THE PRICE OF CHOICE IN SELECTIVE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A vendor may claim that *some* rule from a predeclared menu—a score threshold or a union of K score intervals—lets its selective classifier keep a fraction κ of inputs while erring on at most a fraction α of them. Refuting such an existential claim means ruling out every permitted rule at once. The intersection–union test (IUT) does this without a multiplicity correction, but its sample cost as the menu grows, and how close that cost is to the best valid audit, are not captured by the level argument. We characterize both. For a fixed interior risk target α, menus of K intervals and a risk gap γ, in the small-gap, large-menu regime of our theorem the minimax sample size at each fixed coverage floor $\kappa \in (0, 1)$ is $K(1-\kappa)/(\kappa\gamma^2)$ up to logarithmic factors, and the exact IUT attains it: *menu richness costs K times the abstention odds*, even though validity costs nothing. The lower bound holds for audits that know the score distribution; the upper bound extends to any menu of VC dimension d. We then explain the linear cost. Every certificate that never rejects while some rule looks feasible on the audit data—the IUT, Bonferroni, DKW and uniform-convergence bounds—pays it even at laws where $\sqrt{K}/\gamma^2$ samples suffice, because the best-looking rule is optimistically biased. We then identify when the gap can and cannot be avoided: against constant-risk alternatives a reference-free audit reaches the testing scale, including the square-root rate on the sign family in its stated small-gap regime, whereas over a rich class with unknown heterogeneous risk the linear cost is unavoidable up to a logarithm. In simulation with $K = 4$, the raw menu cost changes 7.7-fold between $\kappa = .5$ and $.95$; after dividing by the abstention odds and the theorem-motivated localized-VC logarithmic normalization used in Figure 1, the two endpoint values differ by about 5%. Held-out model families show that the same feasible-mixture mechanism is not confined to synthetic constructions: across three contemporary VLM judges, all 60 split-specific confidence-gate families have positive exact-hiding windows and 28/30 contracts selected on audit-A remain hidden on untouched audit-B; on the cleaner multi-annotator Polaris target, hiding persists on all 20 splits but is much narrower and transfers in 3/10 resplits. Frozen image and multimodal classifiers separately show boundary behavior consistent with the exact IUT guarantee and strong power relative to the practical certificates we compare.

1 INTRODUCTION

Before a selective classifier is deployed—in procurement, certification or a regulatory feasibility check—its vendor often claims that it *can be configured* to meet a standard: some permitted acceptance rule retains at least 90% of inputs while erring on at most 10% of the accepted ones. The menu of permitted rules is fixed in advance, for example thresholds on a confidence score or unions of a few score intervals. An auditor with fresh labeled data who wants to refute the claim cannot test one rule of its choice: showing that one rule fails leaves the vendor free to pick another. The null hypothesis is a *union* over the menu, and a rejection must rule out every rule at once. This is the dual of risk control and Learn-then-Test, which certify one rule selected by calibration (Chow, 1970; El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017; Bates et al., 2021; Angelopoulos et al., 2024; 2025).

The intersection–union test (IUT) is the natural audit: reject when every rule fails its own valid test. Because a true claim has one fixed witness rule, no multiplicity correction is needed (Berger, 1982; Berger and Hsu, 1996). This invites a tempting conclusion—that the size of the menu is free. It is

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.

not. We characterize the statistical price of menu richness, compare the exact IUT with the best valid audit, and identify assumptions under which the gap can close.

The main result in one line. For menus of at most K score intervals, a fixed interior risk target α, a fixed coverage floor $\kappa \in (0,1)$ and risk gap γ, the number of audit samples needed to refute every robustly false claim is

$$N \asymp \underbrace{\frac{1-\kappa}{\kappa}}_{\text{abstention odds}} \cdot \frac{K}{\gamma^2} \qquad \text{in the small-gap, large-menu regime, up to logarithmic factors,} \qquad (1)$$

and the exact IUT attains it (Theorem 1). This is the price of the menu, and it arises because the vendor's freedom lies in *which* small part of the score line to abstain on (or, for $\kappa < 1/2$, to accept): the auditor must rule out that the errors happen to be concentrated there, and the error mass of such a region fluctuates on the scale of that region, not of the whole sample. Richness is therefore cheap at high floors: at $K = 4$ in our simulations the menu cost at $\kappa = .9$ is 4.4 times smaller than at $\kappa = .5$ (Figure 1).

Contributions. Table 1 summarizes the results; every lower bound holds for audits that are valid over the full, unrestricted null.

1. *The abstention-odds law and optimality of the IUT* (Section 3). For any menu of VC dimension d the IUT needs $\tilde{O}(d(1-\kappa+\gamma)/(\kappa\gamma^2))$ samples (Proposition 2); for unions of K intervals, in the theorem's small-gap regime, no valid audit, even one that knows the score distribution, needs fewer than $\Omega(K(1-\kappa)/(\kappa\gamma^2 \log K))$ at any fixed floor κ (Theorem 1). The law covers the high floors used in practice ($\kappa = .9$ in our experiments), where richness is several times cheaper than at $\kappa = 1/2$.
2. *A plug-in barrier* (Section 4). We call a certificate EFC if it never rejects while some rule looks feasible on the audit data; the IUT, Bonferroni, DKW and uniform-convergence bounds all are. On explicit families with k effective groups, every valid EFC certificate needs $\Omega(k/\gamma^2)$ samples at a law where $\Theta(\sqrt{k}/\gamma^2)$ suffice (Theorems 2 and 3), and this carries over to interval menus (Remark 1). The obstruction is a winner's curse: the best empirical rule is optimistically biased by $\sqrt{k/n}$.
3. *What a better audit must know* (Section 5). An unknown score distribution does not force the linear cost: against constant-risk alternatives, a reference-free audit valid for the unrestricted null reaches the testing scale and, on the uniform sign family in the stated small-gap regime, the square-root rate (Theorem 4). By contrast, uncertainty about heterogeneous conditional risk can force the linear cost up to a logarithm over the class of Theorem 5.
4. *Evidence* (Section 6). Simulations illustrate the coverage-floor dependence and the testing–estimation separation. Held-out families of frozen classifiers and VLM-judge confidence gates exhibit feasible-mixture hiding; on three VLM judges 28/30 audit-A midpoint contracts remain hidden on untouched audit-B, while a clean multi-annotator control retains positive but much smaller windows. Separate frozen-classifier audits show boundary behavior consistent with exact IUT level and strong finite-sample power against Bonferroni, DKW, total-loss and a valid non-EFC universal-inference audit.

For practitioners the message is short. Use the exact IUT: it needs no correction, runs in $O(MKn)$ time over M score cells, and is minimax up to logarithms in the regime of Theorem 1. Within that regime, budget for the menu with (1): richness is cheap at high coverage floors and the menu penalty grows as the floor decreases. The held-out classifier and VLM-judge studies show why a family-level audit matters in practice: multiple individually failing gates can still hide the reference through feasible mixtures, while the constant-risk square-root shortcut requires additional structure.

Positioning and contributions. Classical analyses of the IUT concern a fixed number of normal means (Lehmann, 1952; Berger, 1989); tests of many moment inequalities and reality checks have an *intersection* null and pay only $\log K$ (White, 2000; Chernozhukov et al., 2019); property testing of function classes separates known-distribution from distribution-free testing (Kearns and Ron, 2000; Blais et al., 2021). Our null is a union of convex sets of laws with a ratio-type risk constraint, validity

Table 1: Minimax sample sizes of valid audits (up to constants; $\tilde{\Omega}, \tilde{O}$ hide logarithms). Row 1: robustly infeasible alternatives $\mathcal{P}_\gamma$ (Section 3); the interval bounds need large K, an interior risk target α bounded away from 0 and 1, and $\gamma \lesssim \min\{\kappa, 1-\kappa\}/\log K$. Rows 2–4: the sign families of Section 4 ($\kappa = 1/2$, $M = 2B$ cells, $k = B$ in rows 3–4, large B, $\gamma \le \gamma_0$). "Valid" means level δ over the *unrestricted* null.

Known to auditor	Power required against	Minimax sample size	EFC certificates
Menu only	$\mathcal{P}_\gamma$ (robust margin γ)	K intervals, fixed κ: $\tilde{\Theta}\big(\frac{K(1-\kappa)}{\kappa\gamma^2}\big)$ (Thm. 1)	IUT: $\tilde{O}\big(\frac{d(1-\kappa+\gamma)}{\kappa\gamma^2}\big)$, VC dim. d
Reference P_0	P_0	$\Theta(\sqrt{k}/\gamma^2)$ (Thm. 2)	$\Omega(k/\gamma^2)$; IUT: $\Theta(k/\gamma^2)$ (Thm. 3)
Cells only	constant risk, any marginal	$\Theta(\sqrt{B}/\gamma^2)$ if $\gamma \le (2B)^{-1/4}$ (Thm. 4)	$\Omega(B/\gamma^2)$ at P_0
Cells and marginal	heterogeneous risk	$\tilde{\Theta}(B/\gamma^2)$ (Thm. 5)	IUT: $O(B/\gamma^2)$

is required over all of it, and the question is how the sample size depends on the menu. The lower-bound tools—mixture second moments, identity and independence testing, moment matching—are classical (Ingster, 1987; Valiant and Valiant, 2017; Diakonikolas and Kane, 2016; Cai and Low, 2011). Our contributions are the EFC barrier, the abstention-odds law, and three constructions they require: a padded moment-matching prior whose null draws stay *existentially* feasible while the alternative draws are robustly infeasible against every measurable set at any prescribed floor; a witness that controls the coverage screen and the risk simultaneously; and a localization of the IUT to the smaller side of the floor.

2 The audit problem

Observe i.i.d. pairs (S, L) of score and binary loss. Freeze a measurable family $\mathcal{A}$ of acceptance rules (the *menu*) and a contract (κ, α) independently of the audit sample. A law Q is feasible if

$$\exists A \in \mathcal{A}: \quad c_Q(A) = Q(S \in A) \ge \kappa, \qquad r_Q(A) = Q(L = 1 \mid S \in A) \le \alpha. \tag{2}$$

Let $\mathcal{H}(\mathcal{A})$ be this existential null. It is *unrestricted*: score probabilities and conditional losses may both change. For a class $\mathcal{P}$ of infeasible laws, the minimax sample size is

$$N(\mathcal{P}, \mathcal{H}; \delta, \beta) = \inf\Big\{n \ge 1 : \exists \phi_n \in [0,1],\ \sup_{Q \in \mathcal{H}} \mathbb{E}_{Q^n}\phi_n \le \delta,\ \inf_{P \in \mathcal{P}} \mathbb{E}_{P^n}\phi_n \ge \beta\Big\}. \tag{3}$$

N_{IUT} is the same quantity with ϕ_n restricted to the exact IUT below. For a single law P_0 we write $N(P_0, \mathcal{H})$; the optimal test may then use P_0, so this is an information benchmark as in local minimax testing (Balakrishnan and Wasserman, 2019). Larger classes $\mathcal{P}$ encode what the auditor does not know.

The exact IUT. For a rule A, let T_A and E_A count the audit observations it accepts and the accepted errors; $T_A \sim \mathrm{Bin}(n, c(A))$ and $E_A \mid T_A = t \sim \mathrm{Bin}(t, r(A))$. Let U_A be the one-sided Clopper–Pearson upper bound for coverage at error ζ_c and L_A the lower bound for risk at error ζ_r (Clopper and Pearson, 1934), and define the *certified frontier*

$$L_{\mathrm{IUT}}(\mathcal{A}, \kappa) = \min_{A:\, U_A \ge \kappa} L_A, \qquad \min \varnothing = +\infty. \tag{4}$$

It reads: with confidence $1 - \delta$, no permitted rule of coverage at least κ has risk below L_{IUT}.

Proposition 1 (Validity without correction)**.** *Rejecting feasibility when $L_{\mathrm{IUT}} > \alpha$ has level at most $\zeta_c + \zeta_r \le \delta$ over the unrestricted null.*

Under the null a fixed population witness A^* exists, and rejection requires failure of one of its two bounds. For unions of at most K intervals over M frozen score cells a dynamic program computes (4) exactly in $O(MKn)$ time (Appendix G). We use $\zeta_c = \zeta_r = \delta/2$.

Plug-in (EFC) certificates. Call A *empirically feasible* if $T_A \geq \kappa n$ and $E_A \leq \alpha T_A$.

Definition 1 (EFC audits)**.** An audit ϕ_n is *empirical-feasibility consistent* (EFC) if $\phi_n = 0$ whenever some $A \in \mathcal{A}$ is empirically feasible.

EFC audits never reject a claim that the vendor can apparently meet on the audit data itself, which makes their decisions easy to explain. Every certificate that screens with a coverage upper bound no smaller than the empirical coverage and rejects via risk lower bounds no larger than the empirical risk is EFC. This includes the IUT (for $\zeta_c, \zeta_r \leq 1/2$), its Bonferroni variant, DKW and Rademacher uniform-convergence certificates, and the total-loss certificate (Appendix B.1). EFC plays the role of monotonicity in classical IUT theory (Perlman and Wu, 1999).

3 THE ABSTENTION-ODDS LAW

For an arbitrary menu we place the power requirement on alternatives that are *robustly* infeasible:

$$\mathcal{P}_\gamma(\mathcal{A}) = \{P : \; r_P(A) \geq \alpha + \gamma \text{ for every } A \in \mathcal{A} \text{ with } c_P(A) \geq \kappa - \gamma\}.$$

The class allows arbitrary score marginals and arbitrary, including informative, risk functions. The coverage slack γ excludes rules that sit just below the floor with very low risk, which no finite audit can resolve.

Proposition 2 (The IUT for every menu)**.** *Let $\mathcal{A}$ have VC dimension d, $0 < \delta < \beta < 1$, $\kappa \in (0, 1]$, $0 < \gamma \leq \kappa/2$ and $\alpha + \gamma \leq 1$. For absolute constants C: (i) the exact IUT has power at least β uniformly over $\mathcal{P}_\gamma(\mathcal{A})$ once $n \geq C\{d + \log(2/\delta) + \log(4/(1-\beta))\}/(\kappa\gamma)^2$; (ii) if some $A_0 \in \mathcal{A}$ has coverage at least κ under some score law, every valid audit needs $n \geq 2(\beta-\delta)^2\alpha(1-\alpha)/\gamma^2$; (iii)* localization: *the exact IUT has power at least β over $\mathcal{P}_\gamma(\mathcal{A})$ whenever*

$$n \geq C\,\frac{(1-\kappa+\gamma)\,d\log(2en/d) + \log\{8/(\delta(1-\beta))\}}{\kappa\gamma^2}.$$

Part (i) combines VC uniform deviations with Clopper–Pearson inversion; part (ii) is a two-point bound (Appendix B.3). Part (iii) is the refinement that matters at high floors. A rule that passes the coverage screen differs from the whole score line only on a region of probability at most $1 - \kappa + \gamma$, and it accepts at least a fraction $\kappa/2$ of the data. Writing its error mass as the overall error mass minus the error mass of the abstained region, only the latter varies over the menu. Relative (Bernstein-type) VC deviations bound it by $\sqrt{(1-\kappa+\gamma)d/n}$ up to logarithms, which for $\kappa \geq 1/2$ is also the order of the risk error. When $\kappa < 1/2$ the accepted region is the small side, its error mass fluctuates by $\sqrt{\kappa d/n}$, and dividing by κ gives $\sqrt{d/(\kappa n)}$; both cases are $\sqrt{(1-\kappa+\gamma)d/(\kappa n)}$ up to a factor two. The Clopper–Pearson bounds adapt to both scales automatically, so the IUT needs no tuning.

For interval menus, the next result matches this dependence up to logarithmic factors at any fixed coverage floor.

Theorem 1 (Interval menus at any fixed coverage floor)**.** *Let $\mathcal{A}_K$ contain all unions of at most K intervals of the score line, $\kappa \in (0, 1)$, $a \in (0, 1/8]$, $\alpha \in [2a, 1-2a]$ and $0 < \delta < \beta < 1$. There are $c, c', C, K_0 > 0$ depending only on (a, δ, β) such that for $K \geq K_0$ and $\gamma \leq c' \min\{\kappa, 1-\kappa\}/\log K$,*

$$\frac{c\,K(1-\kappa)}{\kappa\gamma^2 \log K} \;\leq\; N(\mathcal{P}_\gamma(\mathcal{A}_K), \mathcal{H}) \;\leq\; N_{\mathrm{IUT}}(\mathcal{P}_\gamma(\mathcal{A}_K)) \;\leq\; \frac{C\{K(1-\kappa)\log(e/(\kappa\gamma)) + 1\}}{\kappa\gamma^2}.$$

The lower bound holds even for audits that know the score distribution.

Thus, for every fixed $\kappa \in (0, 1)$ in the stated small-gap regime, the IUT is minimax optimal up to logarithmic factors. Moreover, Proposition 2(i) gives $O(K/\gamma^2)$ at fixed κ, leaving only the $\log K$ gap to the lower bound. *Proof idea.* The lower bound embeds $K - 1$ consecutive score groups, each split into two halves whose losses are $m \pm d_b$ with a random sign and a random magnitude d_b. Under the null and alternative priors the magnitudes match their first $\log K$ even moments but differ in their mean by the best polynomial approximation error of $\sqrt{u}$, so the two mixtures are indistinguishable until $n \gtrsim K/(\gamma^2 \log K)$ (Theorem 5 below). To reach an arbitrary floor, a *padding* region carries the rest of the score mass: for $\kappa \geq 1/2$ it has risk α and is accepted by every good rule, so only a fraction $2(1-\kappa)$ of the data is informative and the risk gap shrinks by $(1-\kappa)/\kappa$; for $\kappa < 1/2$ it has risk

one and is never accepted. The delicate step is robust infeasibility: under the alternative prior every measurable set of coverage at least $\kappa - \gamma$ must have risk at least $\alpha + \gamma$, not only the menu's rules. Because the conditional risk is piecewise constant, the least-risk sets fill the score line in increasing order of risk, and trimming them changes the risk by at most $2\gamma D/\kappa$ (Appendix B.4).

Reading the law. The menu cost $K(1-\kappa)/(\kappa\gamma^2)$ is a cost of *power*, not of validity: the level proof pays nothing for $|\mathcal{A}|$, yet a richer menu gives the vendor more ways to place the abstained region where the audit sample happens to contain few errors. Part (i) holds for any uniform-convergence certificate with radius of order $\sqrt{d/n}$; the endpoint DKW certificate (radius of order $K/\sqrt{n}$) and policywise Bonferroni (of order $\sqrt{K \log M/n}$ over M cells) have weaker guarantees, and part (iii) uses the Clopper–Pearson bounds' adaptation to small proportions. The logarithmic gaps and constants remain open; at $\kappa = 1/2$, with the constants of our proof, the new lower bound exceeds the elementary $\Omega(\sqrt{K}/\gamma^2)$ obstruction from Proposition 3 only for small γ and large K. The next two sections explain *why* the cost is linear in K and when an audit can avoid it.

4 A TESTING–ESTIMATION GAP FOR EXISTENTIAL AUDITS

Sign families. Fix $p \in (0,1)$, $0 < \gamma < \min\{p, 1-p\}$ and $(\kappa, \alpha) = (1/2, p-\gamma)$. The reference P_0 has $S \sim \mathrm{Unif}[0,1]$ and $L \mid S \sim \mathrm{Bern}(p)$; it is infeasible. For positive weights w in the simplex Δ_{B-1}, partition $[0,1]$ into groups G_ℓ of mass w_ℓ, each split into equal halves. Each $\theta \in \{-1,+1\}^B$ accepts one half per group, giving 2^B rules A_θ. A feasible component keeps the score law and sets

$$Q_\theta(L = 1 \mid S = x) = p - \gamma s_\theta(x), \qquad s_\theta(x) = 1 \text{ on } A_\theta,\ -1 \text{ off } A_\theta. \tag{5}$$

Every component has coverage $1/2$, accepted risk $p-\gamma$, and the same distance from P_0; mixing them uniformly gives P_0 exactly. All positive weight profiles share the same single-rule testing problems, incidence structure, VC dimension B and growth function (Propositions 6–7); the families differ only in how their rules overlap. For the plateau profile with k heavy groups, $w_\ell = (1 - B^{-2})/k$ for $\ell \le k$ and the remaining mass B^{-2} spread over the other groups, the effective dimension is $d_{\text{eff}} = \|w\|_2^{-2} = k\{1 + o(1)\}$.

Theorem 2 (Testing scale)**.** *Fix $p \in (0,1)$ and $0 < \delta < \beta < 1$. If $\gamma = \gamma_B < \min\{p, 1-p\}$ and $B^{-2} \le \gamma_B/16$, then as $B \to \infty$, uniformly in $k \le B$, the full-null sample size at the reference is $N(P_0, \mathcal{H}_{B,k}; \delta, \beta) = \Theta(\sqrt{k}/\gamma_B^2)$.*

The lower bound is Ingster's second-moment method. For a prior Π over feasible components with $\mathbb{E}_\Pi Q = P_0$, $1 + \chi^2(\mathbb{E}_\Pi Q^n \| P_0^n) = \mathbb{E}_{Q,Q' \sim \Pi}[1 + \Xi(Q,Q')]^n$ with $\Xi = \int (dQ/dP_0)(dQ'/dP_0) dP_0 - 1$ (Lemma 3); here $\Xi(\theta, \theta') = \lambda \sum_\ell w_\ell \theta_\ell \theta'_\ell$ with $\lambda = \gamma^2/[p(1-p)]$, which gives:

Proposition 3 (Weight-profile obstruction)**.** *Every level-δ audit of a sign family has $\mathbb{E}_{P_0^n} \phi \le \delta + \frac{1}{2}\{\exp(n^2\lambda^2 \|w\|_2^2/2) - 1\}^{1/2}$, hence $N_w = \Omega(\sqrt{d_{\text{eff}}}/\gamma^2)$.*

For the upper bound, every law of the full null—whether it changes scores, labels or both—is $\Omega(\gamma)$-far from P_0 after pushing forward to the $(4k+2)$ half/label cells of the heavy groups, so identity testing needs $O(\sqrt{k}/\gamma^2)$ samples (Valiant and Valiant, 2017) (Appendix D).

Theorem 3 (The estimation cost of EFC audits)**.** *Fix $p \in (0,1)$ and $0 < \delta < \beta < 1$. There are $\gamma_0, c > 0$ depending only on (p, δ, β), and for each split below a $C > 0$, such that on the plateaus of Theorem 2, whenever $\gamma \le \gamma_0$ and $B^{-2} \le \gamma/16$: (i) every level-δ EFC audit has power below β at P_0 for all $n \le ck/\gamma^2$; (ii) the exact IUT with any fixed split $\zeta_c + \zeta_r = \delta$, $\zeta_c, \zeta_r \in (0, 1/2]$, has power at least β for all $n \ge Ck/\gamma^2$. Hence $N_{\text{IUT}} = \Theta(k/\gamma^2)$ and $N_{\text{IUT}}/N(P_0, \mathcal{H}_{B,k}) = \Theta(\sqrt{k})$.*

Mechanism. In each heavy group choose the half containing more *correct* audit observations. Conditional on the correct observations and the error count, this rule accepts each error independently with probability $1/2$, while its accepted correct count exceeds half of all correct observations by G with $\mathbb{E}G \gtrsim \min\{n, \sqrt{nk}\}$. For $n \ll k/\gamma^2$ this gain overwhelms the population gap $n\gamma$ and keeps the empirical coverage at least $1/2$: a sample-dependent rule is empirically feasible, so no EFC audit rejects (Appendix C.2). Choosing the half with fewer errors would not control coverage. The optimal audit instead aggregates quadratic evidence whose fluctuation is $\sqrt{k}/n$ against a signal γ^2. The EFC bound does not rely on the reference coverage sitting exactly at the floor: with q-ary groups and reference coverage $m/q > \kappa$ it persists (Proposition 4, Appendix C.3).

Remark 1 (The barrier on interval menus). Order the $2K$ halves of the uniform sign family ($k = B = K$) along the score line, each group's halves adjacent. Unions of at most K intervals of these cells contain every sign rule, so Theorem 3(i) applies: for $\gamma \le \gamma_0$, every valid EFC audit of this interval menu needs $\Omega(K/\gamma^2)$ samples at P_0, without a logarithm, and the IUT needs $O(K/\gamma^2)$. Theorem 4 below gives a valid non-EFC audit with $O(\sqrt{K}/\gamma^2)$ samples at the same law for $\gamma \le (2K)^{-1/4}$.

5 When can the gap be escaped?

The optimal audit of Theorem 2 knows P_0; the IUT knows nothing about it. Which missing knowledge forces the linear cost? Validity is always required over the *unrestricted* null; only the power requirement is placed on a class $\mathcal{P}$. Let $C_1, \dots, C_M$ be score cells frozen independently of the audit sample and $\mathcal{A}$ any family of unions of cells.

Unknown score distribution, constant risk: a reference-free testing-scale audit. Let $\mathcal{P}_{\mathrm{c}}(\alpha, \gamma)$ contain every law under which L is independent of S and $P(L = 1) \ge \alpha + \gamma$, with an arbitrary score marginal.

Theorem 4 (Reference-free audit at the testing scale)**.** *Fix $\kappa \in (0, 1]$ and $0 < \delta < \beta < 1$. An audit that uses only cell labels and losses, and knows neither the score marginal nor $P(L = 1)$, has level δ over $\mathcal{H}(\mathcal{A})$ and power at least β uniformly over $\mathcal{P}_{\mathrm{c}}(\alpha, \gamma)$ for every $n \ge C_{\delta,\beta}\{\sqrt{M}/(\kappa\gamma)^2 + M^{2/3}/(\kappa\gamma)^{4/3}\}$. On the uniform sign family ($M = 2k$, $\kappa = 1/2$) with $\gamma \le (2k)^{-1/4}$ this is $\Theta(\sqrt{k}/\gamma^2)$, while every valid EFC audit needs $\Omega(k/\gamma^2)$ at $P_0 \in \mathcal{P}_{\mathrm{c}}$.*

The audit rejects when the empirical loss rate exceeds $\alpha + \gamma/2$ *and* an independence tester for cell and loss (Diakonikolas and Kane, 2016) finds no dependence. The key fact is that *feasibility forces dependence*: if Q has a witness A and $e = Q(L = 1) > \alpha$, every product law R on cells and labels has $\mathrm{TV}(Q, R) \ge (e - \alpha)\kappa/3$, because a product law assigns the witness the overall loss rate (Lemma 2). The audit is valid even for the family of all 2^M unions of cells, so it matches the testing scale only where that scale is $\sqrt{M}/\gamma^2$; for small menus the IUT is better. It abstains whenever its independence test detects score–loss dependence. Section 6 separately reports strong score–loss dependence and held-out family-level hiding on frozen model outputs. The next result shows that, over a rich heterogeneous class, no valid audit can uniformly retain the square-root rate.

Unknown heterogeneous risk: the linear cost can be necessary. Consider the uniform sign family with B groups, $\kappa = 1/2$ and the uniform score marginal, *known* to the auditor. For $\eta \in [0, 1]^{2B}$ let P_η have loss probability $\eta_{b,h}$ on half h of group b; the best risk is $r^*(\eta) = B^{-1} \sum_b \min\{\eta_{b,+}, \eta_{b,-}\}$. Let $\mathcal{P}_{\mathrm{het}}(\alpha, \gamma) = \{P_\eta : \eta \in [a, 1 - a]^{2B},\ r^*(\eta) \ge \alpha + \gamma\}$.

Theorem 5 (The price of unknown heterogeneity)**.** *Fix $a \in (0, 1/8]$, $\alpha \in [2a, 1 - 2a]$ and $0 < \delta < \beta < 1$. There are $c_1, c_2, C, B_0 > 0$ depending only on (a, δ, β) such that for $B \ge B_0$ and $\gamma \le c_1/\log B$, $c_2 B/(\gamma^2 \log B) \le N(\mathcal{P}_{\mathrm{het}}, \mathcal{H}) \le N_{\mathrm{IUT}}(\mathcal{P}_{\mathrm{het}}) \le CB/\gamma^2$, whereas $N(P, \mathcal{H}) = O(\sqrt{B}/\gamma^2)$ for every single $P \in \mathcal{P}_{\mathrm{het}}$.*

The null and alternative priors set $\eta_{b,\pm} = m \pm s_b d_b$ with a uniform sign s_b and a magnitude $d_b \in [0, D]$ from ν_0 or ν_1. These match $\mathbb{E}d^{2j}$ for $j \le L \asymp \log B$ but differ in $\mathbb{E}d$, and hence in r^*, by twice the best degree-L approximation error of $\sqrt{u}$ on $[0, D^2]$, of order D/L (Bernstein, 1914; Lepski et al., 1999; Cai and Low, 2011). Under Poissonization the random signs make each group's χ^2 series depend only on the even moments, so all terms up to order L cancel (Appendix B.5). Unlike ℓ_1-functional estimation (Jiao et al., 2018), the priors must keep the null draws *existentially* feasible. Together, Theorems 4 and 5 show that validity alone does not force the linear cost, while uncertainty about *where* the classifier errs can force it on a rich heterogeneous class.

6 Experiments

6.1 The abstention-odds law

We audit menus of $K \in \{1, 2, 4, 8, 16\}$ intervals over $M = 64$ equal-mass score cells at floors $\kappa \in \{.5, .7, .8, .9, .95\}$ (and .3 in Appendix H.4). The alternative has constant risk $\alpha + \gamma$ ($\alpha = .10$,

$\gamma = .04$), so every rule has the same risk and the law lies in $\mathcal{P}_\gamma$ at every floor; any increase of the sample size with K is therefore a pure cost of menu richness. For each (K, κ) the exact IUT audits 250 samples at each of 15 log-spaced sizes between 150 and 15,846, and we record the 80%-power crossing n_{80} and the *menu cost* $n_{80}(K) - n_{80}(1)$.

Figure 1 illustrates the coverage-floor dependence in Proposition 2(iii) and Theorem 1. The raw menu cost varies by almost an order of magnitude across floors. The right panel additionally divides by the theorem-motivated localized-VC logarithmic factor discussed in Appendix B.3; that factor is not used in the proofs, and the resulting variation is substantially reduced until the menu saturates. The flattening itself is the localization mechanism at work: only rules that differ on the small abstained region compete. The IUT's level is not at issue here (it is exact); boundary checks are reported below.

6.2 Sign families: the IUT against the information envelope

We fix $B = 256$ (2^{256} rules), $p = .25$, $\kappa = .5$, $k \in \{1, 4, 16, 64, 256\}$ and $\gamma \in \{.04, .02\}$. The exact dynamic program audits 1,500 samples per grid point. The comparator is the Neyman–Pearson curve for P_0^n against the uniform feasible mixture; its population version bounds the power of every valid audit, so it is an information envelope, not an audit. At $\gamma = .04$ the IUT's crossing rises from 2,342 to 32,796 samples as k grows from 1 to 256, and the envelope's from 1,001 to 7,384; halving the gap multiplies all crossings by 3.91–4.17. Fitting $\gamma^2 n_{80} \approx a_0 + a_1\sqrt{k} + a_2 k$ gives $a_2 \approx 0.13$ for the IUT at both gaps and $a_2 \approx 0$ for the envelope, as Theorems 2–3 predict. Ten boundary-null checks, one per (k, γ) with 20,000 audits each, give rejection rates of 0–4.7%. Across the simulated k values the gap to the envelope is a factor of 2.2–4.4. In the separate fixed-degree ($L = 2$) heterogeneity-envelope experiment, the heterogeneous envelope overtakes the known-reference one only at $B \approx 2.6 \times 10^5$ (Appendix B.7).

6.3 Auditing the auditors: VLM-judge reliability

Automated VLM judges are increasingly used to automate evaluation, and a reliability claim is naturally existential: *some* confidence gate should retain enough evaluations while keeping serious judge errors below a target. We use the public score-token log-probabilities and human scores released by Kumar et al. (2026). For each item let $L = \mathbf{1}\{|\hat{y} - y_{\text{human}}| > 1\}$, matching the source study's ± 1 scoring criterion; a non-1–5 parse is counted as a serious error, with a drop-invalid sensitivity analysis in the Supplementary Note. From the five score-token probabilities we predeclare three label-free confidence channels—maximum probability, negative entropy and top-two margin—and cross each with target coverages .65 and .90, giving six gates at contract floor $\kappa = .60$.

For each judge we use ten deterministic three-way resplits. Thresholds are frozen from calibration scores only; audit-A selects the midpoint of its exact fixed-marginal hiding window $I = [\alpha_H, \alpha_{\text{ind}})$, and untouched audit-B tests transfer. Thus $W_{\text{hide}} > 0$ means no individual gate is feasible at some target even though feasible alternatives can convexify to the empirical reference. Table 2 reports the result. On MLLM-Judge, every one of the 60 split-specific windows is positive and 28/30 midpoint contracts transfer. To test whether this is merely noisy single-annotator supervision, we repeat the same LLaVA-Critic protocol on Polaris, where the released target averages at least three annotators and the same judge is much more accurate (Kumar et al., 2026). All 20 windows remain positive, but the median held-out width falls from 4.97 to 0.33 percentage points and midpoint transfer falls from 10/10 to 3/10. On a representative Polaris audit-B split, singleton families have zero window, 60% of two-gate and 90% of three-gate subfamilies hide, and every subfamily of size at least four hides (Supplementary Note). The mechanism therefore persists under cleaner supervision, while its magnitude and split stability are population dependent. These are finite-target mechanism diagnostics, not an instance-specific minimax verification of Theorem 5.

6.4 Frozen classifier certificates and family geometry

We also treat frozen classifier outputs as populations: CIFAR-10/100 ResNet-18, ImageNetV2 ViT-B/16 and CMU-MOSEI MLP outputs (Krizhevsky, 2009; Recht et al., 2019; Bagher Zadeh et al., 2018). Calibration fixes score cells and the audit half supplies labels. At $\kappa = .9$, the exact IUT has no observed boundary rejection above 1.8% and is the most powerful certificate compared except for one $n = 500$ CIFAR-100 point within Monte Carlo error; Bonferroni needs several times more data,

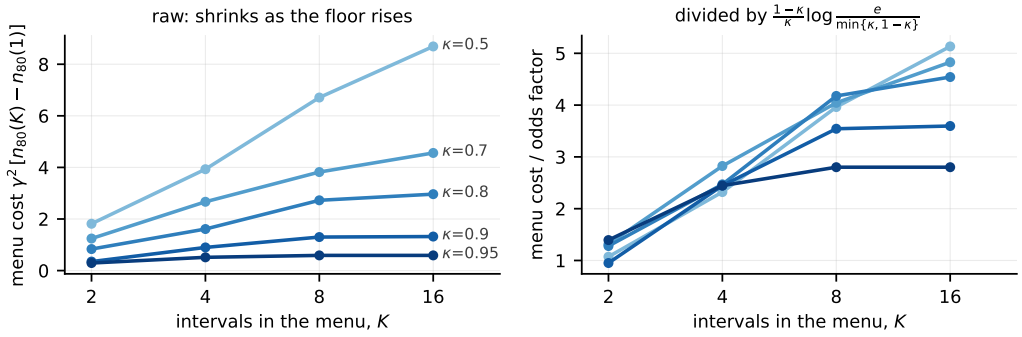

Figure 1: **Menu richness costs K times the abstention odds.** Left: menu cost of the exact IUT, scaled by γ^2, for five coverage floors. At $K = 4$ it falls 7.7-fold from $\kappa = .5$ to $\kappa = .95$. Right: the same costs divided by $\frac{1-\kappa}{\kappa}\log\frac{e}{\min\{\kappa,1-\kappa\}}$ lie within 2.32–2.82 at $K = 4$ and within 0.95–1.39 at $K = 2$. At $\kappa \geq .9$ the curves flatten for $K \geq 8$ because the abstained region spans only $M(1-\kappa) \leq 6.4$ cells, so extra intervals do not enlarge the realized admissible rule set in this finite benchmark. Monte Carlo crossings from 250 audits per point.

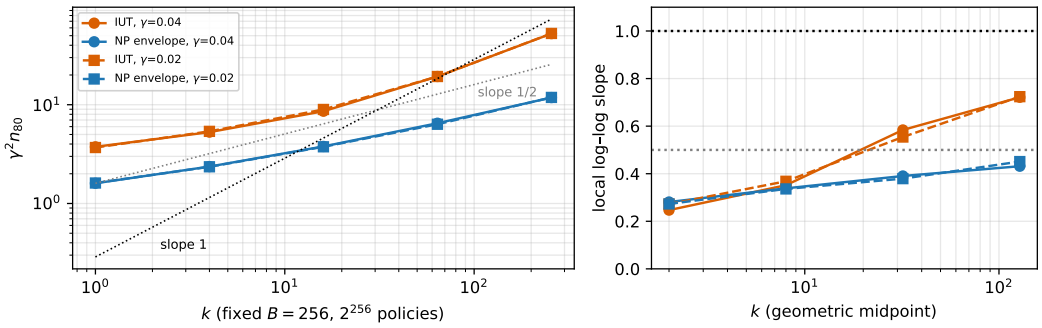

Figure 2: **The IUT and the information envelope have different scales** ($B = 256$). Left: $\gamma^2 n_{80}$ at two gaps; the curves nearly coincide, as γ^{-2} scaling predicts. Right: local log–log slopes in k. The IUT's slope rises from 0.25 to 0.72 towards the proven exponent 1, the envelope's from 0.28 to 0.43 towards $1/2$; the range is pre-asymptotic. Crossings are in Table 7 (Appendix H.1).

structured DKW is vacuous here, and a score-agnostic total-loss certificate fails when total error is small relative to $1 - \kappa$ (Table 8, Appendix H.2).

The constant-risk premise of Theorem 4 is strongly contradicted: a cell–loss independence diagnostic rejects in every one of 200 audits of size 500 on each population. More directly, a separate six-policy, three-way held-out geometry experiment freezes gates from calibration scores only. All 60 audit-A and all 60 audit-B windows are positive; 47/60 midpoint contracts selected on A transfer to untouched B. Singleton subfamilies have zero hiding, 96.7% of pairs hide, and every subfamily of size at least three hides (Supplementary Note, Section S5). Thus the same feasible-mixture mechanism appears on standard classifier families, without claiming that these populations instantiate Theorem 5's asymptotic hard class.

Finally, as a valid non-EFC reference-free competitor we implement cross-fit universal inference (Wasserman et al., 2020). It has no observed boundary rejections, but at $\Delta = .05$ its first 80%-power grid point is $n = 10{,}000$–$50{,}000$, versus 500–2,500 for the IUT; at $\Delta = .02$ it does not reach 80% power by $50{,}000$ (Table 12, Appendix H.5). This evaluates one practical non-EFC construction, not all non-EFC audits.

Table 2: **Auditing the auditors.** Held-out exact-hiding windows for six predeclared confidence gates at $\kappa = .60$ and serious judge error $L = \mathbf{1}\{|\hat{y} - y_{\text{human}}| > 1\}$. Policies are frozen on a disjoint calibration third. "pos. A/B" counts positive windows over audit-A and audit-B; "A→B" asks whether the midpoint contract selected on A remains hidden on untouched B. Widths are median audit-B windows in percentage points.

Target	judge	pos. A/B	A→B	med. W_B
MLLM-Judge	LLaVA-Critic-7B	20/20	10/10	4.97
MLLM-Judge	Gemini 2.5 Flash	20/20	10/10	5.91
MLLM-Judge	Phi-4 reasoning vision	20/20	8/10	4.83
Polaris (multi-annotator)	LLaVA-Critic-7B	20/20	3/10	0.33

7 Related work and limitations

Risk control, Learn-then-Test and Pareto testing certify selected rules (Bates et al., 2021; Angelopoulos et al., 2024; 2025; Laufer-Goldshtein et al., 2023); model falsification, subgroup auditing and PAC verification address adjacent goals (Müller et al., 2025; Cherian and Candès, 2024; Mutreja and Shafer, 2026). Intersection–union validity is classical (Berger, 1982; Berger and Hsu, 1996; Lehmann and Romano, 2005); for a fixed number of normal means the IUT is the likelihood-ratio test and non-monotone tests can be uniformly more powerful (Lehmann, 1952; Berger, 1989), at the price of non-monotone rejection regions that can yield counterintuitive inferences (Perlman and Wu, 1999). Moment-inequality and intersection-bound tests (Andrews and Soares, 2010; Chernozhukov et al., 2013; Romano et al., 2014; Bugni et al., 2015) ask whether some parameter satisfies a set of inequalities; with many inequalities, as in reality checks (White, 2000; Hansen, 2005), the cost is logarithmic in their number (Chernozhukov et al., 2019) because that null is an intersection, whereas ours is a union. Inference on winners corrects the selection bias of an empirical optimum (Andrews et al., 2024), the bias that EFC certificates do not correct; sequential tests of whether some arm lies below a threshold address a related existential question (Kaufmann et al., 2018); confidence sequences give anytime versions of the IUT (Howard et al., 2021) (Appendix G). Universal inference gives reference-free composite tests (Wasserman et al., 2020); recent work on testable hypotheses and safe anytime-valid inference provides broader characterizations and tools for composite testing (Larsson et al., 2026b; Ramdas et al., 2023). In property testing, unions of d intervals are testable from $O(\sqrt{d})$ samples when the input distribution is known but need $\Omega(d/\log d)$ distribution-free (Kearns and Ron, 2000; Balcan et al., 2012; Blais et al., 2021); related interval and flat structure also appears in identity testing of structured distributions (Diakonikolas et al., 2015), and optimal risk can be estimated faster than learned (Kong and Valiant, 2018). Minimax L_1-distance estimation and tolerant distribution testing exhibit related linear-over-log sample-complexity phenomena (Jiao et al., 2018; Canonne et al., 2022). Recent work studies calibrated uncertainty for VLM judges and releases the score-token features used in our auditing application (Kumar et al., 2026). Our null is a union of convex sets of laws with ratio-type constraints, losses are noisy, and an unknown input distribution alone does not force the linear scale (Theorem 4), whereas unknown conditional risk can force it on the heterogeneous class of Theorem 5. The Supplementary Note relates the one-observation geometry to convex-set testing (Goldenshluger et al., 2015).

Limitations. Losses are binary, the contract has one coverage floor, and data are i.i.d.; distribution shift is outside our scope. For each fixed coverage floor, the interval-menu bounds are worst-case and match up to logarithms in the stated small-gap, large-K regime; the lower bound is asymptotic in K; the sharp testing–estimation spectrum and the constant-risk result are proved on sign families; an $\Omega(d)$ bound for every VC-d family is open. Theorem 4 is an existence result with an untuned tester. The natural-family and VLM-judge studies are finite-target mechanism diagnostics, not instance-specific minimax lower bounds for Theorem 5; the six VLM gates are illustrative and use released benchmark outputs rather than production traffic. The non-EFC experiment evaluates one universal-inference construction rather than all possible non-EFC audits. Nonrejection certifies no rule.

AI USE STATEMENT

Generative AI was used to improve the clarity and presentation of manuscript claims, assist with software implementation and debugging, and provide critical feedback on claims and arguments during manuscript development. AI-generated suggestions were treated as editorial and technical assistance rather than as evidence or independent validation. The authors are fully responsible for the final claims, proofs, code, citations, analyses, and submission.

ETHICS STATEMENT

The empirical studies reuse frozen benchmark model outputs and released human annotations and collect no new human-subject data. The work concerns statistical auditability and does not introduce a clinical decision system or an automated evaluator for deployment. Population interpretations require a sampling model appropriate to the intended setting.

REPRODUCIBILITY STATEMENT

The supplement contains this manuscript's figures, analysis scripts, exact rational certificates, frozen internal inputs, and derived VLM-audit summaries; the public third-party VLM feature files are referenced by source and are not redistributed. The analyses for Section 5 and Figure 3 are provided with their tests. The release README maps claims to code and distinguishes recomputed results from retained outputs. Appendix H and the Supplementary Note give protocols and limitations.

REFERENCES

Donald W. K. Andrews and Gustavo Soares. Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection. *Econometrica*, 78(1):119–157, 2010.

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on Winners. *The Quarterly Journal of Economics*, 139(1):305–358, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal Risk Control. In *International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi:10.1214/24-AOAS1998.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2236–2246, 2018. doi:10.18653/v1/P18-1208.

Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *The Annals of Statistics*, 47(4):1893–1927, 2019. doi:10.1214/18-AOS1729.

Maria-Florina Balcan, Eric Blais, Avrim Blum, and Liu Yang. Active Property Testing. In *IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 21–30, 2012.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-Free, Risk-Controlling Prediction Sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi:10.1145/3478535.

Roger L. Berger and Jason C. Hsu. Bioequivalence Trials, Intersection–Union Tests and Equivalence Confidence Sets. *Statistical Science*, 11(4):283–319, 1996.

Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300, 1982. doi:10.1080/00401706.1982.10487790.

Roger L. Berger. Uniformly More Powerful Tests for Hypotheses Concerning Linear Inequalities and Normal Means. *Journal of the American Statistical Association*, 84(405):192–199, 1989.

Serge Bernstein. Sur la meilleure approximation de $|x|$ par des polynômes de degrés donnés. *Acta Mathematica*, 37:1–57, 1914.

Eric Blais, Renato Ferreira Pinto Jr., and Nathaniel Harms. VC Dimension and Distribution-Free Sample-Based Testing. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2021.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of Classification: A Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Federico A. Bugni, Ivan A. Canay, and Xiaoxia Shi. Specification Tests for Partially Identified Models Defined by Moment Inequalities. *Journal of Econometrics*, 185(1):259–282, 2015.

T. Tony Cai and Mark G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011. doi:10.1214/10-AOS849.

Clément L. Canonne, Ayush Jain, Gautam Kamath, and Jerry Li. The Price of Tolerance in Distribution Testing. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, volume 178 of *PMLR*, pp. 573–624, 2022.

John J. Cherian and Emmanuel J. Candès. Statistical Inference for Fairness Auditing. *Journal of Machine Learning Research*, 25(149):1–49, 2024.

Victor Chernozhukov, Sokbae Lee, and Adam M. Rosen. Intersection Bounds: Estimation and Inference. *Econometrica*, 81(2):667–737, 2013.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Inference on Causal and Structural Parameters Using Many Moment Inequalities. *The Review of Economic Studies*, 86(5):1867–1900, 2019.

C. K. Chow. On Optimum Recognition Error and Reject Tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

C. J. Clopper and E. S. Pearson. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413, 1934. doi:10.1093/biomet/26.4.404.

Ilias Diakonikolas and Daniel M. Kane. A New Approach for Testing Properties of Discrete Distributions. In *IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 685–694, 2016.

Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing Identity of Structured Distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1841–1854, 2015. doi:10.1137/1.9781611973730.123.

Ran El-Yaniv and Yair Wiener. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Evarist Giné and Vladimir Koltchinskii. Concentration Inequalities and Asymptotic Results for Ratio Type Empirical Processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

Alexander Goldenshluger, Anatoli Juditsky, and Arkadi Nemirovski. Hypothesis testing by convex optimization. *Electronic Journal of Statistics*, 9(2):1645–1712, 2015. doi:10.1214/15-EJS1054.

Peter Reinhard Hansen. A Test for Superior Predictive Ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, Nonparametric, Nonasymptotic Confidence Sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Yu. I. Ingster. Minimax Testing of Nonparametric Hypotheses on a Distribution Density in the L_p Metrics. *Theory of Probability and Its Applications*, 31(2):333–337, 1987. doi:10.1137/1131042.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax Estimation of the L_1 Distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, 2018. doi:10.1109/TIT.2018.2846245.

R. Kaas and J. M. Buhrman. Mean, median and mode in binomial distributions. *Statistica Neerlandica*, 34(1):13–18, 1980.

Emilie Kaufmann, Wouter M. Koolen, and Aurélien Garivier. Sequential Test for the Lowest Mean: From Thompson to Murphy Sampling. In *Advances in Neural Information Processing Systems 31*, 2018.

Michael Kearns and Dana Ron. Testing Problems with Sublearning Sample Complexity. *Journal of Computer and System Sciences*, 61(3):428–456, 2000.

Weihao Kong and Gregory Valiant. Estimating Learnability in the Sublinear Data Regime. In *Advances in Neural Information Processing Systems 31*, 2018.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. MSc thesis, Department of Computer Science, University of Toronto, 2009.

Divake Kumar, Sina Tayebati, Devashri Naik, Ranganath Krishnan, and Amit Ranjan Trivedi. VLM Judges Can Rank but Cannot Score: Task-Dependent Uncertainty in Multimodal Evaluation. arXiv preprint arXiv:2604.25235, 2026.

Martin Larsson, Johannes Ruf, and Aaditya Ramdas. A Complete Characterization of Testable Hypotheses. *arXiv preprint arXiv:2601.05217*, 2026.

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi Jaakkola. Efficiently Controlling Multiple Risks with Pareto Testing. In *International Conference on Learning Representations*, 2023.

Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2nd edition, 2000. doi:10.1007/978-1-4612-1166-2.

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005. doi:10.1007/0-387-27605-X.

E. L. Lehmann. Testing Multiparameter Hypotheses. *The Annals of Mathematical Statistics*, 23(4):541–552, 1952.

Oleg Lepski, Arkady Nemirovski, and Vladimir Spokoiny. On estimation of the L_r norm of a regression function. *Probability Theory and Related Fields*, 113(2):221–253, 1999.

Manuel M. Müller, Yuetian Luo, and Rina Foygel Barber. Are All Models Wrong? Fundamental Limits in Distribution-Free Empirical Model Falsification. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *PMLR*, pp. 4271–4308, 2025.

Pascal Massart. The Tight Constant in the Dvoretzky–Kiefer–Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi:10.1214/aop/1176990746.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2nd edition, 2018.

Saachi Mutreja and Jonathan Shafer. PAC Verification of Statistical Algorithms. *Theory of Computing*, 22(5):1–40, 2026. doi:10.4086/toc.2026.v022a005.

Michael D. Perlman and Lang Wu. The Emperor's New Tests. *Statistical Science*, 14(4):355–369, 1999.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*, 38(4):576–601, 2023. doi:10.1214/23-STS894.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 5389–5400, 2019.

Joseph P. Romano, Azeem M. Shaikh, and Michael Wolf. A Practical Two-Step Method for Testing Moment Inequalities. *Econometrica*, 82(5):1979–2002, 2014.

Gregory Valiant and Paul Valiant. An Automatic Inequality Prover and Instance Optimal Identity Testing. *SIAM Journal on Computing*, 46(1):429–455, 2017. doi:10.1137/151002526.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020. doi:10.1073/pnas.1922664117.

Halbert White. A Reality Check for Data Snooping. *Econometrica*, 68(5):1097–1126, 2000.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

A ROADMAP AND NOTATION

Local notation. A few symbols are reused locally with meanings fixed by context: L is the loss, but the polynomial degree in Theorem 5; M is the number of cells, but the error count in the proofs of Theorem 3 and Proposition 4; m is a prior midpoint in Theorem 5 and the number of selected parts in the q-ary construction; K counts intervals, and $K(Q, Q')$ is the overlap kernel of Lemma 3; λ is $\gamma^2/[p(1-p)]$ in Section 4 and a Poisson mean in Appendix B.5.

Where the main results are proved. This appendix contains everything the main theorems use. EFC certificates, Theorem 4, Proposition 2 (including the localization in part (iii)), Theorem 1 (Appendix B.4) and Theorem 5: Appendix B. Theorem 3, the Rademacher formula and Proposition 5: Appendix C. Theorem 2 and Propositions 6, 3 and 7: Appendix D. Lemma 3: Appendix E. Proposition 1 and the dynamic programs: Appendix G. The witness power ceiling (Corollary 3) and the collision bound (Theorem 6): Appendix F. Experimental protocols and all tables, including the abstention-odds simulation and the universal-inference audit: Appendix H. A separate Supplementary Note in the supplementary material collects one-observation geometry (exact projections, Hall–TV formulas, the eight-atom pairwise example and natural hiding windows) that the main theorems do not use.

Symbol	Meaning	Sample-computable?
P_0	fixed infeasible reference / audit alternative	reference-dependent
Q	feasible population law under the existential null	no
$\mathcal{A}$	predeclared family of acceptance policies A	yes
$\mathcal{H}(\mathcal{A})$	laws for which at least one $A \in \mathcal{A}$ meets the contract	implicit
κ, α	coverage floor and conditional-risk target	yes
$R_n = \mathbb{E}_{Q\sim\Pi}[Q^n]$	product mixture induced by a prior over feasible laws	only with witness
$\Xi(Q, Q')$	likelihood-ratio overlap kernel	only with components
d_{eff}	participation-ratio summary inside the matched sign construction	diagnostic proxy
$N(P_0, \mathcal{H}; \delta, \beta)$	samples needed for level δ and power β	no
L_{IUT}	reference-free lower risk frontier for an existential contract	yes
EFC	audit that never rejects while some policy is empirically feasible	property
$\mathcal{P}_{\mathrm{c}}, \mathcal{P}_{\mathrm{het}}$	constant-risk and heterogeneous-risk alternative classes	power classes

B EFC CERTIFICATES, REFERENCE-FREE TESTING, AND UNKNOWN HETEROGENEITY

B.1 WHICH CERTIFICATES ARE EFC

Lemma 1 (Clopper–Pearson bounds bracket the empirical proportion)**.** *Let $X \sim \mathrm{Bin}(t, q)$ with observed value x. For $\zeta \le 1/2$ the one-sided Clopper–Pearson upper bound at error ζ is at least x/t and the lower bound is at most x/t.*

Proof. When the mean tq is an integer, it is a median of $\mathrm{Bin}(t, q)$ (Kaas and Buhrman, 1980). At $q = x/t$ we therefore have $P_q(X \le x) \ge 1/2 \ge \zeta$ and $P_q(X \ge x) \ge 1/2 \ge \zeta$. The upper bound is the largest q with $P_q(X \le x) \ge \zeta$, so it is at least x/t; the lower bound is symmetric. □

Let A be empirically feasible: $T_A \ge \kappa n$ and $E_A \le \alpha T_A$.

- *Exact IUT and policywise Bonferroni.* By Lemma 1, $U_A \ge T_A/n \ge \kappa$, so A passes the coverage screen, and $L_A \le E_A/T_A \le \alpha$. Hence $L_{\mathrm{IUT}} \le \alpha$ and the audit does not reject. Bonferroni replaces ζ by $\zeta/|\mathcal{A}| \le 1/2$ and is covered by the same argument.
- *DKW and Rademacher certificates* (Appendices G.5–G.6). Their screen $c_n(A) \ge \kappa - \epsilon$ is passed by A, and their lower bound $[m_n(A) - \epsilon]_+ / \min\{1, c_n(A) + \epsilon\}$ is at most $m_n(A)/c_n(A) = E_A/T_A \le \alpha$.
- *Total-loss certificate.* Writing $\hat{e}$ for the empirical loss rate, $\hat{e} \le (E_A + n - T_A)/n \le 1 - (1-\alpha)T_A/n \le 1 - \kappa(1-\alpha)$. Hence $(\hat{e} + \kappa - 1)/\kappa \le \alpha$; its lower confidence version is smaller by Lemma 1, so it does not reject.

Part (i) of Theorem 3 therefore applies to every certificate compared in Section 6.

B.2 Proof of Theorem 4

Throughout, laws are identified with their pushforwards to $[M] \times \{0,1\}$ (cell index and loss). This loses nothing for the null: every $A \in \mathcal{A}$ is a union of cells, so the contract depends on Q only through this pushforward, and the pushforward of a feasible law is feasible for the induced family. Products on the original space push forward to products.

Lemma 2 (Feasibility forces dependence)**.** *Let Q be feasible with witness A, and let $e = Q(L = 1) > \alpha$. Then $\mathrm{TV}(Q, R) \geq (e-\alpha)Q(A)/3$ for every product law $R = \nu \otimes \mathrm{Bern}(q)$ on $[M] \times \{0,1\}$.*

Proof. Let $t = \mathrm{TV}(Q, R)$. If $t \geq e$ or $t \geq Q(A)$ the claim holds because $(e-\alpha)Q(A)/3 < \min\{e, Q(A)\}$. Otherwise, comparing the events $\{L = 1\}$, $\{S \in A\}$ and $\{S \in A, L = 1\}$ gives $q \geq e - t$, $\nu(A) \geq Q(A) - t$ and $q\nu(A) \leq Q(A, L = 1) + t \leq \alpha Q(A) + t$. Thus $\alpha Q(A) + t \geq (e-t)(Q(A)-t) \geq eQ(A) - t(e + Q(A)) \geq eQ(A) - 2t$, so $3t \geq (e-\alpha)Q(A)$. □

Independence tester. For distributions on $[M] \times [2]$, Diakonikolas and Kane (2016) give a tester that, from $O(\sqrt{M}/\epsilon^2 + M^{2/3}/\epsilon^{4/3})$ samples, accepts every product law and rejects every law ϵ-far in total variation from all product laws, each with probability at least $2/3$ (their ℓ_1 parameter is 2ϵ). Taking the majority of $O(\log(1/\epsilon'))$ independent runs reduces both error probabilities to ϵ'. Denote the amplified output by $\Psi \in \{0,1\}$, with $\Psi = 1$ meaning "dependent".

The audit. Let $\epsilon' = \min\{\delta, (1-\beta)/2\}$ and $\epsilon = \kappa\gamma/12$. Reject feasibility if and only if $\hat{e} \geq \alpha + \gamma/2$ and $\Psi = 0$, where $\hat{e}$ is the empirical loss rate and Ψ uses the independence tester at distance ϵ and error ϵ'. Neither the score marginal nor $P(L = 1)$ is an input. Let n_0 be the larger of $8\log(1/\epsilon')/\gamma^2$ and the amplified tester's sample size; for $n < n_0$ the audit does not reject, and for $n \geq n_0$ it applies the rule above to the first n_0 observations for the tester.

Level over the unrestricted null. Let $Q \in \mathcal{H}(\mathcal{A})$ and $e = Q(L = 1)$. If $e \leq \alpha + \gamma/4$, Hoeffding's inequality gives $Q(\hat{e} \geq \alpha + \gamma/2) \leq \exp(-n\gamma^2/8) \leq \epsilon'$ once $n \geq 8\log(1/\epsilon')/\gamma^2$. If $e > \alpha + \gamma/4$, Lemma 2 and $Q(A) \geq \kappa$ make Q at least $\kappa\gamma/12$-far from every product law, so $Q(\Psi = 0) \leq \epsilon'$. In both cases the rejection probability is at most $\epsilon' \leq \delta$. No assumption beyond i.i.d. sampling and frozen cells is used.

Power. Let $P \in \mathcal{P}_{\mathrm{c}}(\alpha, \gamma)$. Its pushforward is a product law, so $P(\Psi = 1) \leq \epsilon'$, and $P(\hat{e} < \alpha + \gamma/2) \leq \exp(-n\gamma^2/2) \leq \epsilon'$. The power is at least $1 - 2\epsilon' \geq \beta$. Collecting sample sizes, and using $M \geq 2$ and $\kappa \leq 1$ to absorb the $\log(1/\epsilon')/\gamma^2$ term, gives the displayed bound with a constant depending on (δ, β).

Rates on the sign family. With $M = 2k$ and $\kappa = 1/2$, $M^{2/3}\gamma^{-4/3} \leq M^{1/2}\gamma^{-2}$ if and only if $\gamma \leq M^{-1/4}$, so $N(\mathcal{P}_{\mathrm{c}}, \mathcal{H}) = O(\sqrt{k}/\gamma^2)$ in that regime. The uniform reference P_0 with $p = \alpha + \gamma$ lies in $\mathcal{P}_{\mathrm{c}}$, so $N(\mathcal{P}_{\mathrm{c}}, \mathcal{H}) \geq N(P_0, \mathcal{H}) = \Omega(\sqrt{k}/\gamma^2)$ by Proposition 3 with uniform weights, and Theorem 3(i) gives the EFC lower bound at the same law. Conversely, the IUT upper bound of Appendix C.2 extends to every $P \in \mathcal{P}_{\mathrm{c}}$ on the sign family: all policies have risk $P(L = 1) \geq \alpha + \gamma$, and on the uniform-concentration event over the 2^k policies every policy passing the coverage screen has $T_A \geq \kappa n/2$ and $E_A - \alpha T_A \geq \gamma T_A - C\sqrt{nk}$, so $N_{\mathrm{IUT}}(\mathcal{P}_{\mathrm{c}}) = O(k/(\kappa\gamma)^2)$; this justifies the entry in the third row of Table 1. Whether the $M^{2/3}\gamma^{-4/3}$ term is necessary for $\gamma > M^{-1/4}$ is left open.

B.3 Proof of Proposition 2

Two inversions. For $X \sim \mathrm{Bin}(t, q)$ observed at x, Hoeffding's inequality inverted at error ζ gives that the one-sided Clopper–Pearson upper bound is at most $x/t + \sqrt{\log(1/\zeta)/(2t)}$ and the lower bound is at least $x/t - \sqrt{\log(1/\zeta)/(2t)}$.

Uniform deviations. The classes $\mathcal{F}_1 = \{(s,l) \mapsto \mathbf{1}_A(s) : A \in \mathcal{A}\}$ and $\mathcal{F}_2 = \{(s,l) \mapsto l\,\mathbf{1}_A(s) : A \in \mathcal{A}\}$ have VC dimension at most d; a set shattered by $\mathcal{F}_2$ consists of points with $l = 1$ whose scores are shattered by $\mathcal{A}$. The VC bound for Boolean classes (Vershynin, 2018, Chapter 8) gives $\mathbb{E}\sup_{f\in\mathcal{F}_i} |P_n f - Pf| \le C_0\sqrt{d/n}$, and the bounded-differences inequality gives, with probability at least $1-(1-\beta)/2$ for both classes simultaneously,

$$\sup_{A\in\mathcal{A}} |c_n(A) - c(A)| \le \epsilon_n, \qquad \sup_{A\in\mathcal{A}} |m_n(A) - m(A)| \le \epsilon_n,$$
$$\epsilon_n = C_0\sqrt{d/n} + \sqrt{\log(8/(1-\beta))/(2n)},$$

where $c_n(A) = T_A/n$, $m(A) = P(S \in A, L = 1)$ and $m_n(A) = E_A/n$.

Part (i). Let $P \in \mathcal{P}_\gamma(\mathcal{A})$ and work on the event above. Suppose A passes the coverage screen, $U_A \ge \kappa$. By the first inversion $c_n(A) \ge \kappa - \sqrt{\log(2/\delta)/(2n)}$, so $c(A) \ge \kappa - \sqrt{\log(2/\delta)/(2n)} - \epsilon_n \ge \kappa - \gamma$ provided $\sqrt{\log(2/\delta)/(2n)} + \epsilon_n \le \gamma$. The definition of $\mathcal{P}_\gamma$ then gives $m(A) \ge (\alpha+\gamma)c(A)$, and since $c(A) \ge \kappa - \gamma \ge \kappa/2$,

$$\frac{E_A}{T_A} \ge \frac{m(A) - \epsilon_n}{c(A) + \epsilon_n} \ge \alpha + \gamma - \frac{2\epsilon_n}{c(A)} \ge \alpha + \gamma - \frac{4\epsilon_n}{\kappa}.$$

If also $\epsilon_n \le \kappa/4$, then $T_A \ge n\kappa/4$, and the second inversion gives $L_A \ge \alpha + \gamma - 4\epsilon_n/\kappa - \sqrt{2\log(2/\delta)/(n\kappa)}$. All three conditions, and $L_A > \alpha$, hold once $n \ge C\{d + \log(2/\delta) + \log(4/(1-\beta))\}/(\kappa\gamma)^2$ for an absolute C. Then every screened policy has $L_A > \alpha$, and if no policy is screened $L_{\mathrm{IUT}} = +\infty$; either way the IUT rejects. Its power is at least $1-(1-\beta)/2 \ge \beta$.

Part (ii). Choose a score law under which $c(A_0) \ge \kappa$ and let P have constant risk $\alpha+\gamma$. Every policy has risk $\alpha+\gamma$, so $P \in \mathcal{P}_\gamma(\mathcal{A})$. Let Q equal P except that the risk inside A_0 is α; then A_0 witnesses feasibility of Q. Since $\mathrm{KL}(P\|Q) = c(A_0)\,\mathrm{kl}(\alpha+\gamma\|\alpha) \le \gamma^2/\{\alpha(1-\alpha)\}$, Pinsker's inequality gives $\mathrm{TV}(P^n, Q^n) \le \{n\gamma^2/(2\alpha(1-\alpha))\}^{1/2}$. A level-$\delta$ audit with power β at P needs $\mathrm{TV}(P^n, Q^n) \ge \beta - \delta$, hence $n \ge 2(\beta-\delta)^2\alpha(1-\alpha)/\gamma^2$.

Part (iii): localization to the smaller side of the floor. We use Vapnik's relative-deviation inequalities: for a class $\mathcal{F}$ of $\{0,1\}$-valued functions with growth function $S_\mathcal{F}$, with probability at least $1-\eta$, $\sup_f (Pf - P_n f)/\sqrt{Pf} \le \epsilon$ and $\sup_f (P_n f - Pf)/\sqrt{P_n f} \le \epsilon$, where $\epsilon = 2\{(\log S_\mathcal{F}(2n) + \log(8/\eta))/n\}^{1/2}$ (Boucheron et al., 2005, Theorem 5.1). Solving the second inequality for $P_n f$ gives, for every $f \in \mathcal{F}$,

$$|P_n f - Pf| \le \epsilon\sqrt{Pf} + \epsilon^2. \tag{6}$$

Apply this to $\mathcal{F} = \{\mathbf{1}_A\} \cup \{\mathbf{1}_{A^c}\} \cup \{l\mathbf{1}_A\} \cup \{l\mathbf{1}_{A^c}\}$ ($A \in \mathcal{A}$). Each of the four classes has VC dimension at most d, so $\log S_\mathcal{F}(2n) \le \log 4 + d\log(2en/d)$ for $n \ge d$. Take $\eta = (1-\beta)/4$, and let Hoeffding's inequality give $|\hat{e} - e| \le \{\log(8/(1-\beta))/(2n)\}^{1/2}$ for the overall loss rate with probability at least $1-(1-\beta)/4$. On the intersection Ω of these events, which has probability at least $1-(1-\beta)/2$, every $A \in \mathcal{A}$ satisfies, with $c = c(A)$, $m = m(A)$ and c_n, m_n their empirical versions,

$$|c_n - c| \le \epsilon\sqrt{\min\{c, 1-c\}} + \epsilon^2, \tag{D1}$$
$$|m_n - m| \le \min\left\{\epsilon\sqrt{c} + \epsilon^2,\ |\hat{e} - e| + \epsilon\sqrt{1-c} + \epsilon^2\right\}, \tag{D2}$$

using $c_n - c = -(c_n(A^c) - c(A^c))$, $m_n - m = (\hat{e} - e) - (m_n(A^c) - m(A^c))$ and $m(A^c) \le 1 - c$.

Clopper–Pearson widths. Let $X \sim \mathrm{Bin}(t,q)$ be observed at x, $\hat{q} = x/t$. Since $\partial_q \mathrm{kl}(\hat{q}, q) = (q-\hat{q})/\{q(1-q)\}$, we have $\mathrm{kl}(\hat{q}, q) \ge (q-\hat{q})^2/\{2\max_{s\in[\hat{q},q]} s(1-s)\} \ge (q-\hat{q})^2/(2\min\{q, 1-\hat{q}\})$ for $q \ge \hat{q}$. The Chernoff bound $P_q(X \le x) \le \exp\{-t\,\mathrm{kl}(\hat{q}, q)\}$ therefore gives, for the upper bound U at error ζ,

$$U - \hat{q} \le \{2\min\{U, 1-\hat{q}\}\log(1/\zeta)/t\}^{1/2}, \tag{7}$$

and the lower bound satisfies $\hat{q} - L \le \{\log(1/\zeta)/(2t)\}^{1/2}$ as before.

Write $\ell = \log(2/\delta)$, $\sigma = \min\{\kappa, 1-\kappa+\gamma\}$ and $u = \epsilon^2 + \ell/n$. Suppose

$$u \le c_0\gamma^2\kappa/(1-\kappa+\gamma), \qquad |\hat{e} - e| \le \gamma/48, \qquad n \ge 8\ell/(\kappa\gamma^2), \qquad c_0 = 96^{-2}. \tag{8}$$

Because $\kappa^2 \le \kappa \le 1-\kappa+\gamma$ when $\sigma=\kappa$, the first condition implies $u \le c_0\gamma^2/\sigma$, and since $\sigma \ge \gamma$ (as $\gamma \le \kappa/2$) also $u \le c_0\gamma$.

Step 1: the screen admits only policies of coverage at least $\kappa-\gamma$. Suppose $U_A \ge \kappa$ but $t := \kappa - c(A) > \gamma$. By (7) applied to the n coverage trials, $\kappa - c_n \le \{2v\ell/n\}^{1/2}$ with $v = \min\{U_A, 1-c_n\}$, and by (D1), $c_n - c \le \epsilon\sqrt{\min\{c,1-c\}} + \epsilon^2$. Now $\min\{c,1-c\} \le \min\{\kappa, 1-\kappa+t\} \le (t/\gamma)\sigma$ because $t \ge \gamma$. Using $\epsilon\sqrt{x} \le (x+\epsilon^2)/2$ in (D1) and $\sqrt{2U\ell/n} \le U/2 + \ell/n$ in (7) gives $1 - c_n \le 2(1-\kappa+t) + 2u$ and $U_A \le 4\kappa + 4u$, so $v \le 4(t/\gamma)\sigma + 4u$. Adding the two displays,

$$t \le \sqrt{8(t\sigma/\gamma)u} + \sqrt{8}\,u + \sqrt{(t\sigma/\gamma)u} + u \le (\sqrt{8}+1)\{\sqrt{c_0} + c_0\}\,t < t,$$

using $(t\sigma/\gamma)u \le c_0 t\gamma \le c_0 t^2$ and $u \le c_0\gamma \le c_0 t$. This contradiction proves the claim.

Step 2: every screened policy is certified infeasible. Let $U_A \ge \kappa$, so $c = c(A) \ge \kappa - \gamma \ge \kappa/2$, and $P \in \mathcal{P}_\gamma(\mathcal{A})$ gives $m \ge (\alpha+\gamma)c$. With $\varrho = \alpha + \gamma/2$ we have $m - \varrho c \ge \gamma c/2$, and by (D1)–(D2)

$$|(m_n - \varrho c_n) - (m - \varrho c)| \le \Delta := \min\left\{2\epsilon\sqrt{c} + 2\epsilon^2,\ |\hat{e} - e| + 2\epsilon\sqrt{1-c} + 2\epsilon^2\right\}.$$

If $\kappa \le 1/2$, then $\kappa/(1-\kappa+\gamma) \le 2\kappa$, so $\epsilon^2 \le 2c_0\gamma^2\kappa \le \gamma^2 c/256$ and the first term gives $\Delta \le \gamma c/8 + \gamma c/8$. If $\kappa > 1/2$, then $c \ge 1/4$, $1-c \le 1-\kappa+\gamma$ and $\epsilon^2 \le c_0\gamma^2/(1-\kappa+\gamma)$, so the second term gives $\Delta \le \gamma/48 + \gamma/48 + \gamma/48 \le \gamma c/4$. In both cases $m_n - \varrho c_n \ge \gamma c/4 > 0$; also $c_n \ge c/2 \ge \kappa/4$ by (D1). Hence $E_A/T_A \ge \alpha + \gamma/2$ with $T_A \ge n\kappa/4$, and the Clopper–Pearson lower bound satisfies $L_A \ge \alpha + \gamma/2 - \{2\ell/(n\kappa)\}^{1/2} > \alpha$ by the last condition in (8).

Sample size. On Ω, every screened policy has $L_A > \alpha$, so the IUT rejects; its power is at least $1-(1-\beta)/2 \ge \beta$. The conditions (8) hold whenever $n \ge C\{(1-\kappa+\gamma)\,[d\log(2en/d) + \log(8/(1-\beta))] + \log(2/\delta) + \log(8/(1-\beta))\}/(\kappa\gamma^2)$ for an absolute constant C, using $(1-\kappa+\gamma) \le 1$ and $\kappa \le 1$ for the terms without d. The left side minus the right side is increasing in n, which proves part (iii).

Remark on the logarithm. The factor $\log(2en/d)$ comes from the growth function. Localized VC bounds for sets of small probability (Giné and Koltchinskii, 2006) suggest that it can be replaced by $\log(e/\min\{\kappa, 1-\kappa\})$; the experiments of Section 6.1 are consistent with this form, but we do not use it.

B.4 Proof of Theorem 1

Upper bound. Unions of at most K intervals have VC dimension $d = 2K$. Part (iii) of Proposition 2 with $1-\kappa+\gamma \le 2(1-\kappa)$ (as $\gamma \le 1-\kappa$) gives power at least β once $n \ge C\{K(1-\kappa)\log(2en/K) + 1\}/(\kappa\gamma^2)$, with C depending on (δ, β). This condition holds for $n \ge C'\{K(1-\kappa)\log(e/(\kappa\gamma)) + 1\}/(\kappa\gamma^2)$ because the logarithm of such an n/K is at most a constant times $\log(e/(\kappa\gamma))$.

Lower bound: a padded construction. Fix $B = K-1$ and let γ_4, D, m, L, ν_0, ν_1 be as in Steps 1–2 of the proof of Theorem 5 with this B; the conditions imposed below imply those of that theorem. Scores are uniform on $[0,1]$. A *construction region* R of mass π is split into B consecutive groups of mass π/B, each split into a left and a right half, with loss probability $\eta_{b,\pm} = m \pm s_b d_b$ on the two halves as in Step 2. Its complement, the *padding*, has a fixed conditional loss probability η_{pad} that is the same under both priors. The inner best risk is $r^*_{\mathrm{in}}(\eta) = m - B^{-1}\sum_b d_b$, and the lower halves (those with risk $m - d_b$) have total mass $\pi/2$. We distinguish two cases.

Case $\kappa \ge 1/2$. Take $\pi = 2(1-\kappa)$, padding $F = [0, 1-\pi)$ with $\eta_{\mathrm{pad}} = \alpha$, and $R = [1-\pi, 1]$. The set W formed by F and the lower halves has coverage $1 - \pi + \pi/2 = \kappa$, is a union of at most $B + 1 = K$ intervals, and has risk

$$r(W) = \alpha + \tfrac{\pi}{2\kappa}\{r^*_{\mathrm{in}}(\eta) - \alpha\} = \alpha + \tfrac{1-\kappa}{\kappa}\{r^*_{\mathrm{in}}(\eta) - \alpha\}.$$

Under the null prior, on the event $r^*_{\mathrm{in}} \le \alpha$ of Step 2, W witnesses feasibility for $\mathcal{A}_K$. Under the alternative prior, on the event $r^*_{\mathrm{in}} \ge \alpha + \gamma_4$, consider any measurable set. The conditional risk is constant on each half and on F. Every upper half has risk $m + d_b \ge m$, and every point of W has risk at most m, because $m - \alpha = \mathbb{E}_{\nu_0} d - g/4 \ge 3g/4 > 0$. Hence, as in the proof for $\kappa = 1/2$, the least risk at coverage exactly κ is $r^* := r(W) \ge \alpha + (1-\kappa)\gamma_4/\kappa$, larger coverage cannot lower it, and at coverage $\kappa - s$ with $s \le \kappa/2$ the optimal set removes mass s of risk at most m from W,

so its risk is at least $r^* - s(m - r^*)/(\kappa - s) \geq r^* - 2sD/\kappa$, because $m - r^* \leq m - \alpha \leq D$. Put $\gamma = (1 - \kappa)\gamma_4/(2\kappa)$. Taking $s = \gamma$ and using $D \leq 1/16 \leq \kappa/2$, every measurable set of coverage at least $\kappa - \gamma$ has risk at least $\alpha + 2\gamma - \gamma = \alpha + \gamma$: the alternative draws lie in $\mathcal{P}_\gamma(\mathcal{A}_K)$.

Case $\kappa < 1/2$. Take $\pi = 2\kappa$, $R = [0, \pi)$ and padding $H = [\pi, 1]$ with $\eta_{\mathrm{pad}} = 1$. The lower halves form a set W of coverage κ that is a union of at most B intervals, with risk r^*_{in}; under the null prior it witnesses feasibility. Under the alternative, the points of least risk are the lower halves, then the upper halves (risk at least m), then H (risk $1 > m + D$). As before, larger coverage cannot lower the risk and coverage $\kappa - s$ lowers it by at most $2sD/\kappa$. Put $\gamma = \gamma_4/2$ and require $D \leq \kappa/2$; then every set of coverage at least $\kappa - \gamma$ has risk at least $\alpha + \gamma_4 - \gamma = \alpha + \gamma$.

Indistinguishability. Poissonize as in Step 3 of the proof of Theorem 5, with mean n'. Counts in the padding are independent of counts in R and have the same law under both priors, so the total variation between the two mixtures equals that between the mixtures of the R-counts. These are independent Poisson counts with means $\lambda\eta_{b,h}$ and $\lambda(1 - \eta_{b,h})$ in each half, now with $\lambda = \pi n'/(2B)$, so Steps 4–6 apply verbatim with $\rho = \pi n'/\{Bm(1 - m)\}$. The condition $\tau = \rho D^2 \leq (2L + 2)/e^3$ holds whenever $\pi n' \leq c\, B/(\gamma_4^2 \log B)$. In the case $\kappa \geq 1/2$ this reads $n' \leq c\, B(1 - \kappa)/(8\kappa^2\gamma^2 \log B)$, and in the case $\kappa < 1/2$ it reads $n' \leq c\, B/(8\kappa\gamma^2 \log B)$. Both are at least $c''K(1 - \kappa)/(\kappa\gamma^2 \log K)$ for $K \geq 3$. Hence no valid audit with fewer samples has power β over $\mathcal{P}_\gamma(\mathcal{A}_K)$.

Conditions. Theorem 5 needs $\gamma_4 \leq c_1/\log B$, the case $\kappa < 1/2$ needs $D \leq \kappa/2$, and $D = \gamma_4/E_L \leq \gamma_4 L/c_*$ with $L = \lceil \log B \rceil$. In the case $\kappa \geq 1/2$, $\gamma_4 = 2\kappa\gamma/(1 - \kappa) \leq 2\gamma/(1 - \kappa)$; in the case $\kappa < 1/2$, $\gamma_4 = 2\gamma$. All conditions therefore hold when $\gamma \leq c' \min\{\kappa, 1 - \kappa\}/\log K$ for a suitable c' depending only on a, and $K \geq K_0$. The auditor may know the uniform score marginal, the cells and the padding: only the location and size of the risk heterogeneity inside R are random. At $\kappa = 1/2$ the padding is empty and the construction is the one of Theorem 5 with $B = K - 1$ consecutive groups. □

Numerical checks. The artifact test `test_v61.py` verifies the trimming inequality of both cases on 20,000 random padded instances with $\kappa \in (.05, .95)$, the width bound (7) for every outcome with $t \leq 10^4$ trials and $\zeta \in \{.025, .05, .25\}$, and Lemma 4 against constrained numerical optimization.

Consequences. The ratio of the bounds in parts (i) and (ii) of Proposition 2 is $O(d/(\kappa^2\alpha(1 - \alpha)))$ up to (δ, β) terms. On the uniform sign family, the alternative prior draws of Theorem 5 with $r^*(\eta) \geq \alpha + \gamma$ lie in $\mathcal{P}_{\gamma/2}$ by the trimming argument of Theorem 1 (case $\kappa = 1/2$), which uses $d_b \leq D \leq 1/16$; the whole class $\mathcal{P}_{\mathrm{het}}$ need not, since trimming a high-risk lower half can lower the risk by more. Hence Theorem 5's construction also shows that the upper end is attained up to $\log d$ there.

B.5 Proof of Theorem 5

Constants c, C below depend only on (a, δ, β) unless stated otherwise.

Upper bounds. Let $P_\eta \in \mathcal{P}_{\mathrm{het}}$. Every policy has coverage exactly $1/2$ and risk $r_\eta(A_\theta) \geq \alpha + \gamma$. Hoeffding's inequality and a union bound over the 2^B policies give, with probability at least $1 - \eta_0$, $T_A \geq n/2 - C\sqrt{n(B + \log(1/\eta_0))}$ and $E_A - \alpha T_A \geq n\gamma/2 - C\sqrt{n(B + \log(1/\eta_0))}$ for every A. As in the upper-bound part of Appendix C.2, for $n \geq CB/\gamma^2$ every Clopper–Pearson risk lower bound exceeds α, so the IUT rejects with probability at least β uniformly over the class. For a single known $P \in \mathcal{P}_{\mathrm{het}}$ and a feasible Q with witness A, the function $f = \mathbf{1}_A(L - \alpha)$ has oscillation one, $\mathbb{E}_Q f \leq 0$ and $\mathbb{E}_P f = \{r_\eta(A) - \alpha\}/2 \geq \gamma/2$. So every feasible law is $\gamma/2$-far after the $4B$-cell pushforward, and identity testing gives $N(P, \mathcal{H}) = O(\sqrt{B}/\gamma^2)$.

Step 1: moment-matched magnitudes. For $L \geq 1$ let $E_L = \inf_g \sup_{u \in [0,1]} |\sqrt{u} - g(u)|$ over polynomials g of degree at most L. Substituting $u = x^2$, E_L equals the error of best approximation of $|x|$ on $[-1, 1]$ by polynomials of degree $2L$, and Bernstein's theorem gives $2LE_L \to \beta_* \approx 0.2802$ (Bernstein, 1914). Since each $E_L > 0$, $c_* = \inf_{L \geq 1} LE_L > 0$; numerically $LE_L \geq 1/8$ for $L \leq 12$ (Appendix B.6). By duality (Cai and Low, 2011, Lemma 1) there are probability measures π_0, π_1 on

$[0,1]$ with $\int u^j d\pi_0 = \int u^j d\pi_1$ for $j = 1, \dots, L$ and $\int \sqrt{u}\, d\pi_0 - \int \sqrt{u}\, d\pi_1 = 2E_L$. For $D > 0$ let ν_i be the law of $D\sqrt{U}$ with $U \sim \pi_i$. Then ν_0, ν_1 are supported on $[0, D]$, match $\mathbb{E}d^{2j}$ for $j \le L$, and $g := \mathbb{E}_{\nu_0} d - \mathbb{E}_{\nu_1} d = 2DE_L$.

Step 2: priors and their contracts. Set $L = \lceil \log B \rceil$, $D = \gamma / E_L$ and $m = \alpha + \mathbb{E}_{\nu_0} d - g/4$. Under prior i, draw independently for each group a sign s_b uniform on $\{\pm 1\}$ and $d_b \sim \nu_i$, and set $\eta_{b,\pm} = m \pm s_b d_b$. Since $D \le \gamma L / c_*$, the condition $\gamma \le c_1 / \log B$ with c_1 small gives $D \le a/2$; together with $m \in [\alpha - D/4, \alpha + D]$ and $\alpha \in [2a, 1-2a]$ this keeps every $\eta_{b,h}$ in $[a, 1-a]$. Because $r^*(\eta) = m - B^{-1} \sum_b d_b$, Hoeffding's inequality gives

$$\begin{aligned} \Pr_0\{r^*(\eta) > \alpha\} + \Pr_1\{r^*(\eta) < \alpha + \gamma\} &\le 2\exp\{-Bg^2/(8D^2)\} \\ &= 2\exp(-BE_L^2/2) \le 2\exp\{-Bc_*^2/(2L^2)\}, \end{aligned}$$

using $\mathbb{E}_0 r^* = \alpha - g/4$, $\mathbb{E}_1 r^* = \alpha + 3g/4 = \alpha + \gamma + g/4$ and $\gamma = g/2$. Under prior 0 the policy choosing the lower half in each group has coverage $1/2$, so on the event $r^* \le \alpha$ the law P_η is feasible; under prior 1 and $r^* \ge \alpha + \gamma$ it lies in $\mathcal{P}_{\mathrm{het}}$.

Step 3: Poissonization. If a test with n samples has level δ and power β, so does one with any $n'' \ge n$ samples; take $n'' = \max\{n, n_1\}$ with $n_1 = \lceil 4\log(4/(\beta - \delta)) \rceil$. Draw $N \sim \mathrm{Poi}(2n'')$ observations and apply the test to the first n'' if $N \ge n''$, otherwise accept. The level is unchanged and the power drops by at most $\Pr(N < n'') \le e^{-n''/4} \le (\beta - \delta)/4$. Under Poissonization with mean $n' = 2n''$, the numbers of errors and correct observations in the $2B$ halves are independent Poisson variables with means $\lambda\eta_{b,h}$ and $\lambda(1 - \eta_{b,h})$, $\lambda = n'/(2B)$; groups are independent under both priors.

Step 4: one group. Let F_0 be the law of one group's four counts at $\eta \equiv m$, and let μ_i be its law under prior i. For Poisson likelihood ratios, $\mathbb{E}_{F_0}[(dP_x/dF_0)(dP_{x'}/dF_0)] = \exp\{\lambda x x'/m + \lambda x x'/(1-m)\}$ for error-rate shifts x, x' of one half. The two halves have shifts $\pm sd$, so the group-level kernel is $\exp(\rho s s' dd')$ with $\rho = 2\lambda/\{m(1-m)\} = n'/\{Bm(1-m)\}$. Averaging over independent signs gives $\cosh(\rho dd')$ and

$$\int \frac{(d\mu_1 - d\mu_0)^2}{dF_0} = \sum_{j \ge 0} \frac{\rho^{2j}}{(2j)!} \Big(\mathbb{E}_{\nu_1} d^{2j} - \mathbb{E}_{\nu_0} d^{2j} \Big)^2 \le \sum_{j > L} \frac{(\rho D^2)^{2j}}{(2j)!}.$$

With $\tau = \rho D^2$, $\sum_{j>L} \tau^{2j}/(2j)! \le e^\tau \{e\tau/(2L+2)\}^{2L+2}$. If $\tau \le (2L+2)/e^3$ this is at most $\exp\{-(2 - e^{-3})(2L+2)\} \le e^{-3.9(L+1)}$. By Cauchy–Schwarz, $\mathrm{TV}(\mu_0, \mu_1) \le \frac{1}{2} e^{-1.95(L+1)}$.

Step 5: all groups. With $H^2(P, Q) = 1 - \int \sqrt{dP\, dQ}$ we have $H^2 \le \mathrm{TV}$, $H^2(\mu_0^{\otimes B}, \mu_1^{\otimes B}) \le BH^2(\mu_0, \mu_1)$ and $\mathrm{TV} \le \sqrt{2H^2}$. Hence $\mathrm{TV}(\mu_0^{\otimes B}, \mu_1^{\otimes B}) \le \sqrt{B}\, e^{-0.975(L+1)} \le B^{-0.47}$ for $L = \lceil \log B \rceil$.

Step 6: conclusion. Conditioning the priors on the events of Step 2 changes the mixtures by at most their probabilities. A test of level δ over $\mathcal{H}$ and power β over $\mathcal{P}_{\mathrm{het}}$ therefore satisfies

$$\beta - \delta - \tfrac{\beta - \delta}{4} \le B^{-0.47} + 2\exp\{-Bc_*^2/(2L^2)\},$$

which fails for $B \ge B_0$ whenever the condition $\tau \le (2L+2)/e^3$ holds. That condition reads $n' \le (2L+2)Bm(1-m)E_L^2/(e^3\gamma^2)$; with $m(1-m) \ge a/2$, $E_L \ge c_*/L$ and $L \le 2\log B$ it holds whenever $n'' \le c_2 B/(\gamma^2 \log B)$ for a suitable c_2. Since $n'' = \max\{n, n_1\}$ and $n_1 \le c_2 B/(\gamma^2 \log B)$ for $B \ge B_0$, no valid audit with $n \le c_2 B/(\gamma^2 \log B)$ samples has power β. □

Remarks. The auditor in this lower bound knows the family, the cells and the exact uniform score marginal; only the location and size of the risk heterogeneity are random. The unknown quantity is a sum of B nonsmooth terms $\min\{\eta_{b,+}, \eta_{b,-}\}$, which is why the argument parallels minimax estimation of ℓ_1-type functionals (Lepski et al., 1999; Cai and Low, 2011; Wu and Yang, 2016).

B.6 Numerical checks for Theorem 5

The script `heterogeneity_checks.py` (i) computes E_L by a linear program on a grid; (ii) solves the dual linear program for π_0, π_1 and confirms that the optimal moment-matched gap equals $2E_L$; and (iii) compares the χ^2 series of Step 4 with direct summation over Poisson counts for one group (errors only, which has the same structure with $\rho = 2\lambda/m$).

Table 3: Numerical checks behind Theorem 5. Left: best-approximation error of $\sqrt{u}$ on $[0, 1]$ (the monomial LP is reliable in this range; values increase towards Bernstein's constant $\beta_* \approx 0.2802$). Middle: optimal moment-matched gap from the dual LP versus $2E_L$. Right: largest relative error between the closed-form χ^2 series and direct Poisson summation over two parameter settings.

L	E_L	$2LE_L$	L	gap	$2E_L$	L	max rel. error
1	0.12500	0.2500	2	0.13524	0.13524	2	1.4e-13
2	0.06762	0.2705	3	0.09186	0.09186	3	1.8e-11
3	0.04593	0.2756	4	0.06938	0.06938	4	1.6e-09
4	0.03469	0.2775					
6	0.02325	0.2790					
8	0.01747	0.2795					
10	0.01399	0.2797					

B.7 Protocol for the envelope comparison

Setting. Uniform sign family with $B \in \{64, 256, 1024, 4096, 16384, 65536, 262144\}$, uniform score marginal, $\kappa = 1/2$, $\alpha = .2$ and Poissonized sample size n (independent Poisson counts of errors and correct observations in each half).

Heterogeneity envelope. The LP of Appendix B.6 with $L = 2$ gives ν_0 with atoms $\{0.0569, 0.2\}$ and weights $\{0.904, 0.096\}$, and ν_1 with atoms $\{0, 0.1554\}$ and weights $\{0.719, 0.281\}$ ($D = .2$). Then $g = 2DE_2$, $\gamma = g/2 = 0.01352$ and $m = \alpha + \mathbb{E}_{\nu_0} d - g/4 = 0.2639$. The statistic is the exact log-likelihood ratio between the two product mixtures, obtained by summing over atoms and signs in each group. For a valid audit ϕ, $\mathbb{E}_{\mathrm{mix}_0}\phi \le \delta + \Pr_0(G_0^c) = \delta'$ and $\mathbb{E}_{\mathrm{mix}_1|G_1}\phi \le \mathrm{NP}_{\delta'}/\Pr_1(G_1)$. We estimate the conditioning probabilities from 2×10^5 multinomial draws, calibrate the NP threshold at level δ' on independent draws from mix_0, and evaluate power on independent draws from mix_1.

Known-reference envelope. The reference has constant risk $p = \alpha + \gamma$. The null mixture averages the hiding components of Equation (5) with the same γ, so every component is feasible and no conditioning is needed. The NP threshold is calibrated at level $\delta = .05$.

Replicates and crossings. We use 2,000 replicates for calibration and 2,000 for power per grid point when $B \le 4096$, and 1,000 of each when $B \ge 16384$, with sample-size grids around the crossing. The $B = 262{,}144$ runs used a second machine with different NumPy/SciPy versions; rerunning $B = 256$ there changed the crossings by at most 1.2%, within Monte Carlo error. The n_{80} values are log-linear interpolations of Monte Carlo power estimates; the two envelopes use independent seeds. The script `heterogeneity_envelope.py` produces all runs and `merge_het_envelope.py` merges them and draws Figure 3.

C Estimation complexity, IUT inefficiency, and reference adaptation

C.1 Exact distribution-dependent Rademacher complexity

The profiles of Theorem 2 are distinguished by the population Rademacher complexity, $\mathfrak{R}_n \asymp \sum_b \min\{w_b, \sqrt{w_b/n}\}$, of order $\sqrt{k/n}$ on a plateau; but it does so at the *estimation* scale k/γ^2, whereas the optimal audit works at the *testing* scale $\sqrt{k}/\gamma^2$, as in instance-optimal identity testing (Valiant and Valiant, 2017; Balakrishnan and Wasserman, 2019). For $\mathcal{F}_w = \{\mathbf{1}_{A_\theta} : \theta \in \{-1, 1\}^B\}$ define $\mathfrak{R}_n(\mathcal{F}_w) = \mathbb{E}_{S,\sigma} \sup_\theta n^{-1} \sum_{i=1}^n \sigma_i \mathbf{1}_{A_\theta}(S_i)$, with independent Rademacher signs σ_i. Let

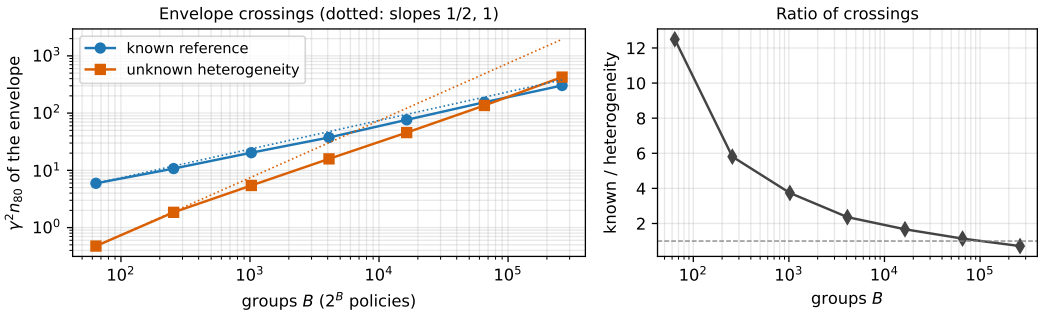

Figure 3: **Unknown heterogeneity changes the exponent and eventually dominates.** Left: 80%-power crossings of the two envelopes, scaled by γ^2. The known-reference crossing grows with fitted exponent 0.48 (theory: $1/2$); the heterogeneity crossing grows with exponent 0.80, approaching $1 - 1/(2L+2) = 0.83$ for fixed $L = 2$ (the theorem lets $L \asymp \log B$). Right: their ratio falls from 12.5 at $B = 64$ to 0.71 at $B = 262{,}144$. Because the constant-risk reference belongs to $\mathcal{P}_{\text{het}}$, the larger crossing is the operative lower bound on $N(\mathcal{P}_{\text{het}}, \mathcal{H})$. Envelopes bound valid audits; they are not audits.

Table 4: Envelope crossings behind Figure 3. Because the constant-risk reference belongs to $\mathcal{P}_{\text{het}}$, the larger of the two crossings is the operative lower bound on $N(\mathcal{P}_{\text{het}}, \mathcal{H})$. The last two columns report how often a prior draw leaves its hypothesis; this probability is charged in the heterogeneity envelope.

B	known n_{80}	heterogeneity n_{80}	ratio	$\Pr_0(G_0^c)$	$\Pr_1(G_1^c)$
64	32,484	2,600	12.49	0.123	0.238
256	58,713	10,108	5.81	0.003	0.055
1,024	111,210	29,743	3.74	0.000	0.001
4,096	205,041	86,773	2.36	0.000	0.000
16,384	417,386	249,855	1.67	0.000	0.000
65,536	846,246	742,665	1.14	0.000	0.000
262,144	1,660,543	2,327,210	0.71	0.000	0.000

$h(s)$ denote the $+/-$ half within its group. Since $\mathbf{1}_{A_\theta}(s) = (1 + \theta_b h(s))/2$ on G_b,

$$\mathfrak{R}_n(\mathcal{F}_w) = \frac{1}{2n} \sum_{b=1}^{B} \mathbb{E}|Z_b|, \qquad Z_b = \sum_{i=1}^{n} \sigma_i h(S_i) \mathbf{1}\{S_i \in G_b\}. \tag{9}$$

The constant term has expectation zero. A summand in Z_b is ± 1 with probability $w_b/2$ each, and zero otherwise. Thus, putting $v = nw_b$,

$$\mathbb{E}Z_b^2 = v, \qquad \mathbb{E}Z_b^4 \le v + 3v^2.$$

Hölder's inequality gives $\mathbb{E}|Z_b| \ge (\mathbb{E}Z_b^2)^{3/2}/(\mathbb{E}Z_b^4)^{1/2} \ge v/\sqrt{1+3v}$, while $\mathbb{E}|Z_b| \le \min\{v, \sqrt{v}\}$. Consequently

$$\frac{1}{4} \sum_b \min\{w_b, \sqrt{w_b/n}\} \le \mathfrak{R}_n(\mathcal{F}_w) \le \frac{1}{2} \sum_b \min\{w_b, \sqrt{w_b/n}\}. \tag{10}$$

If $nw_{\min} \ge 1$, this is $\Theta(\sum_b \sqrt{w_b}/\sqrt{n})$. For a plateau and $n \ge k$, the heavy groups contribute $\Theta(\sqrt{k/n})$ and the residual contributes at most $\varepsilon_B/2$. Therefore, around $n \asymp k/\gamma^2$ with $\varepsilon_B \le \gamma/16$, the complexity is of order γ. Distribution-dependent complexity does distinguish our profiles; it is not one of the quantities matched in Theorem 2.

For indicator policies the squared $L^2(P_0^S)$ distance is

$$\|\mathbf{1}_{A_\theta} - \mathbf{1}_{A_{\theta'}}\|_2^2 = \sum_b w_b \mathbf{1}\{\theta_b \ne \theta'_b\}.$$

At any fixed sufficiently small positive resolution, a plateau has logarithmic covering and packing numbers $\Theta(k)$: ignoring residual coordinates gives a 2^k cover once the squared resolution exceeds ε_B, and a constant-relative-distance binary code on the heavy coordinates gives a packing of size $\exp(ck)$. Thus the usual metric entropy sees the heavy coordinates. The point is a difference between estimation and testing rates, not failure of every distribution-dependent complexity measure.

C.2 Proof of the IUT rate separation

Table 5: Two statistical scales on a k-plateau (constants and the residual suppressed).

Evidence	Signal	Reference fluctuation	Sample scale
Best (minimum) empirical policy risk	γ	$\sqrt{k/n}$	k/γ^2
Quadratic family evidence	γ^2	$\sqrt{k}/n$	$\sqrt{k}/\gamma^2$

Part (ii) of Theorem 3 is proved for the exact Clopper–Pearson IUT with any fixed positive allocation $\zeta_c + \zeta_r = \delta$. Part (i) is proved for every level-δ EFC audit: on the main range the argument below only uses that the audit does not reject while the constructed policy $\widehat{A}$ is empirically feasible, and level δ is used only in the bounded ranges at the end. Constants below may depend on p, δ, β but not on B, k, γ.

Upper bound. Let $\mathcal{R}$ be the residual groups, of reference mass $\varepsilon \le \gamma/16$, with $\varepsilon = 0$ for $k = B$. For a heavy-coordinate sign pattern θ_H, define

$$T_H = \sum_i \mathbf{1}\{S_i \in A_{\theta_H} \setminus \mathcal{R}\}, \qquad W_H = \sum_i (L_i - \alpha)\mathbf{1}\{S_i \in A_{\theta_H} \setminus \mathcal{R}\}.$$

Their expectations are $n(1-\varepsilon)/2$ and $n(1-\varepsilon)\gamma/2$. A union bound over only 2^k heavy patterns and Hoeffding's inequality give, with probability at least $1-\eta$, simultaneously

$$T_H \ge n(1-\varepsilon)/2 - C\sqrt{n(k+\log(1/\eta))}, \qquad W_H \ge n(1-\varepsilon)\gamma/2 - C\sqrt{n(k+\log(1/\eta))}.$$

Also $N_{\mathcal{R}} \le n\varepsilon + C\sqrt{n\log(1/\eta)}$ on an event of the same order of probability. Any full policy has $T_A \ge T_H$ and $E_A - \alpha T_A \ge W_H - \alpha N_{\mathcal{R}}$. Since $\varepsilon \le \gamma/16$, at $n \ge C'(k+\log(1/\eta))/\gamma^2$ these inequalities imply $T_A \ge n/4$ and $E_A/T_A \ge \alpha + c\gamma$ for *every* policy. A one-sided Clopper–Pearson lower bound is at least the empirical proportion minus $\sqrt{\log(1/\zeta_r)/(2T_A)}$ whenever this difference is positive: this follows by inverting the Hoeffding bound for a binomial upper tail. Increasing C' makes every risk lower bound exceed α. Taking $\eta < 1-\beta$ proves $N_{\mathrm{IUT}} \le Ck/\gamma^2$. This argument does not pay for the number of residual sign patterns.

A constructive empirical witness for the lower bound. Condition on the total number of errors M, and write $m = n - M$ for the number of correct observations. In each heavy group choose the half with *more correct observations*; fix the $+$ half in every residual group. This defines a policy $\widehat{A}$ using correct observations alone. Conditional on these observations and M, the error observations have independent uniform scores, and any selected sign policy has score probability $1/2$. Hence

$$E_{\widehat{A}} = M/2 + Z, \qquad \mathbb{E}[Z \mid M, \text{correct data}] = 0, \quad \mathbb{E}[Z^2 \mid M, \text{correct data}] = M/4.$$

If $C_{b,+}, C_{b,-}$ are the heavy correct counts and $C_{\mathcal{R},+}$ is the correct count in the fixed residual halves, then

$$C_{\widehat{A}} = m/2 + G, \qquad G = \tfrac{1}{2}\sum_{b\le k} |C_{b,+} - C_{b,-}| + C_{\mathcal{R},+} - C_{\mathcal{R}}/2.$$

The last term has conditional mean zero. Each heavy difference has variance $v = m(1-\varepsilon)/k$ and fourth moment at most $v + 3v^2$, so

$$\mathbb{E}[G \mid M] \ge \frac{m(1-\varepsilon)}{2\sqrt{1+3m(1-\varepsilon)/k}} \ge c_0 \min\{m, \sqrt{mk}\}.$$

The second inequality holds uniformly for $\varepsilon \le 1/2$. Replacing one correct observation changes the maximum selected correct count by at most one. Efron–Stein therefore gives $\mathrm{Var}(G \mid M) \le m/2$

(the weaker bound $2m$ also suffices). Chebyshev, binomial concentration for M, and the conditional variance of Z imply, except with probability at most $C_0(1/n + 1/k)$,

$$G \geq c_1 \min\{n, \sqrt{nk}\}, \quad |M - np| + 2|Z| \leq c_2 \min\{n, \sqrt{nk}\},$$

where $c_2 > 0$ can be chosen sufficiently small relative to pc_1 by changing C_0. Here the first bound also uses $m \geq (1-p)n/2$; its failure has probability $O(1/n)$.

For $\gamma \leq \gamma_0(p)$ sufficiently small and $n \leq ck/\gamma^2$ with c sufficiently small, $n\gamma \leq c_3 \min\{n, \sqrt{nk}\}$. Since $\alpha = p - \gamma \geq p/2$, the same event yields

$$T_{\widehat{A}} = n/2 + G + Z \geq n/2, \qquad E_{\widehat{A}} - \alpha T_{\widehat{A}} = \tfrac{1}{2}(M - \alpha n) + (1-\alpha)Z - \alpha G \leq 0.$$

Thus $\widehat{A}$ is empirically feasible, so no EFC audit rejects; for the exact IUT this is Lemma 1. For large enough n, k, its rejection probability is below β throughout $n \leq ck/\gamma^2$.

For completeness, the bounded-k cases follow from the information-theoretic lower bound $N \geq c'\sqrt{k}/\gamma^2$, which is at least a constant times k/γ^2 on any fixed bounded range. For bounded n and sufficiently large k, the mixture ceiling in Proposition 3 is below β, uniformly over bounded γ: $\|w\|_2^2 = O(1/k)$ makes its exponent $O(n^2\gamma^4/k)$. Combining these ranges proves the claimed uniform lower rate. This proof controls the coverage screen as well as risk; minimizing error counts alone would not suffice.

C.3 Proof of Proposition 4

Every rule of the sign family has reference coverage exactly κ. The estimation cost does not depend on this boundary case.

Proposition 4 (Estimation cost with coverage slack)**.** *Split each group of a k-plateau into q equal parts and let each policy accept m parts per group (Proposition 8). With reference coverage m/q, a fixed floor $\kappa < m/q$ and residual mass at most $m\gamma/(16q)$, there are $\gamma_0, c_1, C_1 > 0$ depending only on $(p, q, m, \kappa, \delta, \beta)$ such that, for $\gamma \leq \gamma_0$, every level-δ EFC audit has power below β at P_0 for all $n \leq c_1 k/\gamma^2$, while the exact IUT has power at least β for $n \geq C_1 k/\gamma^2$.*

The witness chooses, in each heavy group, the m parts with the most correct observations; with slack the coverage screen is passed automatically, and the gain in correct counts again scales as $\min\{n, \sqrt{nk}\}$. Use the construction of Proposition 8: P_0 has uniform scores and independent $L \sim \mathrm{Bern}(p)$, each group G_ℓ is split into q equal parts, a policy selects m parts per group, $c = m/q > \kappa$, $\alpha = p - \gamma$, and the residual groups have total mass $\varepsilon \leq c\gamma/16$.

Upper bound. Repeat the upper-bound argument of Appendix C.2 with the $\binom{q}{m}^k$ heavy patterns in place of 2^k: the union bound costs $k \log \binom{q}{m}$, every policy has $T_A \geq nc/2$ and $E_A - \alpha T_A \geq nc\gamma/4$ once $n \geq Ck/\gamma^2$, and the Clopper–Pearson inversion makes every risk bound exceed α.

A top-m inequality. If $x_1, \ldots, x_q$ sum to zero, the sum of the m largest is at least $\frac{\min\{m, q-m\}}{2q} \sum_a |x_a|$. Indeed, let $P = \sum_a (x_a)_+ = \frac{1}{2}\sum_a |x_a|$ and let r be the number of positive entries. If $m \leq r$, the m largest positive entries sum to at least $(m/r)P \geq (m/q)P$. Otherwise the top m contain all positive entries and the $m - r$ negative entries of smallest magnitude, whose total magnitude is at most $\frac{m-r}{q-r}P$, so the sum is at least $\frac{q-m}{q-r}P \geq \frac{q-m}{q}P$.

Witness. Condition on the error count M and on all correct observations; let $m_0 = n - M$. In each heavy group choose the m parts with the most correct observations, and fix m parts in every residual group. Conditionally, the error observations have independent uniform scores and every policy has score probability c, so the accepted errors are $E = cM + Z$ with $\mathbb{E}Z = 0$ and $\mathrm{Var}\, Z = Mc(1-c)$. The accepted correct count is $cm_0 + G$, where

$$G = \sum_{b \leq k} \Big(\sum_{a \in \widehat{C}_b} N_{b,a} - cN_b \Big) + R,$$

$N_{b,a}$ is the correct count in part a of group b, $N_b = \sum_a N_{b,a}$, and R is the centred residual contribution. By the top-m inequality applied to $x_a = N_{b,a} - N_b/q$, each heavy term is at least

$\frac{\min\{m,q-m\}}{2q}\sum_a |N_{b,a} - N_b/q|$. Each $N_{b,a} - N_b/q$ is a sum of m_0 independent centred terms bounded by one, with variance $v = m_0 w_b(q-1)/q^2$ and fourth moment at most $v + 3v^2$. Hölder's inequality gives $\mathbb{E}|N_{b,a} - N_b/q| \geq v/\sqrt{1+3v}$, and therefore $\mathbb{E}[G \mid M] \geq c_0 \min\{m_0, \sqrt{m_0 k}\}$ with $c_0 > 0$ depending on (q, m). Moving one correct observation changes the top-m sums by at most one and $c\sum_b N_b$ by at most $2c$, so Efron–Stein gives $\mathrm{Var}(G \mid M) \leq 9m_0/2$.

Empirical feasibility. As in Appendix C.2, Chebyshev's inequality and binomial concentration give, except with probability $C_0(1/n + 1/k)$,

$$G \geq c_1' \min\{n, \sqrt{nk}\}, \qquad |M - np| + 2|Z| \leq c_2' \min\{n, \sqrt{nk}\},$$

with c_2' small relative to $\alpha c_1'$ and to $c - \kappa$. The accepted count is $T = cn + G + Z \geq cn - |Z| \geq \kappa n$, since $|Z| \leq (c-\kappa)n$ on this event. Moreover

$$E - \alpha T = c(M - pn) + c\gamma n + (1-\alpha)Z - \alpha G \leq 0$$

whenever $c\gamma n \leq c_3 \min\{n, \sqrt{nk}\}$, which holds for $n \leq c_1 k/\gamma^2$ and $\gamma \leq \gamma_0$. The selected policy is then empirically feasible, and no EFC audit rejects. The bounded-k and bounded-n ranges are handled exactly as in Appendix C.2, using the lower bound $\Omega(\sqrt{k}/\gamma^2)$ and the product-mixture ceiling of Proposition 8. □

C.4 Explicit constants when only the loss rate is unknown

Theorem 4 relies on a generic independence tester. When the score marginal is known and uniform, a pilot-and-collision audit gives explicit constants.

Proposition 5 (Adaptation to an unknown constant risk)**.** *On the uniform sign family $B = k$, let P_p have uniform scores over the $2k$ halves and independent $L \sim \mathrm{Bern}(p)$, and let $\mathcal{P}_{a,\alpha,\gamma} = \{P_p : p \in [a, 1-a],\ p \geq \alpha + \gamma\}$. Fix $0 < a < 1/3$, $\alpha \in [a, 1-2a]$ and $0 < \delta < \beta < 1$. For sufficiently small $0 < \gamma \leq a$, the audit below has level δ over the unrestricted null, does not take p as input, and has power β over $\mathcal{P}_{a,\alpha,\gamma}$ with $O_{a,\delta,\beta}(\sqrt{k}/\gamma^2)$ samples, which is optimal.*

We prove Proposition 5. The score-space partition is known, but the loss probability p is not an input to the audit. For power, the reference class has uniform scores on the $2k$ halves and independent $L \sim \mathrm{Bern}(p)$ with $p \in [a, 1-a]$ and $p \geq \alpha + \gamma$. Type-I validity continues to cover *every* law in the unrestricted existential null, without uniform-score or score–loss independence assumptions.

Set $\epsilon = \min\{\delta, (1-\beta)/3\}$. Use m pilot observations to estimate p by their mean loss projected onto $[a, 1-a]$, denoted $\widehat{p}$. If $\widehat{p} < \alpha + \gamma/2$, do not reject. Otherwise apply the collision test from Theorem 6 to an independent batch of n observations with the candidate reference $P_{\widehat{p}}$. For each candidate,

$$\Delta_{\widehat{p}} = \frac{(\widehat{p} - \alpha)^2}{2\widehat{p}(1-\widehat{p})} \geq \gamma^2/2, \qquad p_{\min} \geq a/(2k), \qquad d = 4k.$$

All categorical cells are reference-positive. A common sufficient size for the second batch is

$$n \geq \max\left\{2, \left\lceil \frac{32}{\epsilon} \max\left\{ \frac{4\sqrt{k}}{\gamma^2}, \frac{2}{\gamma^2} + \frac{2\sqrt{k}}{\gamma\sqrt{a}} \right\} \right\rceil \right\}. \tag{11}$$

Indeed, $\sqrt{d-1}/\Delta \leq 4\sqrt{k}/\gamma^2$ and $u_0/\Delta = 1/\Delta + 1/\sqrt{\Delta p_{\min}}$ is bounded by the second expression. Conditional on *any* pilot, the collision test has level at most ϵ for every feasible law; a failed pilot screen never rejects. Therefore the entire procedure has level at most δ without assuming the power model.

Under P_p, projection cannot increase squared estimation error, so $\mathbb{E}(\widehat{p} - p)^2 \leq 1/(4m)$. For candidate Bernoulli laws,

$$\mathrm{KL}(P_p \| P_{\widehat{p}}) = \mathrm{kl}(p \| \widehat{p}) \leq \frac{(p - \widehat{p})^2}{a(1-a)}.$$

Pinsker and Jensen yield

$$\mathbb{E}\,\mathrm{TV}(P_p^n, P_{\widehat{p}}^n) \leq \sqrt{\frac{n}{8ma(1-a)}}.$$

Also $P_p(\widehat{p} < \alpha + \gamma/2) \le e^{-m\gamma^2/2}$. Taking

$$m \ge \max\left\{\frac{n}{8a(1-a)\epsilon^2}, \frac{2\log(1/\epsilon)}{\gamma^2}\right\} \tag{12}$$

bounds both terms by ϵ. On a passed screen the test has power at least $1-\epsilon$ under its candidate $P_{\widehat{p}}$; transferring to P_p and averaging gives power at least $1 - 3\epsilon \ge \beta$. The total $m+n$ is $O_{a,\delta,\beta}(\sqrt{k}/\gamma^2)$. The lower bound follows by fixing $p = \alpha + \gamma$ in the reference class and applying the uniform-profile mixture bound, with constants uniform over $[a, 1-a]$.

The statistic can be computed from cell counts in $O(n+k)$ time using $U_n = [\sum_j N_j(N_j - 1)/p_j]/[n(n-1)] - 1$. Its conservative constants are not claimed to be practically competitive. The result establishes the rate and refutes a universal interpretation of k/γ^2 as the unavoidable cost of not knowing p. It does not solve adaptation to an arbitrary unknown score marginal or conditional risk function.

D Proof of the matched same-cardinality complexity spectrum

D.1 Exact singleton sufficiency and audit equivalence

Every component has exactly

$$\mathrm{TV}(P_0, Q_\theta) = \gamma, \tag{13}$$

$$\mathrm{KL}(P_0 \| Q_\theta) = \tfrac{1}{2}\,\mathrm{kl}(p \| p-\gamma) + \tfrac{1}{2}\,\mathrm{kl}(p \| p+\gamma), \tag{14}$$

$$\mathrm{KL}(Q_\theta \| P_0) = \tfrac{1}{2}\,\mathrm{kl}(p-\gamma \| p) + \tfrac{1}{2}\,\mathrm{kl}(p+\gamma \| p). \tag{15}$$

For reference coverage c, accepted risk r, $0 < \kappa \le c$, and $\alpha = r - \gamma$ with $0 < \gamma < r$, the exact projections are as follows. Moreover, putting $d = \mathrm{kl}(r-\gamma \| r)$ and $t^* = \max\{\kappa, ce^{-d}/(1-c+ce^{-d})\}$,

$$\inf_{Q \in \mathcal{H}(\{A\})} \mathrm{TV}(P_0, Q) = c\gamma, \tag{16}$$

$$\inf_{Q \in \mathcal{H}(\{A\})} \mathrm{KL}(P_0 \| Q) = c\,\mathrm{kl}(r \| r-\gamma), \tag{17}$$

$$\inf_{Q \in \mathcal{H}(\{A\})} \mathrm{KL}(Q \| P_0) = \mathrm{kl}(t^* \| c) + t^* d. \tag{18}$$

The four-cell sufficient-statistic reduction is proved below. The constant-risk case used below has $r = s = p$ and gives these formulas as a special case. Note that the canonical hiding component Q_θ in Equation (5) is not the closest feasible singleton law: for $c = 1/2$, Equation (13) gives $\mathrm{TV}(P_0, Q_\theta) = \gamma$, whereas the singleton projection in Equation (16) is $\gamma/2$.

Proposition 6 (Singleton experiments are matched)**.** *For a policy A whose four $(\mathbf{1}\{S \in A\}, L)$ reference cells are positive, the unrestricted singleton testing problem at every sample size is determined by this four-cell law. Hence every member of every sign family has the same singleton minimax complexity.*

We prove Proposition 6 without assuming that L is independent of the score. Let $M = \mathbf{1}\{S \in A\}$ and $T(S, L) = (M, L)$. The singleton contract depends on a law only through the four probabilities of T. Fix once and for all a Markov kernel $K(ds \mid m, l)$ supported on A when $m = 1$ and on A^c when $m = 0$, taking $K(B \mid m, l) = P_0(S \in B, M = m, L = l)/P_0(M = m, L = l)$, the normalized restriction of P_0 to that finite event. Proposition 6 assumes all four cells are positive, so no general disintegration theorem is needed.

Every feasible four-cell law q has a canonical lift Q_q: draw $T \sim q$ and then $S \sim K(\cdot \mid T)$. Conversely, $T_\# Q$ is feasible whenever $Q \in \mathcal{H}(\{A\})$. Hence data processing gives one direction for every divergence and every test.

The testing reduction is exact for every sample size. Given any original-sample test ϕ_n, define

$$\psi_n(t^n) = \int \phi_n(s^n, l^n) \prod_{i=1}^n K(ds_i \mid t_i).$$

For every canonical lift Q_q, $\mathbb{E}_{Q_q^n}\phi_n = \mathbb{E}_{q^n}\psi_n$, and the same identity holds under P_0^n. Therefore every valid original test induces a four-cell test with the same power, while the converse follows by composition with T^n. The unrestricted singleton minimax problem thus depends on (P_0, A) only through $T_\# P_0$. In particular, two policies with identical reference four-cell laws have identical singleton audit complexity even when $\eta(s) = P_0(L = 1 \mid S = s)$ is nonconstant.

For the explicit projections, write

$$c = P_0(M = 1), \quad r = P_0(L = 1 \mid M = 1), \quad s = P_0(L = 1 \mid M = 0), \quad \alpha = r - \gamma.$$

For TV, $f(M, L) = M(L - \alpha)$ has oscillation one. Under P_0, $\mathbb{E}f = c\gamma$, whereas every feasible Q has $\mathbb{E}_Q f \le 0$, so $\mathrm{TV}(P_0, Q) \ge c\gamma$. Equality is attained in the four-cell problem by preserving $P(M = 1) = c$ and the outside conditional risk s while changing the accepted risk from r to $r - \gamma$; the canonical lift preserves TV exactly.

For the KL projections, write $t = Q(M = 1)$, $u = Q(L = 1 \mid M = 1)$, and $v = Q(L = 1 \mid M = 0)$. Data processing and the Bernoulli chain rule give

$$\mathrm{KL}(P_0 \| Q) \ge \mathrm{kl}(c\|t) + c\,\mathrm{kl}(r\|u) + (1 - c)\,\mathrm{kl}(s\|v),$$
$$\mathrm{KL}(Q\|P_0) \ge \mathrm{kl}(t\|c) + t\,\mathrm{kl}(u\|r) + (1 - t)\,\mathrm{kl}(v\|s).$$

The feasible constraints are $t \ge \kappa$ and $u \le r - \gamma$. The forward direction is minimized at $(t, u, v) = (c, r - \gamma, s)$, giving $c\,\mathrm{kl}(r\|r - \gamma)$.

For the reverse direction first set $(u, v) = (r - \gamma, s)$, optimal for every fixed t. With $d = \mathrm{kl}(r - \gamma\|r)$ the remaining objective is

$$h(t) = \mathrm{kl}(t\|c) + td, \qquad t \ge \kappa.$$

Its unconstrained minimizer is $t_0 = ce^{-d}/(1 - c + ce^{-d})$, so the constrained optimizer is $t^* = \max\{\kappa, t_0\}$ and the minimum is $\mathrm{kl}(t^*\|c) + t^*d$. Canonical lifts attain the data-processing bounds because the likelihood ratio depends only on T. This proves Equation (18). The score-independent special case is obtained by $r = s = p$.

D.2 Common component information

Every policy A_θ accepts one half of each group, so $P_0(S \in A_\theta) = 1/2$ and under (5), $Q_\theta(L = 1 \mid S \in A_\theta) = p - \gamma = \alpha$. Half of the score mass shifts the Bernoulli parameter down by γ and half shifts it up by γ, independently of group weights. Summing the two-point channel distances gives (15) and $\mathrm{TV}(P_0, Q_\theta) = \gamma$.

For independent uniform signs θ, θ', the standardized Bernoulli calculation gives

$$\Xi(\theta, \theta') = \frac{\gamma^2}{p(1 - p)} \sum_{\ell=1}^{B} w_\ell \theta_\ell \theta'_\ell = \lambda \sum_{\ell=1}^{B} w_\ell \varepsilon_\ell,$$

where ε_ℓ are i.i.d. Rademacher variables. The nonzero covariance eigenvalues are proportional to w_ℓ, giving $d_{\text{eff}} = (\sum_\ell w_\ell^2)^{-1}$.

Proposition 7 (Combinatorial invariance of the matched sign families)**.** *For any two positive profiles $w, w' \in \Delta_{B-1}$, the corresponding sign families have the same nonempty membership-signature incidence structure up to relabeling of the $2B$ half-group atoms. In particular, both have cardinality 2^B, VC dimension exactly B, and growth function $\Pi_{\mathcal{A}}(m) = 2^{\min\{m,B\}}$. Thus changing the profile changes only the P_0-masses assigned to a common combinatorial set system, not its unweighted class complexity.*

Remark 2 (Pairwise matching inside the plateau spectrum). If $\theta^{(\ell)}$ is obtained from θ by flipping only coordinate ℓ, then $A_\theta \triangle A_{\theta^{(\ell)}} = G_\ell$ and hence $P_0(A_\theta \triangle A_{\theta^{(\ell)}}) = w_\ell$. Thus adjacent pairwise membership laws reveal the plateau weights. A pairwise-matched analogue of Theorem 2 therefore requires a qualitatively different construction; the Supplementary Note gives such a construction for the weaker but already nontrivial question of one-observation family hiding.

D.3 Proof of Proposition 7: fixed incidence and VC complexity

For every positive profile, quotient the score space by policy-membership signature. The nonempty signature atoms are exactly the $2B$ half-groups G_ℓ^+ and G_ℓ^-: policy A_θ contains precisely $G_\ell^{\theta_\ell}$ from

group ℓ. This $2B$-atom incidence matrix depends on B and the sign labels but not on the mass profile w. Hence any two profiles have isomorphic quotient set systems; only the reference masses of the atoms differ.

For the VC dimension, choose one point from G_ℓ^+ for each of $m \le B$ distinct groups. The corresponding m membership bits can be set arbitrarily by the m sign coordinates, so every such m is shattered and $\mathrm{VCdim}(\mathcal{A}) \ge B$. Conversely, among any $B+1$ points two lie in the same group. If they lie in the same half, every policy gives them identical membership; if they lie in opposite halves, every policy gives complementary membership. In either case the pair cannot realize all four binary labelings, so no $B+1$ points are shattered. Thus $\mathrm{VCdim}(\mathcal{A}) = B$.

Finally, for $m \le B$ the shattered construction realizes 2^m traces. For $m > B$, the family contains only 2^B policies, while an m-point set containing one point from each G_ℓ^+ plus additional distinct points in any already represented half-group still distinguishes all 2^B sign vectors. Therefore $\Pi_{\mathcal{A}}(m) = 2^{\min\{m,B\}}$.

D.4 Proof of Proposition 3

Let w be any positive group-mass vector and let $\varepsilon_\ell = \theta_\ell \theta'_\ell$ for independent uniform sign vectors. Then

$$\Xi(\theta, \theta') = \lambda \sum_{\ell=1}^{B} w_\ell \varepsilon_\ell, \qquad \lambda = \frac{\gamma^2}{p(1-p)}.$$

By Lemma 3, $1+\chi^2(R_n \| P_0^n) = \mathbb{E}(1+\Xi)^n$. Since $1+x \le e^x$ for $x > -1$ and the ε_ℓ are independent Rademacher variables,

$$\begin{aligned} 1 + \chi^2(R_n \| P_0^n) &\le \mathbb{E} \exp\left(n\lambda \sum_{\ell} w_\ell \varepsilon_\ell \right) \\ &= \prod_{\ell=1}^{B} \cosh(n\lambda w_\ell) \le \exp\left(\frac{n^2\lambda^2}{2} \sum_{\ell=1}^{B} w_\ell^2 \right), \end{aligned}$$

where $\cosh t \le e^{t^2/2}$. Corollary 3 now gives the ceiling of Proposition 3. Choosing the constant in $n \le c/(\lambda \|w\|_2)$ so that the right-hand side stays below β yields $N_w = \Omega(\sqrt{d_{\mathrm{eff}}}/\gamma^2)$.

Two uniform bounds delimit the entire sign-family class. First, any one canonical component has directed KL $\Theta_p(\gamma^2)$ from P_0, so the standard two-point lower bound (Le Cam and Yang, 2000) gives $N_w = \Omega_p(1/\gamma^2)$ for every profile. Second, for any witness policy A_θ, feasibility implies

$$Q(A_\theta, L=0) \ge \frac{1-\alpha}{2}, \qquad P_0(A_\theta, L=0) = \frac{1-p}{2},$$

so every law in the full existential null is at least $\gamma/2$ away after the $4B$ half-group/label pushforward. Generic known-distribution identity testing on $4B$ categories therefore gives $N_w = O(\sqrt{B}/\gamma^2)$, uniformly in w. Consequently the ratio between any two complexities in this matched sign-family class is at most $O(\sqrt{B})$. Theorem 2 attains every $\sqrt{k}$ order along the plateau subclass and in particular attains this maximal endpoint spread.

D.5 Proof of Theorem 2: plateau spectrum

For $k < B$, the plateau profile $w_\ell^{(k)} = (1-\varepsilon_B)/k$ ($\ell \le k$), $w_\ell^{(k)} = \varepsilon_B/(B-k)$ ($\ell > k$), $\varepsilon_B = B^{-2}$, satisfies

$$\|w^{(k)}\|_2^2 = \frac{(1-\varepsilon_B)^2}{k} + \frac{\varepsilon_B^2}{B-k}, \qquad \varepsilon_B = B^{-2}.$$

Uniformly over $1 \le k < B$, the second term is $o(1/k)$ and $(1-\varepsilon_B)^2 = 1 + o(1)$; for $k = B$, $\|w^{(B)}\|_2^2 = 1/B$. Hence $d_{\mathrm{eff}}(w^{(k)}) = k\{1+o(1)\}$ uniformly in k. Proposition 3 therefore gives

$$N_{B,k} = \Omega\left(\frac{\sqrt{k}}{\gamma_B^2} \right).$$

It remains to prove the matching upper bound against the *full unrestricted* existential null. For $k < B$, let

$$R = \bigcup_{\ell=k+1}^{B} G_\ell, \qquad P_0(R) = \varepsilon_B,$$

and retain separately the two halves of each heavy group $G_1, \dots, G_k$. Push (S, L) to the $(4k+2)$ categories consisting of heavy-group half/label pairs plus $R \times \{0,1\}$. Let Q be any law in the full null and let A_θ be a witness. Write $A_H = A_\theta \cap R^c$ for the accepted halves of the heavy groups.

If $|Q(R) - \varepsilon_B| \geq \gamma_B/16$, the pushed-forward laws already differ in TV by at least $\gamma_B/16$. Otherwise $Q(R) \leq \varepsilon_B + \gamma_B/16$. Feasibility of A_θ gives

$$Q(A_\theta, L=0) \geq (1-\alpha)Q(A_\theta) \geq \frac{1-p+\gamma_B}{2},$$

and hence

$$Q(A_H, L=0) \geq \frac{1-p+\gamma_B}{2} - Q(R),$$
$$P_0(A_H, L=0) = \frac{(1-p)(1-\varepsilon_B)}{2}.$$

Therefore

$$Q(A_H, L=0) - P_0(A_H, L=0) \geq \frac{7\gamma_B}{16} - \frac{1+p}{2}\varepsilon_B \geq \frac{3\gamma_B}{8}$$

whenever $\varepsilon_B \leq \gamma_B/16$. Thus every law in the full null is at TV distance at least $\gamma_B/16$ from P_0 after a pushforward with $4k+2$ categories. Known-distribution identity testing on $m = 4k+2$ categories uses $O(\sqrt{m}/\gamma_B^2) = O(\sqrt{k}/\gamma_B^2)$ samples. For $k = B$, the residual is empty and the same argument reduces to the standard $4B$ pushforward. Combining with the lower bound proves Theorem 2, and $d_{\text{eff}}(w^{(k)}) = k\{1 + o(1)\}$ as computed above.

The condition $\varepsilon_B \leq \gamma_B/16$ also makes the vanishing-gap statement stable when ε_B/γ_B tends to a nonzero constant below $1/16$: the displayed separation remains a fixed positive fraction of γ_B, so no hidden constant degenerates in the pushforward upper bound.

D.6 Balanced q-ary spectrum with positive coverage slack

The binary construction uses reference coverage $1/2$ only for notation. The following balanced construction simultaneously extends the matched spectrum to rational reference coverage and removes the boundary condition $P_0(A) = \kappa$.

Proposition 8 (Matched spectrum with positive coverage slack)**.** *Fix a rational reference coverage $c = m/q \in (0,1)$ with integers $1 \leq m < q$, a contract floor $0 < \kappa \leq c$, $p \in (0,1)$, and $0 < \delta < \beta < 1$. Let $\gamma = \gamma_B$ satisfy*

$$\gamma_B < p, \qquad \frac{c}{1-c}\gamma_B < 1-p, \qquad B^{-2} \leq \frac{c\gamma_B}{16}.$$

For each $k \in \{1, \dots, B\}$ use the plateau profile $w^{(k)}$ of Section 4. There is a family with $J_B = \binom{q}{m}^B$ policies, common reference P_0, common raw support, common contract $(\kappa, p - \gamma_B)$, and exact one-observation hiding such that every policy has reference coverage exactly c and the same unrestricted singleton testing problem, while, uniformly in k,

$$N_{B,k}^{(c,\kappa)} = \Theta_{p,c,\kappa,\delta,\beta}\left(\frac{\sqrt{k}}{\gamma_B^2}\right).$$

Consequently, for every fixed $c > \kappa$ the construction has fixed positive coverage slack $c - \kappa$ for every member, and its endpoint complexity ratio is $\Theta_c(\sqrt{\log J_B})$.

Construction and exact hiding. Partition $[0,1]$ into groups $G_1, \dots, G_B$ of masses $w_1, \dots, w_B$, and split every group into q equal-mass consecutive subintervals $G_{\ell,1}, \dots, G_{\ell,q}$. A local policy chooses an m-subset $C_\ell \subset [q]$; a global policy $\theta = (C_1, \dots, C_B)$ accepts

$$A_\theta = \bigcup_{\ell=1}^{B} \bigcup_{a \in C_\ell} G_{\ell,a}.$$

Thus $P_0(A_\theta) = m/q = c$ for every member. In this one-dimensional realization each policy is a union of at most mB intervals, so the construction does not rely on nonmeasurable or arbitrarily fragmented acceptance sets, although the interval budget grows with B.

Let

$$a_c = \frac{c}{1-c}\gamma_B, \qquad Q_\theta(L=1 \mid S=x) = \begin{cases} p-\gamma_B, & x \in A_\theta, \\ p+a_c, & x \notin A_\theta. \end{cases}$$

The parameter conditions keep both Bernoulli probabilities in $(0,1)$. Each canonical component is feasible because it has coverage $c \geq \kappa$ and accepted risk $p - \gamma_B = \alpha$. Under the uniform prior on the $\binom{q}{m}^B$ policies, each score subinterval is selected with probability $m/q = c$, hence its average perturbation is $-c\gamma_B + (1-c)a_c = 0$. Therefore the mixture barycenter is exactly P_0.

All members have the same singleton problem by Proposition 6. The canonical components are also matched componentwise across profiles:

$$\begin{aligned} \mathrm{TV}(P_0, Q_\theta) &= 2c\gamma_B, \\ \mathrm{KL}(P_0 \| Q_\theta) &= c\,\mathrm{kl}(p \| p-\gamma_B) + (1-c)\,\mathrm{kl}(p \| p+a_c), \\ \mathrm{KL}(Q_\theta \| P_0) &= c\,\mathrm{kl}(p-\gamma_B \| p) + (1-c)\,\mathrm{kl}(p+a_c \| p). \end{aligned}$$

These quantities depend on (p, c, γ_B) but not on the group-mass profile.

Product-mixture lower bound. For a local accepted subset $C \subset [q]$ define

$$h_C(a) = \begin{cases} -1, & a \in C, \\ c/(1-c), & a \notin C, \end{cases} \qquad Z(C, C') = \frac{1}{q}\sum_{a=1}^{q} h_C(a) h_{C'}(a).$$

For independent uniform m-subsets C, C', Z is centered and bounded by a constant depending only on c (equivalently on the fixed pair m, q). Across groups the Z_ℓ are independent, and direct Bernoulli calculation gives

$$\Xi(\theta, \theta') = \frac{\gamma_B^2}{p(1-p)} \sum_{\ell=1}^{B} w_\ell Z_\ell.$$

Applying Hoeffding's lemma inside Lemma 3 yields

$$1 + \chi^2(R_n \| P_0^n) \leq \exp\left(C_{p,c}\, n^2 \gamma_B^4 \sum_{\ell=1}^{B} w_\ell^2 \right).$$

Corollary 3 therefore implies $N_{B,k}^{(c,\kappa)} = \Omega_{p,c,\delta,\beta}(\sqrt{k}/\gamma_B^2)$ because $\sum_\ell (w_\ell^{(k)})^2 = \Theta(1/k)$ uniformly over the plateau spectrum.

Matching upper bound against the full unrestricted null. For $k < B$, let $R = \cup_{\ell > k} G_\ell$, so $P_0(R) = \varepsilon_B = B^{-2}$, retain the q subcells of every heavy group separately, and push (S, L) to the $2(qk+1)$ resulting cell/label categories. Let Q be any law in the full existential null and let A_θ be a witness. Write $D = A_\theta \cap R^c$ and $d_0 = P_0(D) = c(1-\varepsilon_B)$. Put $a = c/16$.

If either $|Q(R) - \varepsilon_B| \geq a\gamma_B$ or $|Q(D) - d_0| \geq a\gamma_B$, the pushed-forward laws already differ in TV by at least $a\gamma_B$. Otherwise $Q(R) \leq \varepsilon_B + a\gamma_B$ and $Q(D) \geq d_0 - a\gamma_B$. Feasibility gives

$$\begin{aligned} Q(D, L=0) &\geq Q(A_\theta, L=0) - Q(R) \\ &\geq (1 - p + \gamma_B) Q(A_\theta) - Q(R) \\ &\geq (1 - p + \gamma_B) Q(D) - Q(R). \end{aligned}$$

Since $P_0(D, L=0) = (1-p)d_0$,

$$\begin{aligned} Q(D, L=0) - P_0(D, L=0) &\geq \gamma_B d_0 - a\gamma_B(2 - p + \gamma_B) - \varepsilon_B \\ &\geq \frac{11 c \gamma_B}{16}, \end{aligned}$$

where the last inequality uses $d_0 \geq 15c/16$, $2 - p + \gamma_B < 3$, and $\varepsilon_B \leq c\gamma_B/16$. Thus every full-null law is separated from P_0 by at least $c\gamma_B/16$ after a pushforward with $O_c(k)$ categories.

Known-distribution identity testing gives $O_{p,c}(\sqrt{k}/\gamma_B^2)$ samples. For $k = B$ the residual cell is absent and the same argument uses the $2qB$ subcell/label categories. Combining the two bounds proves the proposition.

Taking, for example, $c = 9/10$ and any fixed $\kappa < 9/10$ gives a high-coverage construction with nonzero contract slack. The theorem does not settle the different regime in which the number of allowed intervals K is fixed while $B \to \infty$; that structured question is left open.

D.7 Informative nonconstant-reference spectrum

The score-independent reference in Equation (5) is a convenient special case, not a requirement for the separation. The next construction makes the score genuinely informative about risk and makes selective acceptance strictly useful under the reference law.

Proposition 9 (Matched spectrum under informative nonconstant risk)**.** *Fix constants $a, b, \rho > 0$ with $a + b + \rho = 1$, risks $p_C, p_V, p_H \in (0, 1)$, and $h > 0$ such that $p_V \pm h \in (0, 1)$. Put*

$$c = a + \rho/2, \qquad r_A = \frac{ap_C + (\rho/2)p_V}{c}, \qquad \bar{p} = ap_C + \rho p_V + bp_H,$$

and assume $r_A < \bar{p}$. Fix $0 < \kappa \le c$ and $0 < \delta < \beta < 1$. For any sequence $\gamma_B \downarrow 0$ such that

$$\tau_B := \frac{2c\gamma_B}{\rho} < \min\{p_V - h,\, 1 - p_V - h\}, \qquad \rho B^{-2} \le \frac{c\gamma_B}{8},$$

there are k-plateau families with $J_B = 2^B$ policies, a common reference law with nonconstant $\eta(s) = P_0(L = 1 \mid S = s)$, common contract $(\kappa, r_A - \gamma_B)$, exact one-observation hiding, and identical unrestricted singleton testing problems for every member, while uniformly in $k \in \{1, \dots, B\}$,

$$N_{B,k}^{\inf} = \Theta_{a,b,\rho,p_C,p_V,p_H,h,\kappa,\delta,\beta}\left(\frac{\sqrt{k}}{\gamma_B^2}\right).$$

Every policy has reference coverage c and reference conditional risk $r_A < \bar{p}$, so abstention strictly improves risk under P_0.

Reference and policy family. The reference law must remain literally fixed as the plateau profile changes. Let the variable region carry product coordinates $(X, Z) \in [0, 1] \times \{-1, +1\}$, with conditional reference law $X \sim \mathrm{Unif}[0, 1]$, Z uniform and independent of X. Give the disjoint regions C, V, H masses a, ρ, b; set $\eta = p_C$ on C, $\eta = p_H$ on H, and

$$\eta(X, Z) = p_V + hZ \qquad \text{on } V.$$

This defines one common P_0 for every profile. For a mass vector w, partition only the X coordinate into groups $G_1, \dots, G_B$ of masses $w_1, \dots, w_B$ and split each G_ℓ into equal X-halves G_ℓ^+, G_ℓ^-. Each half is crossed with both values of Z, so it always contains equal masses at risks $p_V - h$ and $p_V + h$, independently of its width or of the profile. (A standard nonatomic-space isomorphism can encode these product coordinates in a scalar score if desired; interval structure is not claimed for this informative extension.)

For $\theta \in \{-1, +1\}^B$, let A_θ accept all of C and $(G_\ell^{\theta_\ell} \times \{-1, +1\})$ from every variable group. Then $P_0(A_\theta) = c$ and every member has reference accepted risk r_A; its rejected risk is also profile-independent. In particular the four-cell law of $(\mathbf{1}\{S \in A_\theta\}, L)$ is the same for every θ and every plateau profile, so Proposition 6 gives exact singleton equivalence without any score–loss independence assumption.

Canonical feasible components and exact hiding. Let $u_\theta(X, Z) = +1$ on the selected X-half of a variable group (for either Z) and -1 on the unselected half. Define Q_θ by preserving the score marginal and setting

$$Q_\theta(L = 1 \mid S = x) = \begin{cases} \eta(x), & x \in C \cup H, \\ \eta(x) - \tau_B u_\theta(x), & x \in V. \end{cases}$$

The parameter condition keeps all Bernoulli probabilities in $(0, 1)$. The accepted error mass decreases by $(\rho/2)\tau_B = c\gamma_B$, hence $Q_\theta(L = 1 \mid S \in A_\theta) = r_A - \gamma_B$ and the component is feasible. Under

the uniform prior on signs, each variable point is selected and rejected with equal probability, so the perturbation cancels pointwise and $\mathbb{E}_\theta Q_\theta = P_0$ exactly. Canonical TV and both directed KL divergences are independent of the plateau profile because every group contains the same replicated two-level risk pattern.

Overlap lower bound. Define

$$H_V = \frac{1}{2}\left\{\frac{1}{(p_V - h)(1 - p_V + h)} + \frac{1}{(p_V + h)(1 - p_V - h)}\right\}.$$

By the finite-cell Fisher representation, for independent signs θ, θ',

$$\Xi(\theta, \theta') = \Lambda_B \sum_{\ell=1}^{B} w_\ell \theta_\ell \theta'_\ell, \qquad \Lambda_B = \rho \tau_B^2 H_V = \frac{4c^2 H_V}{\rho}\gamma_B^2.$$

The same Rademacher MGF calculation as Proposition 3 therefore gives $N_{B,k}^{\inf} = \Omega(\sqrt{k}/\gamma_B^2)$ uniformly over the plateau spectrum.

Matching upper bound against the full unrestricted null. For $k < B$, collapse the residual variable groups $R = \cup_{\ell > k} G_\ell$, whose reference mass is $P_0(R) = \rho\varepsilon_B$, while retaining C, H and each heavy half separately; retain the label in every cell. This gives $4k+6$ categories. Let Q be any full-null law with witness A_θ, put $A_H = A_\theta \setminus R$, and define on the pushforward $g = \mathbf{1}\{S \in A_H\}(L - \alpha)$ with $\alpha = r_A - \gamma_B$. Since $P_0(A_\theta)\{r_{P_0}(A_\theta) - \alpha\} = c\gamma_B$,

$$\mathbb{E}_{P_0} g \geq c\gamma_B - P_0(R).$$

Feasibility gives $\mathbb{E}_Q[\mathbf{1}\{S \in A_\theta\}(L - \alpha)] \leq 0$, hence $\mathbb{E}_Q g \leq Q(R)$. If $|Q(R) - P_0(R)| \geq c\gamma_B/4$, the residual-mass coordinate already separates the pushforwards by $c\gamma_B/4$. Otherwise

$$\mathbb{E}_{P_0} g - \mathbb{E}_Q g \geq \frac{3c\gamma_B}{4} - 2P_0(R) \geq \frac{c\gamma_B}{2},$$

using $\rho\varepsilon_B \leq c\gamma_B/8$. Because g has oscillation one, the pushed-forward TV distance is therefore at least $c\gamma_B/4$ in all cases. Known-distribution identity testing on $O(k)$ categories gives $O(\sqrt{k}/\gamma_B^2)$ samples; for $k = B$ there is no residual category. This matches the lower bound. All constants are uniform over k and over every admissible sequence γ_B satisfying the displayed conditions.

A concrete choice is $(a, \rho, b) = (.30, .50, .20)$ and $(p_C, p_V - h, p_V + h, p_H) = (.08, .16, .32, .60)$, for which $c = .55$, the accepted reference risk is about .153, the population risk is .264, and the score carries substantial risk information while the matched spectrum remains unchanged in order.

E Product-overlap identity and Fisher geometry

Lemma 3 (Product-mixture identity; Ingster, 1987)**.** *For dominated feasible components write* $K(Q, Q') = \int (dQ/dP_0)(dQ'/dP_0)dP_0$ *and* $\Xi = K - 1$. *For a prior* Π *with* $\mathbb{E}_\Pi Q = P_0$ *and* $R_n = \mathbb{E}_\Pi Q^n$,

$$1 + \chi^2(R_n \| P_0^n) = \mathbb{E}_{Q,Q' \sim \Pi}[1 + \Xi(Q, Q')]^n \quad \textit{(in the extended sense).} \tag{19}$$

E.1 Proof of Lemma 3

Write $h_Q = dQ/dP_0$. Since $Q \ll P_0$ almost surely, the product-mixture likelihood ratio is

$$\frac{dR_n}{dP_0^n}(x_{1:n}) = \mathbb{E}_Q \prod_{i=1}^{n} h_Q(x_i).$$

All terms are nonnegative, so Tonelli applies without any per-component L^2 assumption. With independent $Q, Q' \sim \Pi$,

$$\begin{aligned} 1 + \chi^2(R_n \| P_0^n) &= \mathbb{E}_{Q,Q'} \prod_{i=1}^{n} \int h_Q(x) h_{Q'}(x)\, dP_0(x) \\ &= \mathbb{E}_{Q,Q'} K(Q, Q')^n. \end{aligned}$$

The barycenter condition implies $\mathbb{E}_Q h_Q(x) = 1$ for P_0-almost every x, hence by Tonelli $\mathbb{E}_{Q,Q'} K(Q,Q') = \int (\mathbb{E}_Q h_Q)^2 dP_0 = 1$. Therefore $K < \infty$ for $\Pi \otimes \Pi$-almost every pair, $R_1 = P_0$, and with $\Xi = K - 1$,

$$\chi^2(R_2 \| P_0^2) = \mathbb{E}K^2 - 1 = \mathbb{E}(K-1)^2 = \mathbb{E}\Xi^2$$

in the extended sense. This is the standard product-mixture second-moment calculation; the paper-specific content is the selective-contract overlap kernel and its role in the matched separation.

Corollary 1 (Approximate hiding under TV mismatch)**.** *Let Π be supported on $\mathcal{H}(\mathcal{A})$, write $\bar{Q} = \mathbb{E}_{Q\sim\Pi} Q$, and suppose $\mathrm{TV}(P_0, \bar{Q}) \le \varepsilon$. Assume $Q \ll \bar{Q}$ almost surely and define $\bar{K}(Q,Q') = \int (dQ/d\bar{Q})(dQ'/d\bar{Q})\, d\bar{Q}$ and $\bar{\Xi} = \bar{K} - 1$. Then every level-δ full-null-valid test satisfies*

$$\mathbb{E}_{P_0^n}\phi \le \delta + 1 - (1-\varepsilon)^n + \frac{1}{2}\sqrt{\mathbb{E}[1+\bar{\Xi}(Q,Q')]^n - 1} \le \delta + n\varepsilon + \frac{1}{2}\sqrt{\mathbb{E}[1+\bar{\Xi}(Q,Q')]^n - 1}. \quad (20)$$

Corollary 2 (Approximate hiding under quadratic mismatch)**.** *Under the same mixture assumptions, additionally suppose $P_0 \ll \bar{Q}$ and $\chi^2(P_0 \| \bar{Q}) \le \rho$. Then every level-$\delta$ full-null-valid test satisfies*

$$\mathbb{E}_{P_0^n}\phi \le \delta + \frac{1}{2}\sqrt{(1+\rho)^n - 1} + \frac{1}{2}\sqrt{\mathbb{E}[1+\bar{\Xi}(Q,Q')]^n - 1}. \quad (21)$$

E.2 Proof of Corollary 1

Let $\bar{Q} = \mathbb{E}_{Q\sim\Pi} Q$ and $R_n = \mathbb{E}_{Q\sim\Pi} Q^n$. For any level-$\delta$ full-null-valid test,

$$\mathbb{E}_{P_0^n}\phi \le \mathbb{E}_{R_n}\phi + \mathrm{TV}(P_0^n, R_n) \le \delta + \mathrm{TV}(P_0^n, R_n).$$

By the triangle inequality,

$$\mathrm{TV}(P_0^n, R_n) \le \mathrm{TV}(P_0^n, \bar{Q}^n) + \mathrm{TV}(\bar{Q}^n, R_n).$$

A maximal coupling of one draw from P_0 and $\bar{Q}$ disagrees with probability $\mathrm{TV}(P_0, \bar{Q}) \le \varepsilon$; coupling the n coordinates independently gives

$$\mathrm{TV}(P_0^n, \bar{Q}^n) \le 1 - (1-\varepsilon)^n \le n\varepsilon.$$

Finally, $\bar{Q}$ is the exact barycenter of Π, so Lemma 3 applied with reference $\bar{Q}$ gives

$$\mathrm{TV}(\bar{Q}^n, R_n) \le \frac{1}{2}\sqrt{\chi^2(R_n \| \bar{Q}^n)} = \frac{1}{2}\sqrt{\mathbb{E}[1+\bar{\Xi}(Q,Q')]^n - 1}.$$

Combining the three displays proves (20). The TV statement is intentionally assumption-light; its linear product penalty is useful only when $n\varepsilon$ is small.

E.3 Proof of Corollary 2

With the same notation and triangle inequality,

$$\mathrm{TV}(P_0^n, R_n) \le \mathrm{TV}(P_0^n, \bar{Q}^n) + \mathrm{TV}(\bar{Q}^n, R_n).$$

The second term is controlled exactly as above by Lemma 3 with reference $\bar{Q}$. For the first term, product likelihood ratios give

$$1 + \chi^2(P_0^n \| \bar{Q}^n) = \{1 + \chi^2(P_0 \| \bar{Q})\}^n \le (1+\rho)^n.$$

Therefore $\mathrm{TV}(P_0^n, \bar{Q}^n) \le \frac{1}{2}\sqrt{(1+\rho)^n - 1}$. Combining the two terms and test validity proves (21). When $n\rho = o(1)$, $(1+\rho)^n - 1 = O(n\rho)$, giving an $O(\sqrt{n\rho})$ mismatch contribution.

When exact one-observation hiding is possible, minimizing the witness bound in Corollary 3 over all hiding priors yields the family-level obstruction

$$\Omega_n(P_0, \mathcal{H}) = \inf_{\Pi:\, \Pi(\mathcal{H})=1,\ \mathbb{E}_{Q\sim\Pi}[Q]=P_0} \mathrm{TV}(P_0^n, \mathbb{E}_{Q\sim\Pi} Q^n)\,, \quad (22)$$

so every level-δ full-null-valid audit has power at most $\delta + \Omega_n(P_0, \mathcal{H})$. Lemma 3 evaluates any chosen hiding witness, while the infimum above removes dependence on a particular decomposition. We do not assume this infimum is attained without additional compactness; in the finite constructions used below, the displayed witness prior is explicit and order-tight.

E.4 Finite-cell Fisher representation

Assume component Q_k preserves cell masses μ_j and has Bernoulli error rate $\theta_{k,j}$ in cell j, while P_0 has rate $\eta_j \in (0,1)$. Directly evaluating the two label states shows

$$\Xi(Q_k, Q_\ell) = \sum_j \mu_j \frac{(\theta_{k,j} - \eta_j)(\theta_{\ell,j} - \eta_j)}{\eta_j(1-\eta_j)} = \sum_j \mu_j g_k(j) g_\ell(j).$$

Thus the overlap matrix is exactly a Gram matrix in the local Bernoulli Fisher metric.

F Auxiliary testing bounds

For a prior Π over feasible laws with $\mathbb{E}_\Pi Q = P_0$ (a hiding prior) and $R_n = \mathbb{E}_\Pi Q^n$:

Corollary 3 (Witness power ceiling)**.** *For every full-null-valid level-δ test and any hiding prior,*

$$\mathbb{E}_{P_0^n}\phi \le \delta + \mathrm{TV}(P_0^n, R_n) \le \delta + \tfrac{1}{2}\sqrt{\mathbb{E}[1 + \Xi(Q,Q')]^n - 1}. \tag{23}$$

Proof. Every component of the prior is feasible, so a level-δ test satisfies $\mathbb{E}_{R_n}\phi \le \delta$. Hence $\mathbb{E}_{P_0^n}\phi \le \delta + \mathrm{TV}(P_0^n, R_n)$; $\mathrm{TV} \le \frac{1}{2}\sqrt{\chi^2}$ and Lemma 3 give the second inequality. □

F.1 A collision bound for supported nulls

Let $p_i > 0$ on $d \ge 2$ reference-positive cells and define $\mathcal{H}_+$ as the feasible categorical laws that put zero mass on all reference-zero cells. Assume $\mathcal{H}_+$ is nonempty and

$$\Delta = \min_{q \in \mathcal{H}_+} \sum_i (q_i - p_i)^2/p_i > 0, \quad p_{\min} = \min_i p_i, \quad u_0 = 1 + \sqrt{\Delta/p_{\min}}.$$

The infimum is over members, not their one-observation convex hull.

Theorem 6 (Supported collision bracket)**.** *Set $\epsilon = \min\{\delta, 1-\beta\}$ and $c_{\delta,\beta} = \min\{1/2, \frac{1}{4}\log[1/(1-(\beta-\delta)^2)]\}$. Then*

$$\frac{c_{\delta,\beta}}{\Delta} \le N(p, \mathcal{H}_+; \delta, \beta) \le \max\left\{2, \left\lceil \frac{32}{\epsilon}\frac{\max\{\sqrt{d-1}, u_0\}}{\Delta} \right\rceil \right\}.$$

The lower bound also holds for the full null. The upper bound is achieved by rejecting feasibility when $U_m < \Delta/2$, where

$$U_m = \binom{m}{2}^{-1} \sum_{a<b} \left(\frac{\mathbf{1}\{X_a = X_b\}}{p_{X_a}} - 1 \right).$$

Proof. Let $f(i,j) = \mathbf{1}\{i=j\}/p_i - 1$. Under p, $\sum_j p_j f(i,j) = 0$ and $\mathbb{E}_{p\otimes p} f^2 = d-1$. Thus $\mathbb{E}_p U_m = 0$ and $\mathrm{Var}_p(U_m) = (d-1)/\binom{m}{2}$. Under a component q, let $D = \chi^2(q\|p) \ge \Delta$, $r_i = q_i/p_i$ and $u = \max_i r_i$. Then $\mathbb{E}_q U_m = D$,

$$\zeta_1 = \mathrm{Var}_q(r_i - 1) \le uD, \qquad \zeta_2 \le \mathbb{E}_{q\otimes q} f^2 \le 4(d-1) + 4D/p_{\min} + 1.$$

For the last inequality split the sum $\sum_i r_i^2(1-p_i)^2$ at $r_i = 2$, use $r_i^2 \le 4(r_i-1)^2$ on the second part, and bound the off-diagonal contribution by one. The U-statistic decomposition gives

$$\mathrm{Var}_q(U_m) \le 4\zeta_1/m + 2\zeta_2/[m(m-1)].$$

Chebyshev therefore bounds the reference error by $16(d-1)/(m^2\Delta^2)$ and the component error by $16u/(mD) + 16\zeta_2/(m^2D^2)$ for $m \ge 2$. Since $D \ge p_{\min}(u-1)^2$, we have $u/D \le u_0/\Delta$. Also

$$\sqrt{\zeta_2}/D \le [2\sqrt{d-1} + 2(u_0 - 1) + 1]/\Delta \le 5\max\{\sqrt{d-1}, u_0\}/\Delta.$$

The stated constant controls both errors by ϵ. Validity is componentwise and therefore survives mixing. For the lower bound use a minimizing q and $H^2(p,q) \le \chi^2(q\|p) = \Delta$. The two-point inequality gives $N \ge (1-H^2)/(2H^2)\log[1/(1-(\beta-\delta)^2)]$. If $\Delta \le 1/2$ this is at least the claimed lower bound; if $\Delta > 1/2$, use $N \ge 1$ and $c_{\delta,\beta} \le 1/2$. □

Reference-zero cells. The collision statistic is evaluated only on reference-positive cells. For a full-null extension let Z be the reference-zero cells, $\rho(q) = q(Z)$, and

$$\Delta(\rho_0) = \inf_{q \in \mathcal{H}: \rho(q) \le \rho_0} \chi^2(q(\cdot \mid Z^c) \| p).$$

Combine the collision rejection with the requirement that no sample falls in Z. If $\rho(q) > \rho_0$, rejection probability is at most $(1 - \rho(q))^m \le e^{-m\rho_0}$. Otherwise, conditional on avoiding Z, the samples are i.i.d. from $q(\cdot \mid Z^c)$ and the collision proof applies with $\Delta(\rho_0)$. Thus a valid full-null upper bound uses the maximum of the collision sample bound at this radius and $\log(1/\epsilon)/\rho_0$, whenever $\Delta(\rho_0) > 0$. We do not substitute the supported Δ for this radius in our empirical claims.

G Reference-free intersection–union certificate

For a fixed policy A, let $c(A) = P(S \in A)$ and $r(A) = P(L = 1 \mid S \in A)$. Then $T_A \sim \mathrm{Bin}(n, c(A))$ and, conditional on $T_A = t$, $E_A \sim \mathrm{Bin}(t, r(A))$. Let U_A be a one-sided $(1 - \zeta_c)$ exact upper bound on $c(A)$ and L_A a one-sided $(1 - \zeta_r)$ exact lower bound on $r(A)$, with $\zeta_c + \zeta_r \le \delta$.

G.1 Proof of Proposition 1

If the existential null is true, choose a fixed population witness A^* satisfying $c(A^*) \ge \kappa$ and $r(A^*) \le \alpha$. This witness depends on the population, not on the sample. If the procedure rejects, then either $U_{A^*} < \kappa \le c(A^*)$, or $U_{A^*} \ge \kappa$, in which case A^* participates in the minimization and rejection forces $L_{A^*} > \alpha \ge r(A^*)$. The two one-sided bounds fail with probabilities at most ζ_c and ζ_r, respectively. A union bound over these two events gives at most $\zeta_c + \zeta_r \le \delta$. No event is required to hold simultaneously for all searched policies. The equal split used in our experiments is therefore a convenience rather than a requirement; any predeclared allocation is valid. More generally, a fixed-policy joint confidence region for $(c(A), r(A))$ can replace the rectangular pair whenever its simultaneous failure probability is at most δ and the corresponding coverage/risk projections are used in the same screening logic.

Proposition 10 (Anytime-valid existential certificate)**.** *For each fixed A, replace (U_A, L_A) by one-sided confidence sequences $(U_{A,t}, L_{A,t})$ whose anytime failure probabilities are at most ζ_c for coverage and ζ_r for risk. Reject at the first time t for which $\min_{A: U_{A,t} \ge \kappa} L_{A,t} > \alpha$. Under the existential null, the probability of ever rejecting is at most $\zeta_c + \zeta_r \le \delta$, again without a multiplicity correction over $\mathcal{A}$.*

G.2 Proof of Proposition 10

If the existential null is true, fix a population witness A^*. Let $\mathcal{E}_c$ be the event that the coverage upper confidence sequence ever falls below the true $c(A^*)$, and let $\mathcal{E}_r$ be the event that the risk lower confidence sequence ever rises above the true $r(A^*)$. By assumption, their probabilities are at most ζ_c and ζ_r, respectively. If the anytime procedure ever rejects while neither event occurs, then $U_{A^*,t} \ge c(A^*) \ge \kappa$, so A^* is in the screened family, while $L_{A^*,t} \le r(A^*) \le \alpha$, contradicting rejection. Hence ever-rejection is contained in $\mathcal{E}_c \cup \mathcal{E}_r$ and has probability at most $\zeta_c + \zeta_r \le \delta$. The argument requires simultaneous-in-time validity only for the one fixed population witness, not for all searched policies.

Why a one-policy split-sample comparator is not a valid existential-null test. Selecting one A on one split and rejecting that policy on an independent split proves only that the selected member is infeasible. The null is a union over all $A \in \mathcal{A}$, so another member may still be the population witness. A valid split-sample existential test would still need to reject every potential witness (or otherwise control the union-null logic), restoring the structural problem addressed by the IUT. A "select one, then test one" comparator therefore tests a different null hypothesis, and we do not report it.

G.3 Operational power on theorem-admissible matched families

An additional finite-size control uses $B = 16$, $p = .25$, $\gamma = .08$, $\alpha = .17$, $\kappa = .5$, and $k = 1, 4, 16$. The residual mass is $B^{-2} = 1/256$, which is at most $\gamma/16 = .005$, as required by Theorem 2. The

$k = 16$ profile is uniform. For every (k, n), generate independent multinomial counts over the $2B$ group halves and then binomial errors with common probability p. The boundary-null control instead uses error probability $p - \gamma$ on one selected half of each group and $p + \gamma$ on the other; the selected policy is a population witness.

The sample grid is $n \in \{300, 500, 750, 1000, 1400, 2000\}$ for alternatives and $n \in \{500, 1400\}$ for boundary checks, with 2,000 replicates per point. The artifact records all deterministic seeds. Error bars are pointwise Wilson 95% intervals for Monte Carlo rejection probabilities. The reported n_{80} values linearly interpolate neighboring power estimates; no confidence interval or minimax interpretation is attached to a crossing. The endpoint ratio is $1318.11/572.29 = 2.303$. All grid values, including null controls, are retained below.

Table 6: Additional sign-family simulation: observed rejection percentages, 2,000 replicates per cell. Null rows refer to a feasible boundary component.

Source	n	$k = 1$	$k = 4$	$k = 16$
Alternative	300	42.75	16.95	0.75
Alternative	500	75.20	47.60	7.00
Alternative	750	91.80	77.00	26.85
Alternative	1000	97.45	92.95	54.75
Alternative	1400	99.90	98.90	86.50
Alternative	2000	99.95	99.95	98.35
Boundary null	500	3.50	2.15	0.25
Boundary null	1400	3.95	2.80	1.30

Sign-family dynamic program. Write n_{bh}, e_{bh} for counts and errors in half $h \in \{0, 1\}$ of group b. Initialize $D_0(0) = 0$ and $D_0(t) = +\infty$ otherwise. The exact recurrence is

$$D_b(t) = \min_{h \in \{0,1\}} \{D_{b-1}(t - n_{bh}) + e_{bh}\},$$

with invalid indices assigned infinity. After all B groups, $D_B(t)$ is the least errors among policies accepting exactly t observations. The lower binomial bound is monotone in errors, so applying the coverage screen and risk bound to D_B computes the IUT over all 2^B policies in $O(Bn)$ time and $O(n)$ memory. Exhaustive tests compare every state and final decision with direct policy enumeration on small instances. This recurrence changes computation only; Proposition 1 supplies validity.

Coverage-floor and audit-size sensitivity. On the held-out ImageNetV2 audit with $K = 16$ and cells frozen on disjoint calibration data, we also evaluate $\kappa \in \{.50, .60, .75, .90\}$ and audit sizes $n \in \{500, 1000, 2500, 5000\}$. For the full $n = 5000$ audit the exact 95% IUT lower frontiers are 4.06%, 6.28%, 10.89%, and 17.71%, respectively. In 40 deterministic held-out subsamples, the corresponding median frontiers at $n = 500$ are 0.97%, 2.62%, 6.70%, and 13.55%; at $n = 1000$ they are 2.41%, 4.20%, 8.79%, and 15.64%. This is a finite-benchmark sensitivity analysis, not a population learning curve, but it makes explicit how the certificate weakens at lower coverage floors and smaller audit sizes.

G.4 Weighted interval dynamic program

Let cell m contain n_m audit observations and e_m errors. A policy using at most K contiguous intervals selects a subset of cells that is a union of at most K contiguous intervals. Define

$$e_K(t) = \min \left\{ \sum_{m \in A} e_m : \sum_{m \in A} n_m = t,\ A \text{ uses at most } K \text{ intervals} \right\}.$$

A standard dynamic program over cell index, selected count, number of opened intervals, and whether the current cell is inside an interval computes all $e_K(t)$ exactly. If zero-count cells are possible, an additional one-bit state distinguishes the structurally empty mask from a nonempty rule that happens to accept no audit observation; this leaves the $O(MKn)$ state and $O(Kn)$ memory bounds unchanged. Because the one-sided binomial lower bound is monotone in the observed error count,

the IUT objective reduces to the one-dimensional minimization reported in Section 4. In the reported experiments the coverage floors are large enough that a $t = 0$ rule cannot pass the coverage screen, but the implementation check retains the distinction for completeness.

G.5 Structured DKW simultaneous baseline

To compare the IUT against a policy-count-free baseline, let $F(x) = P(S \le x)$ and $G(x) = P(S \le x, L = 1)$. The latter is the CDF on the extended line of the random variable equal to S when $L = 1$ and $+\infty$ otherwise. By the Dvoretzky–Kiefer–Wolfowitz inequality with Massart's constant (Massart, 1990), allocating failure probability $\delta/2$ to each CDF gives, simultaneously,

$$\|F_n - F\|_\infty \le \epsilon_n, \qquad \|G_n - G\|_\infty \le \epsilon_n, \qquad \epsilon_n = \sqrt{\frac{\log(4/\delta)}{2n}}.$$

For a union of at most K intervals, telescoping its $2K$ endpoints gives $|c_n(A) - c(A)| \le \rho_K$ and $|m_n(A) - m(A)| \le \rho_K$ with $\rho_K = 2K\epsilon_n$, where $m(A) = P(S \in A, L = 1)$. Hence every population policy with $c(A) \ge \kappa$ must satisfy $c_n(A) \ge \kappa - \rho_K$ and

$$r(A) = \frac{m(A)}{c(A)} \ge \frac{[m_n(A) - \rho_K]_+}{\min\{1, c_n(A) + \rho_K\}}.$$

The same weighted interval DP minimizes this expression exactly over the empirical $(t, e_K(t))$ frontier. This comparator exploits the one-dimensional interval structure and contains no factor $|\mathcal{A}_{M,K}|$; it is nevertheless conservative when $K\epsilon_n$ is not small.

G.6 VC/Rademacher simultaneous baseline

The class of unions of at most K intervals on a line has VC dimension $2K$. To compare against a standard uniform-convergence route rather than only endpoint DKW or policywise Bonferroni, define $\mathcal{F}_K = \{(s,l) \mapsto \mathbf{1}\{s \in A\} : A \in \mathcal{A}_{M,K}\}$ and $\mathcal{G}_K = \{(s,l) \mapsto l\mathbf{1}\{s \in A\} : A \in \mathcal{A}_{M,K}\}$. For either $[0,1]$-valued class $\mathcal{F}$, the standard empirical-Rademacher inequality gives, with probability at least $1 - \eta$,

$$\sup_{f \in \mathcal{F}} \{Pf - P_n f\} \le 2\widehat{\mathfrak{R}}_n(\mathcal{F}) + 3\sqrt{\frac{\log(2/\eta)}{2n}},$$

and the reverse deviation follows by applying the same bound to $-\mathcal{F}$ (Mohri et al., 2018). We allocate the data failure budget 0.049 over the four one-sided events for $\mathcal{F}_K$ and $\mathcal{G}_K$. The two empirical Rademacher expectations are evaluated by 200,000 independent sign draws; for each draw the supremum over at most K intervals is solved exactly by a vectorized interval DP. A Hoeffding upper confidence radius with failure probability 5×10^{-4} is added to each Monte-Carlo mean, leaving total data-plus-numerical failure probability at most 0.05.

Let the resulting two-sided radii be $\epsilon_c(K)$ for coverage and $\epsilon_m(K)$ for accepted-error mass. Then every population policy with $c(A) \ge \kappa$ obeys

$$c_n(A) \ge \kappa - \epsilon_c(K), \qquad r(A) \ge \frac{[m_n(A) - \epsilon_m(K)]_+}{\min\{1, c_n(A) + \epsilon_c(K)\}},$$

and the same weighted interval DP minimizes this lower bound over the class. This is a generic VC/Rademacher comparator, not a claim that the bound is sharp for interval unions.

H Experimental details and complete empirical tables

H.1 Matched-family efficiency simulation

We fix $B = 256$, $p = .25$, $\kappa = .5$, $k \in \{1, 4, 16, 64, 256\}$ and $\gamma \in \{.04, .02\}$. The finite sample-size grid for IUT multiplies $(.12k + .9\sqrt{k} + 1.5)/\gamma^2$ by $.4, .6, .8, 1, 1.25, 1.6, 2$, rounded to multiples of 50; every crossing is bracketed. These grids are computational design choices, not fitted rate claims. Each of the 70 alternative grid points uses 1,500 independent audits. Ten boundary checks, one per (k, γ), use the grid point nearest observed 80% power, a boundary-null law and independent

Table 7: Measured 80%-power sample crossings for fixed $B = 256$. NP denotes the simple-mixture envelope, not a full-null-valid test. Ratios compare IUT to that envelope; no minimax constant is inferred.

	$\gamma = .04$			$\gamma = .02$	
k	IUT	NP	ratio	IUT	NP
1	2,342	1,001	2.34	9,216	4,030
4	3,298	1,478	2.23	13,437	5,883
16	5,370	2,363	2.27	22,386	9,374
64	12,065	4,060	2.97	48,268	15,859
256	32,796	7,384	4.44	131,679	29,637

randomness, with 20,000 audits each (`boundary_all.py`). For $k = 1, 4, 16, 64, 256$ the rejection rates are 4.73, 2.75, 1.44, 0.37 and 0.00% at $\gamma = .04$ and 4.73, 2.59, 1.66, 0.29 and 0.00% at $\gamma = .02$; the largest Wilson upper limit is 5.03%. The alternatives use 105,000 IUT audits in total. Seeds are recorded in the scripts and depend on the configuration.

For the NP envelope, the grid multiplies $(.7\sqrt{k} + 1)/\gamma^2$ by $.5, .7, .9, 1.1, 1.35, 1.65, 2$, rounded to multiples of 25. At each of 70 points, 12,000 independent mixture draws calibrate a nominal 5% likelihood-ratio threshold; 6,000 fresh reference draws estimate power and 6,000 fresh mixture draws estimate size. The factorized likelihood ratio is invariant under flipping any group half, so sampling the all-+ feasible component gives the same LR distribution as sampling the uniform component mixture. We test this symmetry directly. Calibration is independent of both evaluations. Pointwise Wilson intervals for the evaluations condition on the calibrated threshold and do not include its calibration uncertainty. The empirical size column is retained; the fitted threshold is not asserted to have exact 5% size. The population NP envelope bounds full-null-valid power, but the estimated curve is descriptive.

For exact computation, subtract the smaller half count in every group from the accepted-count state. The reachable width is then at most $1 + \sum_b |n_{b,+} - n_{b,-}|$, rather than $n + 1$. A compiled implementation performs the same two-choice recurrence over this interval, with unreachable states at infinity. This compression is exact and is checked against the independent uncompressed dynamic program, including final binomial decisions. Source code and a compiler invocation are supplied; no prebuilt binary is needed.

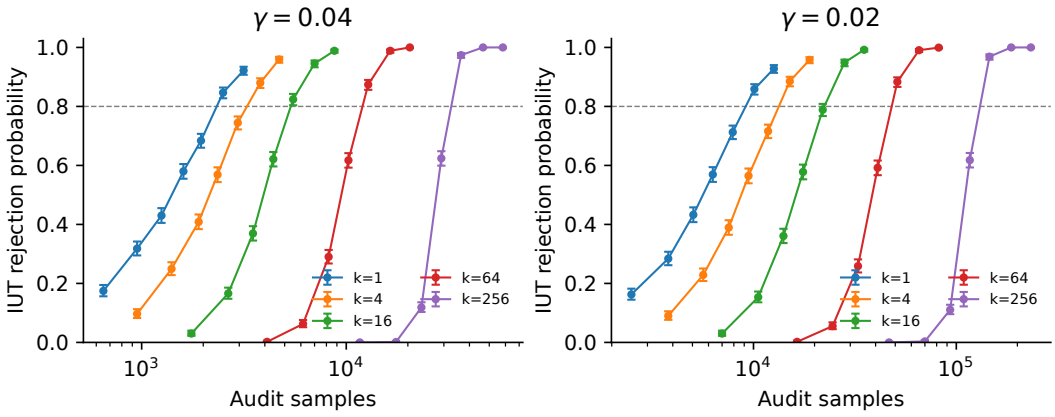

Figure 4: Exact IUT power on the two risk-gap grids. Bars are pointwise Wilson 95% intervals over 1,500 audits. These are Monte Carlo uncertainties, not confidence bands over the sample-size grid.

The recorded constructive-witness fraction selects the half with more correct observations in each heavy group and a fixed half in residual groups. Whenever it has accepted count at least $n/2$ and empirical risk at most α, IUT rejection is impossible; the simulation checks this deterministic implication. The theorem does not require this particular witness to be optimal.

H.2 STRUCTURED CERTIFICATE COMPARISON ON ALL FROZEN OUTPUTS

Semi-synthetic power study. The protocol is described in Section 6. Scripts: `semisynthetic_power.py`, `make_semisyn_table.py`, `make_semisyn_full.py`. Table 8 lists the IUT for every K and the other certificates at $K = 4$.

Table 8: Semi-synthetic audits, all configurations ($\kappa = .9$; 500 audits per cell for $n \leq 5{,}000$, 300 at $n = 10{,}000$, 200 at $n = 20{,}000$). Entries are rejection rates at $\alpha = r^* - \Delta$; the column $\Delta = 0$ is the level at the true boundary claim. For CIFAR-10, $\Delta = .10$ gives $\alpha < 0$ and is rejected trivially.

Population	Method	n	K	$\Delta = 0$	.02	.05	.10	median frontier
CIFAR-10	IUT	500	1	0.006	0.13	0.92	1.00	0.044
CIFAR-10	IUT	500	4	0.000	0.06	0.92	1.00	0.037
CIFAR-10	Bonferroni	500	4	0.000	0.00	0.00	1.00	0.003
CIFAR-10	DKW	500	4	0.000	0.00	0.00	1.00	0.000
CIFAR-10	Total loss	500	4	0.000	0.00	0.03	1.00	0.000
CIFAR-10	IUT	500	16	0.000	0.05	0.92	1.00	0.037
CIFAR-10	IUT	1,000	1	0.004	0.24	1.00	1.00	0.052
CIFAR-10	IUT	1,000	4	0.000	0.20	1.00	1.00	0.048
CIFAR-10	Bonferroni	1,000	4	0.000	0.00	0.00	1.00	0.013
CIFAR-10	DKW	1,000	4	0.000	0.00	0.00	1.00	0.000
CIFAR-10	Total loss	1,000	4	0.000	0.00	0.02	1.00	0.003
CIFAR-10	IUT	1,000	16	0.000	0.19	1.00	1.00	0.048
CIFAR-10	IUT	2,500	1	0.004	0.67	1.00	1.00	0.061
CIFAR-10	IUT	2,500	4	0.000	0.73	1.00	1.00	0.058
CIFAR-10	Bonferroni	2,500	4	0.000	0.00	0.90	1.00	0.030
CIFAR-10	DKW	2,500	4	0.000	0.00	0.00	1.00	0.000
CIFAR-10	Total loss	2,500	4	0.000	0.00	0.03	1.00	0.010
CIFAR-10	IUT	2,500	16	0.000	0.73	1.00	1.00	0.058
CIFAR-10	IUT	5,000	1	0.000	0.95	1.00	1.00	0.064
CIFAR-10	IUT	5,000	4	0.000	0.99	1.00	1.00	0.062
CIFAR-10	Bonferroni	5,000	4	0.000	0.00	1.00	1.00	0.041
CIFAR-10	DKW	5,000	4	0.000	0.00	0.00	1.00	0.000
CIFAR-10	Total loss	5,000	4	0.000	0.00	0.01	1.00	0.013
CIFAR-10	IUT	5,000	16	0.000	0.99	1.00	1.00	0.062
CIFAR-100	IUT	500	1	0.008	0.06	0.42	0.99	0.212
CIFAR-100	IUT	500	4	0.002	0.02	0.23	0.97	0.199
CIFAR-100	Bonferroni	500	4	0.000	0.00	0.00	0.00	0.090
CIFAR-100	DKW	500	4	0.000	0.00	0.00	0.00	0.000
CIFAR-100	Total loss	500	4	0.004	0.03	0.25	0.94	0.195
CIFAR-100	IUT	500	16	0.002	0.02	0.21	0.96	0.197
CIFAR-100	IUT	1,000	1	0.002	0.10	0.78	1.00	0.228
CIFAR-100	IUT	1,000	4	0.002	0.04	0.63	1.00	0.217
CIFAR-100	Bonferroni	1,000	4	0.000	0.00	0.00	0.01	0.134
CIFAR-100	DKW	1,000	4	0.000	0.00	0.00	0.00	0.000
CIFAR-100	Total loss	1,000	4	0.000	0.02	0.38	1.00	0.207
CIFAR-100	IUT	1,000	16	0.002	0.04	0.63	1.00	0.217
CIFAR-100	IUT	2,500	1	0.004	0.30	1.00	1.00	0.241
CIFAR-100	IUT	2,500	4	0.000	0.21	0.99	1.00	0.234
CIFAR-100	Bonferroni	2,500	4	0.000	0.00	0.00	0.96	0.179
CIFAR-100	DKW	2,500	4	0.000	0.00	0.00	0.00	0.000
CIFAR-100	Total loss	2,500	4	0.000	0.01	0.68	1.00	0.217
CIFAR-100	IUT	2,500	16	0.000	0.20	0.99	1.00	0.234
CIFAR-100	IUT	5,000	1	0.014	0.62	1.00	1.00	0.249
CIFAR-100	IUT	5,000	4	0.008	0.55	1.00	1.00	0.243
CIFAR-100	Bonferroni	5,000	4	0.000	0.00	0.09	1.00	0.203
CIFAR-100	DKW	5,000	4	0.000	0.00	0.00	0.00	0.000
CIFAR-100	Total loss	5,000	4	0.000	0.00	0.91	1.00	0.223
CIFAR-100	IUT	5,000	16	0.008	0.54	1.00	1.00	0.243
CIFAR-100	IUT	10,000	1	0.010	0.92	1.00	1.00	0.255
CIFAR-100	IUT	10,000	4	0.000	0.90	1.00	1.00	0.249
CIFAR-100	Bonferroni	10,000	4	0.000	0.00	0.98	1.00	0.221
CIFAR-100	DKW	10,000	4	0.000	0.00	0.00	0.00	0.042
CIFAR-100	Total loss	10,000	4	0.000	0.00	1.00	1.00	0.226
CIFAR-100	IUT	10,000	16	0.000	0.90	1.00	1.00	0.249
CIFAR-100	IUT	20,000	1	0.015	0.99	1.00	1.00	0.260
CIFAR-100	IUT	20,000	4	0.000	0.99	1.00	1.00	0.254
CIFAR-100	Bonferroni	20,000	4	0.000	0.00	1.00	1.00	0.231
CIFAR-100	DKW	20,000	4	0.000	0.00	0.00	0.00	0.104
CIFAR-100	Total loss	20,000	4	0.000	0.00	1.00	1.00	0.228
CIFAR-100	IUT	20,000	16	0.000	0.99	1.00	1.00	0.254
ImageNetV2	IUT	500	1	0.002	0.05	0.44	1.00	0.146
ImageNetV2	IUT	500	4	0.000	0.02	0.34	1.00	0.135
ImageNetV2	Bonferroni	500	4	0.000	0.00	0.00	0.00	0.047
ImageNetV2	DKW	500	4	0.000	0.00	0.00	0.00	0.000
ImageNetV2	Total loss	500	4	0.000	0.02	0.19	0.95	0.123
ImageNetV2	IUT	500	16	0.000	0.02	0.34	0.99	0.134
ImageNetV2	IUT	1,000	1	0.004	0.11	0.82	1.00	0.162
ImageNetV2	IUT	1,000	4	0.002	0.09	0.80	1.00	0.153

Population	Method	n	K	$\Delta = 0$	.02	.05	.10	median
ImageNetV2	Bonferroni	1,000	4	0.000	0.00	0.00	0.16	0.083
ImageNetV2	DKW	1,000	4	0.000	0.00	0.00	0.00	0.000
ImageNetV2	Total loss	1,000	4	0.000	0.00	0.28	1.00	0.135
ImageNetV2	IUT	1,000	16	0.002	0.08	0.80	1.00	0.153
ImageNetV2	IUT	2,500	1	0.010	0.27	1.00	1.00	0.174
ImageNetV2	IUT	2,500	4	0.000	0.31	1.00	1.00	0.169
ImageNetV2	Bonferroni	2,500	4	0.000	0.00	0.00	1.00	0.120
ImageNetV2	DKW	2,500	4	0.000	0.00	0.00	0.00	0.000
ImageNetV2	Total loss	2,500	4	0.000	0.00	0.52	1.00	0.144
ImageNetV2	IUT	2,500	16	0.000	0.31	1.00	1.00	0.169
ImageNetV2	IUT	5,000	1	0.004	0.52	1.00	1.00	0.180
ImageNetV2	IUT	5,000	4	0.000	0.66	1.00	1.00	0.176
ImageNetV2	Bonferroni	5,000	4	0.000	0.00	0.28	1.00	0.140
ImageNetV2	DKW	5,000	4	0.000	0.00	0.00	0.00	0.000
ImageNetV2	Total loss	5,000	4	0.000	0.00	0.79	1.00	0.149
ImageNetV2	IUT	5,000	16	0.000	0.66	1.00	1.00	0.176
ImageNetV2	IUT	10,000	1	0.013	0.79	1.00	1.00	0.187
ImageNetV2	IUT	10,000	4	0.003	0.94	1.00	1.00	0.181
ImageNetV2	Bonferroni	10,000	4	0.000	0.00	1.00	1.00	0.156
ImageNetV2	DKW	10,000	4	0.000	0.00	0.00	0.00	0.000
ImageNetV2	Total loss	10,000	4	0.000	0.00	0.97	1.00	0.152
ImageNetV2	IUT	10,000	16	0.003	0.94	1.00	1.00	0.181
ImageNetV2	IUT	20,000	1	0.015	0.99	1.00	1.00	0.193
ImageNetV2	IUT	20,000	4	0.000	1.00	1.00	1.00	0.184
ImageNetV2	Bonferroni	20,000	4	0.000	0.00	1.00	1.00	0.167
ImageNetV2	DKW	20,000	4	0.000	0.00	0.00	0.00	0.045
ImageNetV2	Total loss	20,000	4	0.000	0.00	1.00	1.00	0.155
ImageNetV2	IUT	20,000	16	0.000	1.00	1.00	1.00	0.184
CMU-MOSEI	IUT	500	1	0.012	0.12	0.72	1.00	0.104
CMU-MOSEI	IUT	500	4	0.002	0.03	0.53	1.00	0.088
CMU-MOSEI	Bonferroni	500	4	0.000	0.00	0.00	0.09	0.025
CMU-MOSEI	DKW	500	4	0.000	0.00	0.00	0.00	0.000
CMU-MOSEI	Total loss	500	4	0.000	0.00	0.03	0.79	0.052
CMU-MOSEI	IUT	500	16	0.002	0.02	0.48	1.00	0.086
CMU-MOSEI	IUT	1,000	1	0.006	0.23	0.97	1.00	0.116
CMU-MOSEI	IUT	1,000	4	0.000	0.10	0.97	1.00	0.103
CMU-MOSEI	Bonferroni	1,000	4	0.000	0.00	0.00	0.98	0.049
CMU-MOSEI	DKW	1,000	4	0.000	0.00	0.00	0.00	0.000
CMU-MOSEI	Total loss	1,000	4	0.000	0.00	0.02	0.98	0.058
CMU-MOSEI	IUT	1,000	16	0.000	0.09	0.96	1.00	0.102
CMU-MOSEI	IUT	2,500	1	0.004	0.60	1.00	1.00	0.127
CMU-MOSEI	IUT	2,500	4	0.000	0.52	1.00	1.00	0.117
CMU-MOSEI	Bonferroni	2,500	4	0.000	0.00	0.08	1.00	0.079
CMU-MOSEI	DKW	2,500	4	0.000	0.00	0.00	0.00	0.000
CMU-MOSEI	Total loss	2,500	4	0.000	0.00	0.00	1.00	0.067
CMU-MOSEI	IUT	2,500	16	0.000	0.50	1.00	1.00	0.116
CMU-MOSEI	IUT	5,000	1	0.018	0.91	1.00	1.00	0.132
CMU-MOSEI	IUT	5,000	4	0.010	0.91	1.00	1.00	0.124
CMU-MOSEI	Bonferroni	5,000	4	0.000	0.00	0.96	1.00	0.095
CMU-MOSEI	DKW	5,000	4	0.000	0.00	0.00	0.00	0.000
CMU-MOSEI	Total loss	5,000	4	0.000	0.00	0.01	1.00	0.072
CMU-MOSEI	IUT	5,000	16	0.010	0.91	1.00	1.00	0.124

Fixed-split frontiers on 12 outputs. We use one predeclared split, seed 3401, for each of the 12 frozen model outputs. Calibration uses half the scores and no audit labels; audit counts use the disjoint half. We report every combination of $K = 1, 4, 16$ at $\kappa = .9$. For each configuration, IUT, policywise Bonferroni, and DKW use the nominal error budget $.05$. The Rademacher calculation uses 200,000 signs for each of the coverage and accepted-error classes; Hoeffding upper confidence corrections cost $.0005$ each, with the remaining $.049$ allocated over four one-sided data events. Guarantees are pointwise in K, not simultaneous over all reported configurations. These methods have different inferential targets internally but each resulting frontier is valid for the same existential contract. The global-loss comparator is deliberately described as score-agnostic, and no data-dependent maximum of unadjusted certificates is used.

Table 9: Fixed-split structured comparison at $\kappa = .9$. All frontiers are percentages, pointwise 95% under the stated sampling assumptions. Numerical uncertainty is included for the empirical-Rademacher comparator.

Output / K	IUT	Bonf.	DKW	Rad.	Global
CIFAR-10 / 1	6.40	5.64	0.99	0.00	1.37
CIFAR-10 / 4	6.22	4.16	0.00	0.00	1.37

Output / K	IUT	Bonf.	DKW	Rad.	Global
CIFAR-10 / 16	6.22	3.28	0.00	0.00	1.37
CIFAR-100 / 1	24.56	23.16	18.87	8.35	22.25
CIFAR-100 / 4	24.29	20.37	0.00	4.29	22.25
CIFAR-100 / 16	24.29	18.82	0.00	1.93	22.25
CIFAR-100 (2) / 1	24.11	21.62	17.75	8.28	21.50
CIFAR-100 (2) / 4	23.70	19.74	0.00	3.93	21.50
CIFAR-100 (2) / 16	23.70	17.82	0.00	1.44	21.50
CIFAR-100 (3) / 1	23.58	22.20	17.46	8.09	21.45
CIFAR-100 (3) / 4	23.58	19.77	0.00	3.77	21.45
CIFAR-100 (3) / 16	23.58	17.77	0.00	1.63	21.45
CIFAR-10 (2) / 1	6.61	5.26	1.20	0.00	1.39
CIFAR-10 (2) / 4	6.57	4.39	0.00	0.00	1.39
CIFAR-10 (2) / 16	6.57	3.41	0.00	0.00	1.39
CIFAR-10 (3) / 1	6.81	5.24	0.84	0.00	1.17
CIFAR-10 (3) / 4	6.17	3.91	0.00	0.00	1.17
CIFAR-10 (3) / 16	6.17	3.36	0.00	0.00	1.17
ImageNetV2 / 1	17.78	16.55	12.45	2.95	14.89
ImageNetV2 / 4	17.71	14.13	0.00	0.00	14.89
ImageNetV2 / 16	17.71	12.58	0.00	0.00	14.89
ImageNetV2 (2) / 1	24.83	22.61	18.39	8.27	22.53
ImageNetV2 (2) / 4	24.12	20.22	0.00	4.24	22.53
ImageNetV2 (2) / 16	24.12	18.83	0.00	2.04	22.53
ImageNetV2 (3) / 1	27.43	25.26	21.18	10.35	25.68
ImageNetV2 (3) / 4	27.38	23.04	0.00	6.24	25.68
ImageNetV2 (3) / 16	27.24	21.20	0.00	3.69	25.68
MOSEI / 1	12.44	9.94	3.80	0.00	6.53
MOSEI / 4	11.58	7.20	0.00	0.00	6.53
MOSEI / 16	11.58	5.45	0.00	0.00	6.53
MOSEI (2) / 1	10.80	8.72	2.89	0.00	5.68
MOSEI (2) / 4	10.80	6.56	0.00	0.00	5.68
MOSEI (2) / 16	10.80	4.86	0.00	0.00	5.68
MOSEI (3) / 1	13.75	11.47	5.24	0.00	8.65
MOSEI (3) / 4	13.51	8.56	0.00	0.00	8.65
MOSEI (3) / 16	13.30	6.55	0.00	0.00	8.65

H.3 HELD-OUT IMAGENETV2 IUT AUDIT

For the reference-free certificate, we use the frozen ImageNetV2 ViT-B/16 output and split it 50/50 with seed 3401. The calibration half defines 57 ordered cells from the stored `max_softmax_score` channel; the audit half contains 5,000 images and supplies all labels/counts. We search unions of at most K contiguous cell intervals and report the 90%-coverage contract. Alongside the exact IUT we report policywise Bonferroni, the structured DKW comparator of Appendix G.5, and the VC/Rademacher comparator of Appendix G.6.

Table 10: Held-out ImageNetV2 audit at $\kappa = .90$. Lower frontiers are 95% valid under each method's stated probability budget. "Rad." is the Monte-Carlo-certified empirical-Rademacher uniform-convergence comparator.

K	policy count	empirical	IUT	Bonf.	DKW	Rad.
1	1.65×10^3	19.996%	17.783%	16.552%	12.451%	2.947%
2	4.26×10^5	19.410%	17.783%	15.242%	4.660%	1.085%
4	1.96×10^9	19.348%	17.712%	14.135%	0.000%	0.000%
8	9.10×10^{13}	19.348%	17.712%	13.321%	0.000%	0.000%
16	1.23×10^{17}	19.348%	17.712%	12.584%	0.000%	0.000%

The IUT value stabilizes once additional intervals no longer improve the relevant frontier. DKW accumulates endpoint error linearly in K, while the empirical-Rademacher baseline explicitly controls the VC-$2K$ interval-union class without a raw policy-count correction but is still conservative at this sample size. Thus the comparison does not rest on Bonferroni as the only alternative. A one-policy split-sample select-then-test procedure is not listed because it tests only the selected member and is

not level-valid for the existential union null. We do not claim any displayed comparator is minimax optimal for the existential problem. On the supplied CPU implementation ($M = 57, n = 5000$), median DP times are 1.15, 1.60, 2.43, 4.34, 8.13, and 14.80 ms for $K = 1, 2, 4, 8, 16, 32$, respectively; absolute timing is environment-specific.

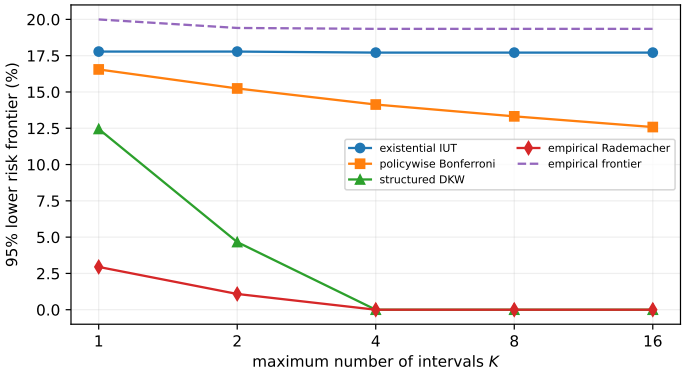

Figure 5: Held-out ImageNetV2 audit at $\kappa = .90$. The exact IUT frontier is shown with policywise Bonferroni, structured DKW, empirical-Rademacher uniform convergence, and the empirical frontier. This illustrates Proposition 1; it is not a minimax-optimality comparison.

Boundary-null implementation check. We additionally simulate 1,000 i.i.d. audits from a finite-cell null containing a known boundary witness: 12 equally likely score cells, A^* equal to the first nine cells, $c(A^*) = .75$, $r(A^*) = .20$, $K = 3$, and $n = 300$. The IUT+DP implementation rejects the true null in 9/1,000 runs, an empirical rate of 0.9% with 95% Wilson interval 0.47–1.70%. This is intentionally only an implementation regression check; Proposition 1 supplies the exact finite-sample level guarantee.

Operating-characteristic check. To distinguish conservativeness at the boundary from lack of power, we keep the same $M = 12$, $K = 3$, $n = 300$, $\kappa = .75$, and target $\alpha = .20$ design and vary the best population risk. Over 2,000 simulations per point, rejection rates are:

best population risk	.20	.22	.25	.30	.35
IUT rejection rate	.0115	.0555	.3015	.8620	.9955

This is a descriptive power curve, not a replacement for Proposition 1's exact level guarantee. It shows that the 0.9–1.2% boundary rejection rate is a feature of conservative exact one-sided bounds rather than a procedure that remains inert away from the boundary.

Family-flexibility power cost. The phrase "no multiplicity correction" concerns the level proof, not immunity of power to the size of the searched family. To isolate this distinction, we simulate a homogeneous alternative with $M = 12$, $n = 300$, $\kappa = .75$, target $\alpha = .20$, and population risk .25 in every score cell. Hence every policy in every interval-union class has exactly the same population risk; increasing K creates no better population witness. Over 3,000 simulations, IUT rejection falls from 23.1% at $K = 1$ to 7.6% at $K = 2$, 4.7% at $K = 3$, and 4.4% at $K = 4$ and 6. The median lower frontier simultaneously falls from 18.17% to 16.03%. Thus family flexibility can materially reduce power even though the exact type-I argument pays no raw $|\mathcal{A}|$ factor.

H.4 Abstention-odds simulation

Design. Scores are uniform over $M = 64$ equal-mass cells and $L \sim \mathrm{Bern}(.14)$ independently of the score, so every union of cells has risk $.14 = \alpha + \gamma$ with $\alpha = .10$, $\gamma = .04$; the law lies in $\mathcal{P}_\gamma(\mathcal{A}_K)$ for every K and κ, and every rule is exactly as bad as every other. The menus are unions of at most $K \in \{1, 2, 4, 8, 16\}$ intervals of cells, and the IUT uses $\zeta_c = \zeta_r = .025$. On a grid of log-spaced sizes from 150 to 15,846 (the first 15 points of a 19-point grid up to 60,000; the run was stopped

once every power had reached 1) we draw 250 audits per size, with seeds 6101 and 6102 for the two interleaved halves of the grid. One dynamic program per audit serves all (K, κ). Crossings are log-linear interpolations of Monte Carlo power; with 250 audits the power standard error is at most .032, which moves a crossing by a few percent.

Results. Table 11 lists all crossings. The normalization divides by the abstention odds and by $\log(e/\min\{\kappa, 1-\kappa\})$, a logarithmic factor suggested by localized VC bounds (Appendix B.3); without that logarithm the normalized costs still vary much less than the raw ones (at $K = 4$: 3.9 at $\kappa = .5$ and 9.8 at $\kappa = .95$, against a raw variation of 7.7-fold in the opposite direction). At $\kappa = .3$ the normalized costs are 0.56–0.64 times those at $\kappa = .5$, within the constant factor that separates $(1-\kappa)/\kappa$ from $\min\{\kappa, 1-\kappa\}/\kappa^2$. At $\kappa \geq .9$ and $K \geq 8$ the crossings stop growing: the abstained region contains only $M(1-\kappa) \leq 6.4$ cells, so no new rule competes. The constant-risk law is not the least favorable law of Theorem 1; the experiment checks the κ-dependence of the IUT's cost, which is governed by the upper bound. Script: `kappa_law.py`; figure and table: `make_kappa_figure.py`, `make_kappa_table.py`.

Table 11: Abstention-odds simulation: 80%-power crossings n_{80} of the exact IUT ($M = 64$ equal cells, constant risk .14, $\alpha = .10$, $\gamma = .04$, 250 audits per grid point, log-linear interpolation), and normalized menu costs $\gamma^2\{n_{80}(K) - n_{80}(1)\}/\{\frac{1-\kappa}{\kappa}\log\frac{e}{\min(\kappa,1-\kappa)}\}$.

$K \setminus \kappa$	.3	.5	.7	.8	.9	.95
1	3,832	1,958	1,070	881	755	656
2	5,764	3,092	1,847	1,402	973	839
4	8,252	4,414	2,737	1,887	1,316	977
8	11,961	6,151	3,458	2,584	1,568	1,024
16	14,050	7,389	3,921	2,732	1,580	1,024
norm., $K = 2$	0.60	1.07	1.32	1.28	0.95	1.39
norm., $K = 4$	1.38	2.32	2.82	2.47	2.45	2.45
norm., $K = 8$	2.53	3.96	4.04	4.18	3.54	2.80
norm., $K = 16$	3.18	5.13	4.83	4.54	3.60	2.80

H.5 A VALID NON-EFC AUDIT: CROSS-FIT UNIVERSAL INFERENCE

Identify each law with its pushforward to the $2M$ cell-by-loss categories; as in Appendix B.2, this loses nothing for the null because every rule is a union of cells. For $A \in \mathcal{A}$ let $\mathcal{H}_A$ be the laws under which A meets the contract, so $\mathcal{H} = \bigcup_A \mathcal{H}_A$.

Lemma 4 (Profile likelihood of the existential null)**.** *Let x be category counts with total n_0, and let $\hat{c}$ be the empirical coverage of A. When $\hat{c} > 0$, let $\hat{r}$ be its empirical risk; when $\hat{c} = 0$, define the product $\hat{c}\,\mathbf{1}\{\hat{r} > \alpha\}\,\mathrm{kl}(\hat{r}, \alpha)$ below to be zero. Then*

$$\sup_{Q \in \mathcal{H}_A} \sum_i x_i \log Q_i = \sum_i x_i \log \frac{x_i}{n_0} - n_0 \Big[\mathbf{1}\{\hat{c} < \kappa\}\,\mathrm{kl}(\hat{c}, \kappa) + \hat{c}\,\mathbf{1}\{\hat{r} > \alpha\}\,\mathrm{kl}(\hat{r}, \alpha)\Big].$$

Proof. The constraints of $\mathcal{H}_A$ involve Q only through the four masses $Q(A,0), Q(A,1), Q(A^c,0), Q(A^c,1)$. For fixed masses, the likelihood is maximized by splitting each mass in proportion to the counts, which gives $\sum_i x_i \log(x_i/n_0) - n_0\,\mathrm{KL}(\hat{p}\|Q_4)$, where $\hat{p}$ and Q_4 are the empirical and model four-cell laws. Parametrize Q_4 by $c = Q(A)$, $r = Q(L = 1 \mid A)$ and $s = Q(L = 1 \mid A^c)$. The chain rule gives $\mathrm{KL}(\hat{p}\|Q_4) = \mathrm{kl}(\hat{c}, c) + \hat{c}\,\mathrm{kl}(\hat{r}, r) + (1-\hat{c})\mathrm{kl}(\hat{s}, s)$, with zero-weight conditional terms defined as zero (so the empirical conditional proportion on an empty block may be chosen arbitrarily). The constraints are $c \geq \kappa$ and $r \leq \alpha$, and the three terms are minimized separately by $c = \max\{\hat{c}, \kappa\}$, by $r = \min\{\hat{r}, \alpha\}$ when $\hat{c} > 0$, and by $s = \hat{s}$ when $\hat{c} < 1$; the corresponding zero-weight choices are immaterial. □

For $t > 0$, the bracket is increasing in the error count for a fixed accepted count, so the minimal error counts $e_K(t)$ of the interval dynamic program (Appendix G) suffice. If a nonempty frozen score cell receives zero observations in the evaluated split, a nonempty rule with $t = 0$ is also available and contributes the coverage-only penalty $\mathrm{kl}(0, \kappa)$; we include this candidate explicitly. The structurally empty mask is not a witness for the existential null. On small instances the formula agrees with

numerical constrained optimization to 10^{-9}, and the dynamic-program minimum, including the zero-count edge case, agrees with brute-force enumeration.

The audit. Split the audit sample into halves D_0, D_1. Let $\hat{P}_1$ be the add-$1/2$ smoothed empirical law of D_1 over the $2M$ categories, $\Lambda_0 = \hat{P}_1^{n_0}(D_0)/\sup_{Q\in\mathcal{H}} Q^{n_0}(D_0)$ the split likelihood ratio, Λ_1 its mirror image, and reject when $(\Lambda_0 + \Lambda_1)/2 \geq 1/\delta$. For every $Q \in \mathcal{H}$, $\mathbb{E}_Q[\Lambda_0 \mid D_1] \leq \mathbb{E}_Q[\hat{P}_1^{n_0}(D_0)/Q^{n_0}(D_0) \mid D_1] \leq 1$: explicitly, the latter expectation sums $\hat{P}_1^{n_0}$ only over sample sequences with positive Q^{n_0} mass, and equals one when $\hat{P}_1 \ll Q$. The same holds for Λ_1, so the average is an e-value and Markov's inequality gives level δ over the unrestricted null (Wasserman et al., 2020). The audit is not EFC: it may reject although a rule is empirically feasible on the full sample. It is a natural reference-free competitor, and it is not the testing-scale audit of Theorem 4, which requires constant risk.

Results. Table 12 compares it with the IUT on the same draws. No boundary rejection was observed for the universal-inference audit. The IUT's largest boundary rejection rate in these runs is 4.0% (ImageNetV2, $K = 1$, $n = 50{,}000$, 200 audits), within Monte Carlo error of its exact level; up to $n = 20{,}000$ it matches the at most 1.8% of Table 8 within that error. The χ^2 independence test rejected in all 200 audits at $n = 500$ on every population. The numerator must fit the full cell-by-loss law, which has up to $2M - 1 = 113$ free parameters here (105 for CIFAR-10 after ties are merged); this is a substantial finite-sample burden that the IUT does not incur. Script: `nonefc_semisyn.py`; table: `make_nonefc_table.py`.

Table 12: A valid non-EFC audit on the frozen populations ($\kappa = .9$, $K = 4$, 200 audits per point, $n \in \{500, 1000, 2500, 5000, 10^4, 2\times10^4, 5\times10^4\}$). Entries give the smallest grid size at which the rejection rate of the false claim $\alpha = r^* - \Delta$ reaches 80%. "Level" is the largest rejection rate of a true boundary claim over all n and $K \in \{1, 4, 16\}$. The last column is the median p-value of a χ^2 test of cell–loss independence at $n = 500$.

		$\Delta = .05$		$\Delta = .02$		level		independence
Population	r^*	IUT	split LR	IUT	split LR	IUT	split LR	median p
CIFAR-10	0.074	500	10,000	5,000	$> 50{,}000$	2.0%	0.0%	2e-17
CIFAR-100	0.262	2,500	50,000	10,000	$> 50{,}000$	3.0%	0.0%	3e-19
ImageNetV2	0.193	2,500	50,000	10,000	$> 50{,}000$	4.0%	0.0%	3e-17
CMU-MOSEI	0.136	1,000	50,000	5,000	$> 50{,}000$	3.5%	0.0%	5e-10

I ARTIFACT PROVENANCE AND AUTOMATED CHECKS

The release includes the frozen dependency artifact plus an analysis directory containing the operational benchmark, sign-family simulation, full-null window reconstruction, exact parity certificates, validation tests, and generated manuscript tables. All computational configurations have recorded deterministic seeds. Figures and numerical controls computed from frozen outputs are identified as such; their source scripts and input arrays are included. The parity certificates are verified with rational arithmetic; the sign dynamic program and binomial rejection lookup are checked against independent exhaustive calculations; singleton radii are checked against constrained four-cell numerical optimization. The release README supplies commands, software versions, and claim-to-file mappings.