite truncation upper bound).

Let H_j be the probabilists' Hermite polynomials, normalized by

$$E[H_j(X)H_k(X)] = j!\, 1_{\{j=k\}}.$$

Then

$$g_t(X) = \sum_{j=1}^{\infty} \frac{t^j}{j!} H_j(X)$$

in $L^2(\mu)$. If

$$v_r(X) = \sum_{j=1}^{r} \frac{t^j}{j!} H_j(X),$$

then

$$\|g_t - v_r\|_1 \le \|g_t - v_r\|_2 = \left(\sum_{j=r+1}^{\infty} \frac{t^{2j}}{j!} \right)^{1/2}.$$

For $\lambda = t^2 = 2d$ and $r = \lceil e\lambda \rceil$, the right-hand side tends to zero. Hence, for every fixed $\tau > 0$,

$$\mathrm{dist}_{L^1}(g_t, V_r) \le \tau$$

for all sufficiently large d, with

$$r \le (2e + o(1))d.$$

Proof. The generating function identity

$$e^{tx - t^2/2} = \sum_{j=0}^{\infty} \frac{t^j}{j!} H_j(x)$$

holds in $L^2(\mu)$ because the squared coefficient sum is

$$\sum_{j=0}^{\infty} \frac{t^{2j}}{j!} = e^{t^2} < \infty.$$

Subtracting the $j = 0$ term gives the expansion of g_t. Orthogonality gives the exact residual norm. Put $a_j = \lambda^j / j!$. For $j \ge r + 1$ with $r \ge e\lambda$,

$$\frac{a_{j+1}}{a_j} = \frac{\lambda}{j+1} \le \frac{1}{e} + o(1) < 1.$$

Thus the tail is bounded by a constant multiple of its first term. Stirling's lower bound gives, for $j \ge e\lambda$,

$$a_j = \frac{\lambda^j}{j!} \le \frac{1}{\sqrt{2\pi j}} \left(\frac{e\lambda}{j} \right)^j \le \frac{1}{\sqrt{2\pi j}}.$$

Therefore

$$\sum_{j=r+1}^{\infty} a_j = O(\lambda^{-1/2}),$$

and the L^2 residual is $O(\lambda^{-1/4}) \to 0$. Since $r = \lceil e\lambda \rceil = \lceil 2ed \rceil$, the degree statement follows.

□

Theorem 18.4 (Polynomial monitor degree threshold).

For fixed $\tau \in (0, 1/4)$, define

$$r^*_{L^1}(d, \tau) = \min\{r \in \mathbb{N}_0 : \mathrm{dist}_{L^1(\mu)}(g_t, V_r) \le \tau\}.$$

Then

$$r^*_{L^1}(d,\tau) = \Theta_\tau(d).$$

More precisely,

$$\liminf_{d\to\infty} \frac{r^*_{L^1}(d,\tau)}{d} \geq \frac{1}{2\log 3}, \qquad \limsup_{d\to\infty} \frac{r^*_{L^1}(d,\tau)}{d} \leq 2e.$$

Proof. For every $c_- < 1/(2\log 3)$, Theorem 18.2 shows that degree at most $c_- d$ has distance at least $1/4 > \tau$ eventually. Letting c_- increase to the boundary proves the liminf. Lemma 18.3 proves the limsup.

□

Corollary 18.5 (Constant invisible loss below linear degree).

For every $0 \leq \rho \leq 1/2$ and every $c_- < 1/(2\log 3)$, if $r \leq c_- d$ and d is large, then

$$\inf_{m \in \mathcal{M}_{\rho, V_r}} \Delta_{q_t}(m) \leq -\rho/4.$$

Proof. Combine Theorems 14.1 and 18.2.

□